Linearly-Involved Moreau-Enhanced-Over-Subspace Model: Debiasd Sparse Modeling and Stable Outlier-Robust Regression

Masahiro Yukawa, Senior Member, IEEE, Hiroyuki Kaneko, Member, IEEE, Kyohei Suzuki, Graduate Student Member, IEEE, and Isao Yamada, Fellow, IEEE

Abstract—We present an efficient mathematical framework to derive promising methods that enjoy “enhanced” desirable properties. The popular minimax concave penalty for sparse modeling subtracts, from the \(\ell_1\) norm, its Moreau envelope, inducing nearly unbiased estimates and thus yielding considerable performance enhancements. To extend it to underdetermined linear systems, we propose the projective minimax concave penalty, which leads to “enhanced” sparseness over the input subspace. We also present a promising regression method which has an “enhanced” robustness and substantial stability by distinguishing outlier and noise explicitly. The proposed framework, named the linearly-involved Moreau-enhanced-over-subspace (LiMES) model, encompasses those two specific examples as well as two others: stable principal component pursuit and robust classification. The LiMES function involved in the model is an “additively nenseparable” weakly convex function, while the “inner” objective function to define the Moreau envelope is “separable”. This mixed nature of separability and nonseparability allows an application of the LiMES model to the underdetermined case with an efficient algorithmic implementation. Two linear/affine operators play key roles in the model: one corresponds to the projection mentioned above and the other takes care of robust regression/classification. A necessary and sufficient condition for convexity of the smooth part of the objective function is studied. Numerical examples show the efficacy of LiMES in applications to sparse modeling and robust regression.

Index Terms—Convex optimization, weakly convex function, proximity operator, Moreau envelope.

I. INTRODUCTION

The primal goal of this article is to present a unified mathematical framework to derive promising methods that enjoy “enhanced” desirable properties. The main body is divided into two parts. The first part concerns two specific tasks of signal processing. Specifically, the first task is finding sparse solutions of underdetermined linear systems with small biases, and we present a certain data-dependent penalty function yielding “enhanced” sparseness. The second task is a robust regression task in the presence of sparse outliers and large Gaussian noise, and we present an efficient formulation that leads to “enhanced” robustness and substantial stability. The second part presents the proposed framework which contains the linearly-involved Moreau-enhanced-over-subspace (LiMES) model at its core. The proposed framework covers the two methods studied in the first part as well as many others, including two more examples presented in the second part. The background of the study of the first part is presented below, followed by the details of each part.

A. Background

Sparsity awareness and outlier robustness are two key aspects of paramount importance in regression (linear estimation), which has a wide range of applications in many fields including signal processing and machine learning [1], [2]. The \(\ell_1\) penalty and the \(\ell_1\) loss, a.k.a. the least absolute deviation (LAD), are known to yield sparse solutions [3], [4] and outlier-robust estimates [5], [6], respectively, as opposed to the squared \(\ell_2\) norm which has widely been used for the Tikhonov regularization or the squared errors. The \(\ell_1\) norm is a convex relaxation of the \(\ell_0\) pseudo-norm (which is a direct discrete measure of sparsity counting the number of nonzero entries); i.e., the \(\ell_1\) norm is the largest convex minorant of \(\ell_0\) in a vicinity of the origin. For better relaxations/approximations to ameliorate the performance, a plethora of nonconvex alternatives to the \(\ell_1\) norm have been proposed [7], [8], [9], [10], [11], including the \(\ell_p\) quasi-norm for \(p \in (0, 1)\) (e.g., [12], [13], [14] among many others), capped \(\ell_1\) [15], log-sum function [16], minimax concave (MC) [17], smoothly clipped absolute deviation (SCAD) [18], continuous exact \(\ell_0\) (CEL0) [19], to name a few. See also the survey paper [20] for more references. Among those penalties, MC and SCAD are well known to be weakly convex; i.e., those functions become convex by adding a scaled squared \(\ell_2\) norm.

The notion of “convexity-preserving” nonconvex penalties using weakly convex functions can be found in the literature [21], [22]. The idea is to preserve overall convexity of the whole objective function by exploiting strong convexity of the other term(s); cf. difference of convex (DC) programming [23]. See, e.g., [24], [25] for more recent advances. In addition that the
weakly convex penalties induce sparsity with small estimation biases, the optimization problems involving a quadratic function and such weakly convex penalties can be solved by powerful convex-analytic algorithms with convergence guarantee to a global minimizer. It is widely known that the $\ell_p$ quasi-norm resides between the $\ell_0$ and $\ell_1$ norms. It has been shown recently that the (properly-normalized) MC penalty bridges the $\ell_0$ and $\ell_1$ norms by a single parameter [26, Example 2]. This, together with its nice experimental performances, motivates us to focus on the MC penalty. Let us consider the squared-error fidelity (the least square loss) penalized by a weakly convex function in linear regression. The overall convexity can be preserved in the overdetermined case by choosing the regularization parameter properly. In some important applications including high dimensional data analysis and compressed sensing, however, the overall convexity cannot be preserved because the number of measurements is much smaller than the number of variables.

To overcome this strict limitation, the generalized MC (GMC) penalty has been proposed [27]. Based on the fact that the MC penalty can be expressed as a difference between the $\ell_1$ norm and its Moreau envelope, GMC inserts a matrix-valued tuning parameter in the quadratic term of the Moreau envelope. We refer to the $\ell_1$ norm as “the seed function” of the GMC penalty. The GMC penalty has been extended (i) from $\ell_1$ norm to a more general convex seed-function satisfying certain mild conditions and (ii) to a composition of linear operator [26], [28]. The extended function is called linearly involved generalized Moreau enhanced (LiGME) penalty [26], covering the Moreau enhanced penalties for the nuclear norm and total variation among many others. The important property common to those generalized penalties is nonseparability even if its seed function is additively separable; i.e., those penalties are not necessarily expressed as a sum of individual functions each of which depends solely on each variable. Thanks to its nonseparability, GMC/LiGME can be applied to underdetermined linear systems. While it has rigorous theoretical backbones, its use in robust regression has not been investigated so far. Although a number of nonconvex loss functions have been proposed [29], [30], [31], [32], [33], [34], [35] indeed as alternatives to the convex ones such as LAD or Huber’s loss [5], [6], global optimality has not been discussed in those previous works.

B. Contributions — Part I

There are three research questions that motivate the present study, two of which are stated in this part.

(Q1) What is a function that is maximally close to the MC penalty while being able to possess overall convexity in underdetermined situations?

We would like to reserve such a region, as much as possible, on which the newly developing penalty coincides with the MC penalty. In the underdetermined case, the fidelity function is not strongly convex in the whole space. Precisely, while it is strongly convex on the subspace spanned by the input vectors, it is “flat” (it has no strong convexity at all) on its orthogonal complement. This immediately implies that the penalty function needs to be convex on the orthogonal complement to preserve the overall convexity. This simple observation is the key for our first contributions summarized below.

- We propose the projective minimax concave (PMC) penalty in which the projection operator onto the input subspace is used to annihilate the Moreau enhancement effects on its orthogonal complement. PMC reduces to the original MC penalty on the input subspace while it reduces to the $\ell_1$ norm (a convex relaxation of the $\ell_0$ pseudo-norm) on its orthogonal complement (see Proposition 1). This means that PMC gives an answer to the first question shown above.
- The formulation involving PMC enhances sparsity with small estimation biases in the underdetermined case, and thus it is referred to as debiased sparse modeling.
- While the PMC penalty itself is “additively nonseparable,” the “internal” objective function to define the Moreau envelope is “separable” as long as the seed function is separable. This mixed nature of separability and nonseparability allows PMC to preserve overall convexity in the underdetermined case with its efficient implementation using no extra variable.

(Q2) Can we build a regression method that is highly robust against huge outliers and stable even in severely noisy environments?

- We propose stable outlier-robust regression (SORR) under the assumption that the noise is Gaussian and the outlier is sparse. An additional variable vector is introduced to model the Gaussian noise on top of the adoption of the MC-based fidelity function to evaluate the sparse outliers, thereby reflecting the noise Gaussianity and the outlier sparsity in a reasonable way.
- SORR is a promising approach because (a) it is highly robust and stable even in severely noisy environments, and (b) it can be implemented efficiently by the operator splitting methods since overall convexity of the whole cost is preserved under a certain condition. This indicates that SORR resolves a certain intrinsic tradeoff existing in the conventional approaches (see Section II-B1).

C. Contributions — Part II

The two methods proposed in the first part are based on weakly convex functions. This gives rise to the third question.

(Q3) Can we build a mathematical modeling framework to treat weakly convex functions in a unified fashion for regression/classification tasks such as those studied in the first part?

- We propose the LiMES model which encompasses the debiased sparse modeling and SORR as its particular examples. The other examples of LiMES presented in this paper are stable principal component pursuit (SPCP) [36] and robust classification. For the latter application, in particular, the popular hinge loss is enhanced by the LiMES model with its expression as a composition of the support function of a closed interval $[-1, 0]$ and some affine operator.
- A necessary and sufficient condition for the smooth part of the whole cost to be convex is presented under a certain assumption (Proposition 5).
The structure of LiMES admits its decomposition into a sum of smooth and nonsmooth (proximable) convex functions, allowing an application of the efficient operator splitting methods to solve the posed problem. The gradient of the smooth part produces an implicit proximity operator, which contributes to reducing the estimation bias caused by the proximity operator appearing explicitly in the original form of the optimization algorithm.

Numerical examples show the efficacy of the LiMES framework. Specifically, the PMC penalty achieves debiased sparse modeling for underdetermined systems as well as outperforming GMC, and SORR\(^1\) achieves stable and remarkably robust performances in the presence of both heavy Gaussian noise and sparse outlier as well as outperforming the existing robust methods.

D. Notation and Mathematical Tools

Let \(\mathbb{R}, \mathbb{R}_+, \) and \(\mathbb{N}\) denote the sets of real numbers, strictly positive real numbers, and nonnegative integers, respectively. Let \((\mathcal{H}, \langle \cdot, \cdot \rangle)\) be a real Hilbert space equipped with inner product \(\langle \cdot, \cdot \rangle\), of which the induced norm is denoted by \(\| \cdot \|\).

Throughout the paper, we focus on the finite dimensional case, although many of the arguments given in this section apply to the infinite dimensional case. We denote by \(I : \mathcal{H} \to \mathcal{H}\) the identity operator, and by \(0 \in \mathcal{H}\) and \(O : \mathcal{H} \to \mathcal{H} : x \mapsto 0\) the zero vector of \(\mathcal{H}\) and the zero operator, respectively. We may use the same notation of inner product, norm, vector, and zero operator for other Hilbert spaces, whenever it causes no confusion. A subset \(C \subset \mathcal{H}\) is convex if \(ax + (1 - a)y \in C\) for all \((x, y, a) \in C \times C \times [0, 1]\). Given a nonempty closed convex set \(C \subset \mathcal{H}\), the projection operator is defined by \(P_C : \mathcal{H} \to \mathcal{H} : x \mapsto \text{argmin}_{y \in C} \| x - y \|\).

An operator \(T : \mathcal{H} \to \mathcal{H}\) is Lipschitz continuous with constant \(L \in \mathbb{R}_+^+\) if \(\| T(x) - T(y) \| \leq L \| x - y \|\) for every \(x, y \in \mathcal{H}\). The projection operator \(P_C\) is Lipschitz continuous with constant 1 (i.e., nonexpansive).

A function \(f : \mathcal{H} \to (-\infty, +\infty) := \mathbb{R} \cup \{+\infty\}\) is convex on \(\mathcal{H}\) if \(f(ax + (1 - a)y) \leq a f(x) + (1 - a) f(y)\) for all \((x, y, a) \in \text{dom } f \times \text{dom } f \times [0, 1]\), where \(\text{dom } f := \{ x \in \mathcal{H} : f(x) < +\infty \}\). If in addition \(\text{dom } f \neq \emptyset\), \(f\) is a proper convex function. For \(\eta \in \mathbb{R}_+^+\), \(f\) is \(\eta\)-strongly convex if \(f - 0.5\eta\| x \|^2\) is convex, and it is \(\eta\)-weakly convex if \(f + 0.5\eta\| x \|^2\) is convex. A convex function \(f : \mathcal{H} \to (-\infty, +\infty)\) is lower semi-continuous (or closed) on \(\mathcal{H}\) if the level set \(\text{lev}_{\leq a} f := \{ x \in \mathcal{H} : f(x) \leq a \}\) is closed for every \(a \in \mathbb{R}\). The set of all proper lower-semicontinuous convex functions defined over \(\mathcal{H}\) is denoted by \(\Gamma_0(\mathcal{H})\). Given a function \(f \in \Gamma_0(\mathcal{H})\), the Fenchel conjugate of \(f\) is \(\Gamma^*(\mathcal{H}) \ni f^* : x \mapsto \sup_{y \in \mathcal{H}} \langle x, y \rangle - f(y)\). The Moreau envelope (smooth convex approximation) of \(f\) of index \(\gamma \in \mathbb{R}_+^+\) is defined by \([38],[39],[40]\)

\[
\gamma f : \mathcal{H} \to \mathbb{R} : x \mapsto \min_{y \in \mathcal{H}} \left( f(y) + 0.5\gamma^{-1} \| x - y \|^2 \right)
\]

which is the proximity operator of \(f\) of index \(\gamma\). The gradient of the Moreau envelope \(\nabla \gamma f\) is given by \([38],[39],[40],[41]\)

\[
\nabla \gamma f = \gamma^{-1} (I - \text{Prox}_{\gamma f}) ,
\]

where \(\text{Prox}_{\gamma f} : \mathcal{H} \to \mathcal{H} : x \mapsto \text{argmin}_{y \in \mathcal{H}} \left( f(y) + 0.5\gamma^{-1} \| y - x \|^2 \right)\),

\[
(2)
\]

is the proximity operator of \(f\) of index \(\gamma\). The gradient of the Moreau envelope \(\nabla \gamma f = \gamma^{-1} (I - \text{Prox}_{\gamma f})\), which is Lipschitz continuous with constant \(\gamma^{-1}\). The following identity holds in general \([42], \text{Theorem 14.3}\):

\[
\gamma f + \frac{1}{\gamma} (f^*) \circ \gamma^{-1} I = 0.5\gamma^{-1} \| \cdot \|^2 .
\]

(3)

For any \(n, m \in \mathbb{N}^* := \mathbb{N} \setminus \{0\}\), the \(n \times n\) identity and zero matrices are denoted by \(I_n\) and \(O_n\), respectively, and the \(n \times m\) zero matrix is denoted by \(O_{n \times m}\). Matrix transpose is denoted by \(\langle \cdot \rangle^\top\). The \(\ell_1\) and \(\ell_2\) norms of Euclidean vectors \(x := \left[ x_1, x_2, \ldots, x_n \right]^\top \in \mathbb{R}^n\) are defined respectively by \(\| x \|_1 := \sum_{i=1}^n |x_i|\) and \(\| x \|_2 := \left( \sum_{i=1}^n x_i^2 \right)^{1/2}\).

II. TWO NOVEL FORMULATIONS FOR LINEAR REGRESSION

Two specific situations in linear regression are considered. We first present the PMC penalty to obtain debiased estimates for sparse modeling under possibly underdetermined systems. We then present SORR to combat the noise and outlier in a separate fashion. Given a coordinate system, a function is said to be “additively separable” when it is a superposition of individual functions of each parameter.\(^2\) The \(\ell_1\) norm is a simple example of separable functions.

A. PMC Penalty for Debiased Sparse Modeling

1) Sparse Modeling: We consider sparse modeling under the standard linear model \(y := Ax + \varepsilon\). Here, \(x \in \mathbb{R}^m\) is the sparse unknown vector to be estimated, \(\varepsilon \in \mathbb{R}^m\) is the Gaussian noise vector, and \(A := \left[ a_1, a_2, \ldots, a_m \right]^\top \in \mathbb{R}^{m \times n}\) and \(y := \left[ y_1, y_2, \ldots, y_m \right]^\top \in \mathbb{R}^m\) are the input matrix and the output vector, respectively, with the \(i\)th input vector \(a_i \in \mathbb{R}^n\), \(i = 1, 2, \ldots, m\), and its corresponding output \(y_i \in \mathbb{R}\). The task is the following: find the sparse vector \(x \in \mathbb{R}^m\) given \(A\) and \(y\). The linear system is supposed to be possibly underdetermined; i.e., \(A^\top A \in \mathbb{R}^{n \times n}\) might be singular.

2) The PMC Penalty: To reduce the estimation bias while preserving the overall convexity, we propose the following formulation (which we refer to as debiased sparse modeling\(^3\)):

\[
\min_{x \in \mathbb{R}^n} \left( 0.5 \| Ax - y \|^2_2 + \mu \left[ \| x \|_1 - \gamma \| 1 \| (P_M x) \right] \right) ,
\]

where \(P_M = A^\top A \in \mathbb{R}^{n \times n}\) is the orthogonal projection operator onto \(M := \text{null}(A) = \{ \text{range} A^\top \} \subset \mathbb{R}^n\), \(\mu \in \mathbb{R}_+^+\) is the regularization parameter, and

\[
\Phi_{\gamma \text{PMC}}(x) := \| x \|_1 - \gamma \| 1 \| (P_M x)
\]

\[
(5)
\]

\(^1\)Partial results (the SORR estimator and a special case of the LiMES model) of this work have been presented at a conference \([37]\) with no detailed discussions nor proofs for theoretical results.

\(^2\)Additive separability depends on the coordinate system.

\(^3\)It differs from debiased lasso estimator studied in statistics \([43]\) which “desparsifies” the estimate to reduce the estimation bias by adding a Newton step to the lasso estimate.
is the proposed PMC penalty. Here, $(\cdot)^{1}$ and $(\cdot)^{-1}$ denote the Moore-Penrose pseudoinverse and the orthogonal complement of subspace, respectively.

Using the identity $(3)$, the standard MC penalty $[17, [27]$ can be written as $\Phi_{MC}^\gamma(x) := \|x\|_1 - \gamma \cdot \|1(x)\|_1 = \|x\|_1 + \gamma \cdot (\|\cdot\|_\gamma - 0.5\gamma^{-1}\|x\|)^2$. Here, the subtraction of the Moreau envelope $\|\cdot\|_\gamma - 1(\|\cdot\|_1)$ from $\|x\|$ leads to nearly unbiased estimation $[17]$, and it hence enhances the performance significantly. As the conjugate function $\|\cdot\|_\gamma$ of $\|\cdot\|$ is convex, so is its Moreau envelope $\gamma^{-1}(\|\cdot\|_\gamma)$, and thus $\Phi_{MC}^\gamma(x)$ is $\gamma^{-1}$-weakly convex. The MC penalty cannot therefore be applied to the overdetermined case when $A^T A$ is singular, because $0.5\|A x - y\|_2^2 + \mu \Phi_{MC}^\gamma(x)$ cannot be convex for any $\mu \in \mathbb{R}_{++}$. Intuitively, the convexity of the fidelity term $0.5\|A x - y\|_2^2$ since the former function is flat (i.e., possesses zero curvature) over $M^\perp(= \text{null } A)$, or any of its translations. Here comes the idea of inserting $P_M$ into the penalty in $(4)$. The projection operator $P_M$ restricts the convexity to $M(= \text{null } A)$, on which the fidelity function is strongly convex, so that the overall convexity can be preserved. As a result, the Moreau enhancement effect is restricted to $M$ as well. A formal discussion about the convexity issue is postponed to Section II-A4. In the overdetermined case, PMC reduces to the standard MC penalty, as $M = \mathbb{R}^n$ and thus $P_M = I$.

We mention that the PMC penalty $\Phi_{MC}^\gamma$ depends on the input subspace $M$. This comes from a requirement for the preservation of overall convexity. This design strategy also has a more positive aspect in such specific situations when the desired solution belongs to a known input subspace at least with high probability (and possibly one is allowed to generate the input vectors so that it spans that particular subspace).

3) Properties of the PMC Penalty: Some properties of PMC are given below.

**Remark 1 (Separability and nonseparability):** The PMC penalty $\Phi_{PMC}^\gamma$ in $(4)$ is “additively nonseparable” as a function of $x$ with respect to the Cartesian coordinate system (i.e., PMC is not represented as a sum of individual functions of each component of $x$), unless $P_M$ is a diagonal matrix. In contrast, the second term of $\Phi_{PMC}^\gamma$ is given by $\gamma\|\cdot\|_1(P_M x) = \min_{u \in \mathbb{R}^n}\|u\|_1 + 0.5\gamma^{-1}\|P_M x - u\|_2^2 = \min_{u_1, u_2, \ldots, u_n \in \mathbb{R}} \sum_{i=1}^n \phi_i(u_i)$, in which the objective function is “separable” as a function of $u$. Here, $\phi_i(u_i) := \|u_i\| + 0.5\gamma^{-1}(p_i - u_i)^2$ with $p_i \in \mathbb{R}$ denoting the $i$th component of $P_M x$. This mixed nature of separability and nonseparability is crucial. It is known indeed that, to preserve the overall convexity when $A^T A$ is singular, a nonconvex penalty needs to be nonseparable, excluding a trivial case $[44]$. At the same time, thanks to the separability mentioned above, the minimizer of the objective function $\|\cdot\|_1 + 0.5\gamma^{-1}\|P_M x - u\|_2^2$ is given simply by $\text{soft}_\gamma(P_M x)$. Since $\|\cdot\|_1(P_M x)$ is merely the composite of the linear operator $P_M$ and the Moreau envelope of the $\ell_1$ norm, an application of the chain rule with $P_M = P_M$ gives the gradient $\nabla(\|\cdot\|_1 \circ P_M)(x) = \gamma^{-1} P_M \circ (I - \text{Prox}_{\gamma^{-1}/2}) \circ P_M(x)$, where the gradient operator $\nabla(\|\cdot\|_1 \circ P_M)$ is Lipschitz continuous with constant $\gamma^{-1}$. This smoothness property simplifies the optimization procedure, as shown in Section II-A4.

**Proposition 1:** For the PMC penalty, the following hold:

a) The PMC penalty $\Phi_{PMC}^\gamma$ coincides with the MC penalty on the input subspace $M$; i.e., $\Phi_{PMC}^\gamma(M) = \|x\|_1 - \gamma \cdot \|1(x)\|_1$ for $x \in M$.

b) The PMC penalty $\Phi_{PMC}^\gamma$ is reduced to the $\ell_1$ norm on the orthogonal complement $M^\perp(= \text{null } A)$; i.e., $\Phi_{PMC}^\gamma(M) = \|x\|_1$ for $x \in M^\perp$.

**Proof:** Clear from $(4)$.

**Remark 2 (PMC penalty bridges $\ell_0$ and $\ell_1$ over $M$):** An important implication of Proposition 1(a) is that the PMC penalty gives a bridge by a single parameter $\gamma$ between the direct measure $\|\cdot\|_0$ of sparsity and its convex relaxation $\|\cdot\|_1$ on the subspace $M$. To be specific, we define $\Phi_{PMC}^\gamma := \theta_\gamma \Phi_{MC}^\gamma$, where $\theta_\gamma := \left\{ \begin{array}{ll} \frac{\gamma}{\gamma} & \text{if } \gamma \in (0, 2) \\ 1 & \text{if } \gamma \in [2, +\infty) \end{array} \right.$

Then, it follows, for any $x \in \mathbb{R}^n(> M)$, that $\lim_{\gamma \to 0^+} \Phi_{PMC}^\gamma(x) = \|x\|_0$ [26], and $\lim_{\gamma \to +\infty} \Phi_{PMC}^\gamma(x) = \|x\|_1$. Here, the latter argument can be justified by observing that $0 \leq \gamma \|1\|_1(x) = \min_{u \in \mathbb{R}^n}\|u\|_1 + 0.5\gamma^{-1}(0 - x_2)^2 \leq \|0\|_1 + 0.5\gamma^{-1}0 - x_2^2 \to 0$ as $\gamma \to +\infty$ for any $x \in \mathbb{R}^n$.

We emphasize that those remarkable properties given in Remarks 1 and 2 and Proposition 1 come from the “structure” of PMC (see Remark 1), not from the use of the projection operator. We mention that PMC is non-monotonic, and it is decreasing in some direction(s) so that it may take negative values. Although this may cause overestimation, PMC tends to perform better than GMC (which would suffer from underestimation owing to a shrinking bias to a certain extent), as shown by simulations in Section IV-A. In fact, all those properties make PMC be significantly different from GMC and its related works.

4) Iterative Shrinkage and Debiasing Algorithm: The problem in $(4)$ can be viewed as

$$\min_{x \in \mathbb{R}^n} 0.5\|Ax - y\|_2^2 - \mu \gamma \|\cdot\|_1(P_M x) + \mu \|x\|_1,$$

(6)

Since the gradient of the smooth part in $(6)$ and the proximity operator of the $\ell_1$ norm are available (see Remark 1), the proximal gradient method [45, 46] can be applied, under the convexity condition presented in Proposition 2 below, directly to $(6)$ to obtain the following algorithm. Given an initial point $x_0 \in \mathbb{R}^n$, generate a sequence $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^n$ by (cf. Section III-D)

$$x_{k+1} := \text{soft}_\beta \mu[x_k + \beta_k \gamma^{-1} P_M(x_k - \text{soft}_\gamma(P_M x_k))]
- \beta_k A^T (Ax_k - y), k \in \mathbb{N},$$

(7)

where $\beta_k \in (0, 2/(\lambda_{\max}(A^T A) + \mu \gamma^{-1}))$ is the step size, and, for any $\delta \in \mathbb{R}_{++}$, \text{soft}_\beta := \text{Prox}_{\beta \delta}(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$. $x := [x_1, x_2, \ldots, x_n]^\top \mapsto \{\varphi_1(x_1), \varphi_2(x_2), \ldots, \varphi_n(x_n)\}^\top$, $n \in \mathbb{N}^*$, is the shrinkage (soft thresholding) operator. Here, $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue, and $\varphi_\delta : \mathbb{R} \to \mathbb{R}$ : $a \mapsto \text{sign}(a) a \max\{0, |a| - \delta\}$, where $\text{sign}(a) := 1$ if $a \geq 0$; $\text{sign}(a) := -1$ otherwise.

**Remark 3:** The algorithm in $(7)$ involves no auxiliary vector thanks to the “separability” discussed in Remark 1. This is in contrast to the GMC-based formulation [27] for which auxiliary vectors (with a saddle-point problem considered) need to be
used, because the objective function of the minimization problem involved in the generalized Huber function (the generalized Moreau envelope) is typically “nonseparable”. This auxiliary-vector-free nature of PMC could potentially reduce the memory requirement with respect to that for the algorithm in [27] applied to the GMC-based formulation.

To understand the behaviour of the algorithm in (7) geometrically, let us first consider the case when it is applied to the original MC penalty; i.e., the case of $P_M = I$. In this case, the second term in the bracket of (7) reduces to $\beta_k \mu \gamma_1 (x_k - \soft\gamma(x_k))$, which actually reduces the shrinking bias caused by the shrinkage operator $\soft\beta_k \mu$ in the algorithm. Each nonzero component of $\beta_k \mu \gamma_1 (x_k - \soft\gamma(x_k))$ shares the same sign as the corresponding component of $x_k$. Hence, this term debiases the estimate by enhancing the magnitudes of the nonzero components prior to the operation of $\soft\beta_k \mu$, while maintaining zero components.

In the case of PMC, the projection $P_M$ restricts the “debiasing” effect (the Moreau enhancement effect) to the subspace $\mathcal{M}$. Here, this restriction is due to a requirement for ensuring convexity of the whole cost of (4). We therefore refer to the algorithm in (7) as the iterative shrinkage and debiasing algorithm (ISDA),\footnote{Although (7) can be regarded as a specific instance of the iterative shrinkage-thresholding algorithm (ISTA) [47], we call it ISDA due to its remarkable debiasing property. A stochastic version of ISDA has been presented in [48] with its geometric interpretation.} which converges to a minimizer of (4) provided that the smooth part in (6) is convex. The convexity condition is given below.

**Proposition 2 (Convexity condition for (4)):** The smooth part $0.5 \|Ax - y\|^2_2 - \gamma \|\theta\|^1 (P_M \theta)$ is convex if and only if $\mu \leq \gamma \lambda_{\min}^+ (A^\top A)$, where $\lambda_{\min}^+ (\cdot)$ denotes the smallest strictly-positive eigenvalue.

**Proof:** The proof is based on the results to be presented in Section III-C, and it is given in Appendix A.

B. SORR Estimator for Outlier-Robust Regression

Robust regression concerns the case when some components of $y$ contaminate outliers as follows [49]:
\begin{equation}
\phi := Ax + \varepsilon + \omega.
\end{equation}

Here, $x \in \mathbb{R}^m$ and $\varepsilon \in \mathbb{R}^m$ are the unknown and noise vectors which are mutually uncorrelated and both of which obey i.i.d. zero-mean normal distributions with variances $\sigma_x^2 \in \mathbb{R}_{++}$ and $\sigma_\varepsilon^2 \in \mathbb{R}_{++}$, respectively, and $\omega \in \mathbb{R}^m$ is the sparse outlier vector [5].

1) A Tradeoff Between Robustness and Mathematical Tractability, and Stability Aspect: Popular Huber’s loss [5] is more insensitive to outliers than the least square (LS) loss, and it is mathematically tractable owing to its convexity at the same time. Regarding stability with respect to fluctuations caused by Gaussian noise, the Huber’s and LS losses are equally stable. Nevertheless, the robustness of Huber’s loss against huge outliers is limited. This can be seen by inspecting its derivative to which the so-called influence function is proportional [50]. See Fig. 1. It can be seen that the derivative of Huber’s loss stays constant above (or below) the threshold. This means that large outliers give a constant amount of influence to the estimate, thus causing extra estimation bias.

In contrast, Tukey’s biweight loss [51] has a “redescending property (vanishing derivative)”, i.e., the derivative vanishes at some point on each side of the real line. This implies that such outliers that have magnitudes exceeding the threshold would give no influence to the estimate, thus leading to remarkable robustness to huge outliers. Unfortunately, however, Tukey’s biweight is mathematically intractable owing to its nonconvexity. So, how can we break the tradeoff between robustness against huge outliers and mathematical tractability?

To find an answer to this question, let us consider the following question first: can we find such a convex loss that has a vanishing derivative (for sufficiently large values)? This is hopeless actually in a certain sense. To be precise, we restrict ourselves to such a class of loss functions $\phi : \mathbb{R} \rightarrow [0, +\infty)$ such that (i) $\phi(0) = 0$, (ii) $\phi(\varepsilon) > 0$ if $\varepsilon \neq 0$, and (iii) $\phi$ is differentiable everywhere but the origin. (This assumption is reasonably mild. See, e.g., [52].) Within this class of functions, $\phi$ is continuous if it is convex, because range $\phi = [0, +\infty) \subset \mathbb{R}$ owing to conditions (i) and (iii). In fact, no convex loss has a vanishing derivative in this case. The derivative $\psi := \phi’$ of $\phi$ has the following properties: (i) $\psi(0) = 0$ if $\phi$ is also differentiable at $\varepsilon = 0$ which minimizes $\phi$ (or $0 \in \partial \phi(0)$ in general), and (ii) $\psi(\varepsilon) > 0$ when $\varepsilon$ increases from zero infinitesimally. Hence, there is no way for $\psi$ to vanish again because the derivative of a convex function is monotonically non-decreasing.

The above arguments encourage us to explore nonconvex loss functions. In fact, the MC loss $\Phi_{MC}$ has a vanishing derivative (as can be seen from Fig. 1), and it is mathematically tractable at the same time. Let us now highlight the behaviour of the derivative $\psi$ for each loss in the vicinity of the origin. It can be seen that the derivative vanishes at the origin for the Huber, Tukey’s biweight, and LS losses, while it does not vanish for the MC and LAD losses. This implies a direct use of the MC loss may cause instability with respect to small fluctuations generated by Gaussian noise. We therefore present another formulation using an additional variable vector to model the Gaussian noise vector $\varepsilon$, in the following.

2) Stable Outlier-Robust Regression: We introduce the variable vector $\varepsilon \in \mathbb{R}^m$ to model the noise $\varepsilon_o$. The SORR formulation is given as follows:
\begin{equation}
\min \left\{ \min_{x \in \mathbb{R}^m, \varepsilon \in \mathbb{R}^m} \mu \left[ \|y - (Ax + \varepsilon)\|_1 - \theta \|\theta\|_1 (y - (Ax + \varepsilon)) \right] \Phi_{MC}(y - (Ax + \varepsilon)) \right. \left.+ 0.5 \sigma_x^{-2} \|x\|_2^2 + 0.5 \sigma_\varepsilon^{-2} \|\varepsilon\|_2^2, \right. \end{equation}
where \( \sigma^2_x \in \mathbb{R}_{++} \) and \( \sigma^2_{\epsilon_x} \in \mathbb{R}_{++} \) are estimates of \( \sigma^2_x \) and \( \sigma^2_{\epsilon_x} \), respectively. If such estimates are unavailable, \( \sigma^2_x \) and \( \sigma^2_{\epsilon_x} \) are considered as tuning parameters. An extreme case of SORR with \( \sigma_x \downarrow 0 \) (or with a sufficiently small \( \sigma_x > 0 \)) makes the optimal \( \varepsilon \) of (8) be the zero vector, reducing SORR to the following simple formulation:

\[
\min_{x \in \mathbb{R}^n} \mu \left( \|Ax - y\|_1 - \gamma \|\cdot\|_1 (Ax - y) \right) + 0.5 \|x\|_2^2. \tag{9}
\]

We shall refer to the formulation in (9) as outlier-robust regression (ORR).\(^5\)

The first term \( \Phi^{MC}_x(y - (Ax + \varepsilon)) \) of (8) is the MC loss encouraging sparsity of the estimation residual \( y - (Ax + \varepsilon) \) which can be regarded as an estimate of the sparse outlier. The last two terms \( 0.5 \sigma^2_x \|x\|_2^2 \) and \( 0.5 \sigma^2_{\epsilon_x} \|\varepsilon\|_2^2 \) reflect the Gaussianity of \( x \) and \( \epsilon_x \), playing double roles of convexification and regularization (in the Tikhonov sense). In particular, when the noise power \( \sigma^2_{\epsilon_x} \) is large, \( \|\varepsilon\|_2^2 \) tends to be large as well. In this case, the inverse \( \sigma^2_x \) of the noise-power estimate would be small, and it allows \( \|\varepsilon\|_2^2 \) to be large so that \( \varepsilon \) mimics \( \epsilon_x \) well, yielding efficient mitigation of the MC loss \( \Phi^{MC}_x(y - (Ax + \varepsilon)) \). This leads to the “stability” of the SORR estimator in the spirit of [36]. A primal-dual splitting algorithm which can solve some class of linearly-involved nonsmooth convex optimization problems including (8) will be presented in Section III-D. The algorithm relies on convexity of the smooth part of the objective function in (8), for which the condition is given below.

Proposition 3 (Convexity condition for SORR (8)): The smooth part \( 0.5 \sigma^2_x \|x\|_2^2 + 0.5 \sigma^2_{\epsilon_x} \|\varepsilon\|_2^2 - \mu \|\cdot\|_1 (y - (Ax + \varepsilon)) \) is convex in \( (x, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}^m \) if and only if

\[
\mu (\sigma^2_x + \sigma^2_{\epsilon_x} \lambda_{\max} (A^* A)) \leq \gamma. \tag{10}
\]

Proof: The proof is based on the results to be presented in Section III-C, and it is given in Appendix B.

Remark 4 (SORR resolves the tradeoff efficiently): The SORR estimator breaks the tradeoff between robustness and mathematical tractability. In particular, SORR enjoys (i) remarkable robustness against huge outliers and (ii) insensitivity to small fluctuations, while the posed problem in (8) is still tractable because the whole cost is convex under (10). Those advantages come mainly from the use of the MC loss and the additional vector \( \varepsilon \).

We mention that the introduction of the additional vector \( \varepsilon \) does not increase computational complexity essentially, although a larger amount of memory is required than the case of ORR (the \( \varepsilon \)-parameter free formulation) to store the length- \( (n + m) \) vector \( \varepsilon \) as well as some other intermediate vectors that are required in the algorithm (see [37] or Section III-D2). Specifically, the algorithm iteration to compute the SORR estimator requires \( O(mnQ) \) complexity, in addition to the computation of the largest eigenvalue of \( A^* A \) (or \( A A^* \)) as preprocessing, where \( Q \) is the number of iterations.

The SORR formulation has been extended to sparse modeling under Gaussian and impulsive noises [54].

III. LiMES Model: Convexity Condition, Algorithm, and Applications

To give the convexity conditions for the debiased sparse modeling and SORR in a unified way, we present a generalized model called “LiMES” and show the necessary and sufficient condition for its convexity. Applications of the LiMES model can be classified into two categories: type-sparse and type-robust. We derive the proximal debiasing-gradient algorithm (which requires no auxiliary variable) for the former type and the primal-dual debiasing algorithm for the latter type. We finally give a couple of other applications than debiased sparse modeling and SORR.

Let \( \mathcal{L} : \mathcal{H} \to \mathcal{K} \) be a bounded linear operator from a Hilbert space \( \mathcal{H} \) to another Hilbert space \( \mathcal{K} \). The adjoint operator of \( \mathcal{L} \) is denoted by \( \mathcal{L}^* \). The operator norm is then defined by \( \|\mathcal{L}\| = \sup \{\|\mathcal{L}x\| : x \in \mathcal{H}, \|x\| \leq 1\} \). Given a bounded linear operator \( \mathcal{L} : \mathcal{H} \to \mathcal{L} \), \( \mathcal{L} \geq 0 \) means that \( \mathcal{L} \) is positive semidefinite, i.e., \( \langle \mathcal{L}x, x \rangle \geq 0 \) for all \( x \in \mathcal{H} \). Given any bijective bounded linear operator \( \mathcal{L} : \mathcal{H} \to \mathcal{K} \) and any function \( f \in \Gamma_0(\mathcal{H}) \), it holds that

\[
(f \circ \mathcal{L})^* = f^* \circ (\mathcal{L}^*)^{-1}. \tag{11}
\]

A. LiMES: A Class of Weakly Convex Functions

Definition 1 (The LiMES Model): Let \( \mathcal{X}, \mathcal{Y}, \) and \( \mathcal{Z} \) be finite-dimensional Hilbert spaces. Let \( \mathcal{A}_1 : \mathcal{X} \to \mathcal{Y} : x \mapsto M_1 x + c_1 \) and \( \mu \in \mathbb{R}_{++} \), where \( (\mathcal{O} \neq)M_1 : \mathcal{X} \to \mathcal{Y} \) is a bounded linear operator and \( c_1 \in \mathcal{Y} \) is a vector. Let \( (\mathcal{O} \neq)\mathcal{L} : \mathcal{Z} \to \mathcal{Z} \) be a bounded linear operator,\(^7\) \( D : \mathcal{Z} \to \mathcal{Z} \) be a diagonal positive-definite operator, and \( \mathcal{A}_2 : \mathcal{X} \to \mathcal{Z} : x \mapsto M_2 x + c_2 \), where \( (\mathcal{O} \neq)M_2 : \mathcal{X} \to \mathcal{Z} \) is a bounded linear operator and \( c_2 \in \mathcal{Z} \) is a vector. Let \( \Psi \in \Gamma_0(\mathcal{Z}) \), which is referred to as a seed function. The linearly-involved Moreau-enhanced-over-subspace (LiMES) model is defined as the minimization of the following function:

\[
J_{\Omega} : \mathcal{X} \to (-\infty, \infty] : x \mapsto 0.5 \|\mathcal{A}_1 x\|_2^2 + \mu \Psi^D_{\mathcal{L}} (\mathcal{A}_2 x), \tag{12}
\]

where \( \Omega := (\mathcal{A}_1 : \Psi^D_{\mathcal{L}} \circ \mathcal{A}_2) \), and

\[
\Psi^D_{\mathcal{L}} : \mathcal{Z} \to (-\infty, +\infty] : \|z\|_{\mathcal{L}} \mapsto \Psi(z) - \min_{v \in \mathcal{Z}} \left[ \Psi(v) + 0.5 \|D(\mathcal{L} z - v)\|_2^2 \right]. \tag{13}
\]

We refer to \( \Psi^D_{\mathcal{L}} \circ \mathcal{A}_2 : \mathcal{X} \to (-\infty, +\infty] \) as the LiMES function.

Define the subspace \( M_1 := \text{range } M_1 \). The debiased sparse modeling in (4) is reproduced by letting \( \mathcal{X} := \mathbb{R}^n, \mathcal{Y} := \mathbb{R}^m, \mathcal{A}_1 := A \cdot -y \) \( (M_1 : A) \), \( \Psi := \|\cdot\|_1 \), \( \mathcal{A}_2 := I_n \), \( \mathcal{L} := F_{M_1} = F_{M} = A^* A \in \mathbb{R}^{n \times n}, \) and \( D := \gamma^{-1/2} I_n \). On the other hand, ORR in (9) is reproduced by letting \( \mathcal{X} := \mathbb{R}^n, \mathcal{Z} := \mathbb{R}^m, \mathcal{A}_1 := I_n, \) \( \Psi := \|\cdot\|_1 \), \( \mathcal{A}_2 := A \cdot -y \), \( \mathcal{L} := I_m \), and \( D := \gamma^{-1/2} I_m \). In the former example, here, the quadratic term \( 0.5 \|\mathcal{A}_1 x\|_2^2 \) of (12) represents the data fidelity,

\(^5\)Unlike SORR, the ORR formulation in (9) does not distinguish the Gaussian noise \( \epsilon_x \) and the sparse outlier \( c_1 \) explicitly. ORR in (9) can also be viewed as a particular case of the model proposed in [53] for robust recovery of jointly sparse signals.

\(^7\)The letter \( \mathcal{L} \) will be used to denote the linear operator of LiMES, distinguished from the general linear operator \( \mathcal{L} \) (which was used to denote the linear operator of LiME in [26]).
TABLE I

| $\mathcal{L}$ | preserving the convexity of the smooth term (default: $P_{\text{range } M}$) |
| $D$ | assigning individual weights to the variables (Section III-F) |
| $\omega_1$ | used to define data fidelity for type-S applications |
| $\omega_2$ | used to define data fidelity for type-R applications |

TABLE II

| $\iota_C$ | the indicator function of $C$ |
| $\sigma_C$ | the support function of $C$ |
| $f$ | the Fréchet conjugate of $f$ |
| $\|\cdot\|_*$ | the dual norm of $\|\cdot\|$ |
| $\gamma_f$ | the Moreau envelope of $f$, e.g., $\gamma\|\cdot\|$ |
| $\text{Prox}_\Gamma$ | the proximity operator w.r.t. $\Gamma$ |
| $P_M$ | the projection operator onto subspace $M$ |
| $\Sigma^*$ | the adjoint of linear operator $\Sigma$ |
| $\Psi^M_\Gamma$ | the LiMES function ($\Psi^M_\Gamma = \Psi_P$); see (13) |
| $J_\Omega$ | the LiMES model for $\Omega := (\omega_1; \Psi^M_D \circ \omega_2)$; see (12) |

For consistent notation with [26], $\Psi^M_D := \Psi^M_{I_D}$ will be used when $\mathcal{L} := I$. The question now is: what is the role of the term $\min_{v \in Z} [\Psi(v) + 0.5 \|D(\mathcal{L} z - v)\|^2]$ in (13)? The following proposition, which generalizes Proposition 1, answers this question for the case of $\mathcal{L} := P_{M_1}$. Proposition 4:

(a) The particular LiMES function $\Psi^M_\Gamma$ coincides with the generalized Moreau enhanced penalty $\Psi^M_D$ on the subspace $M_1$; i.e., $\Psi^M_{D_1}(z) = \Psi^M_{I_D}(z) = \Psi^M_D(z) = \Psi^M_D(z) - \min_{v \in Z} [\Psi(v) + 0.5 \|D(\mathcal{L} z - v)\|^2]$ for $z \in M_1$.

(b) Let $\mathcal{L}$ satisfy $\mathcal{L} = \text{L} \circ P_{M_1}$. Then, $\Psi^M_D$ reduces to $\Psi$ (up to constant) on $M_1$; i.e., $\Psi^M_{D_1}(z) = \Psi(z) - \min_{v \in Z} [\Psi(v) + 0.5 \|Dv\|^2]$ for $z \in M_1$.

Proof: (a) The assertion can be verified by applying $P_{M_1}z = z$ for all $z \in M_1$ to (14) with $\mathcal{L} := P_{M_1}$.

(b) Use $\mathcal{L} = \mathcal{L} \circ P_{M_1}, z = \mathcal{L}0 = 0, \forall z \in M_1$ in (13).

Proposition 4 states, under the use of $\mathcal{L} := P_{M_1}$, that $\Psi^M_D$ is an “exact” Moreau-enhanced model over the subspace $M_1$ in the sense of the generalized Moreau enhanced (GME) penalty [26]. Note here that minimum $[\Psi(v) + 0.5 \|D(\mathcal{L} z - v)\|^2]$ can be regarded as a generalized Moreau envelope of $\Psi$. In all applications presented in this article, $\mathcal{L} := P_{M_1}$ will be used. Nevertheless, we would not exclude the possibility of using other choices of $\mathcal{L}$ such as those presented in [25, 26], although the Moreau enhancement over $M_1$ could be “inexact” in this case (see Example 1 in Section III-B). Proposition 4(b) states that $\Psi^M_D$ coincides with $\Psi$ over $M_1$ up to constant under the condition (which $\mathcal{L} := P_{M_1}$ satisfies).

As shown in Section II-A, the PMC penalty preserves the convexity of the smooth part even when $A^\top A$ is singular, and at the same time it enjoys the mixed nature of separability and nonseparability. Such a penalty can be generated systematically by the LiMES function with $\mathcal{L} := P_{M_1}$ given a separable function $\Psi$. We emphasize here that the diagonal $D$ induces the separability of the function $\Psi(v) + 0.5 \|D(\mathcal{L} z - v)\|^2$ in (13) in terms of $v$, which makes the gradient computation of the smooth part $F$ in (16) simple (see Remark 1). In particular, for typical type-S applications (such as debiased sparse modeling and SPCP to be presented in Section III-F), $\omega_2$ is also a diagonal operator, and thus the computationally efficient proximal gradient algorithm can be applied which requires no auxiliary vector to compute the LiMES model (see Section III-D).

To show an active role of the diagonal operator $D$ briefly, suppose that the variable vector $x \in X$ consists of several subvectors. In this case, $D$ can be used to give an individual weight to the regularizer of each subvector (see Section III-F).

B. Examples of LiMES Function: Penalty and Loss

In this subsection, we simply let $D := \gamma^{-1/2} I$ for $\gamma \in \mathbb{R}_{++}$, which reduces (14) to

$$\Psi^M_D(z) = \Psi_{z^{-1/2}I}(z) = \Psi(z) - \gamma \Psi(\mathcal{L} z).$$

(17)

Some examples of the LiMES function are listed below.
Example 1 (LiMES penalty): We let $X := \mathbb{R}^n$ and $Z := \mathbb{R}^m$ in (a)–(d) below.

a) (MC penalty [17, 27]) Let $\Psi := \| \cdot \|_1$, $\mathcal{L} := \Phi_\gamma$, $I_m (m = n)$, and $\theta_1 (\| \cdot \|_1, \gamma) := \Phi_{\gamma,MC}$. In particular, the MC penalty, or $\theta_1 (\| \cdot \|_1, \gamma) = \Phi_{\gamma,MC}$ more specifically, gives a parametric bridge between $\| \cdot \|_1$ and $\| \cdot \|_1$ and $\| \cdot \|_1$ and $\| \cdot \|_1$ (see Remark 2 for definition of $\theta_1$).

b) (PME and PMC) Let $\mathcal{L} := \mathcal{L}_m$ and $\mathcal{A} := I_m (m = n)$, where $\mathcal{L} = \mathbb{R}^n$ is a linear subspace of $\mathbb{R}^n$. Then, $\Phi_{\gamma,PMC} = \Psi := \gamma - \Psi \circ \mathcal{P}_{\mathcal{L}_m}$, which we call the Projective Moreau enhanced (PME) function. In particular, letting $\Psi := \| \cdot \|_1$ yields $\{ | \| \cdot \|_1, \gamma \} := 1 - \| \gamma \|_1 = 1$, $\| \cdot \|_1 = \Phi_{\gamma,PMC}$, which is the PMC penalty presented in Section II-A. An alternative choice of $\mathcal{L}$ to the $\mathcal{L}_m$ used in $\Phi_{\gamma,PMC}$ is given by $\mathcal{L} := \sqrt{\gamma/n} \Delta \log (a_1, 1/2, a_2, 1/2, \ldots, m_1, n)$, shown (cf. [25]), where $a_i \in [0, 1], i = 1, 2, \ldots, n$, are tuning parameters, and $A := V \Delta \gamma/dt$ is an eigenvalue decomposition with some orthogonal matrix $V \in \mathbb{R}^{n \times n}$ and some diagonal matrix $\sigma_1 \Delta \gamma \succeq 0$. (This choice actually satisfies the convexity condition to be presented in Section III-C.)

c) (MC-W) Let $\Psi := \| \cdot \|_1$, $\mathcal{L} := I_m (m = n)$, and $\mathcal{A} := \mathbb{R}^m$, where $\mathbb{R}^m$ is the popular wavelet transform [58]. Then, $\{ | \| \cdot \|_1, \gamma \} := \Phi_{\gamma,WM} = \Psi$.

d) (MC-TV) Let $\Psi := \| \cdot \|_1$, $\mathcal{L} := I_{m-1} (m = n-1)$, and $\mathcal{A} := \mathcal{D}_0 := \{ 0_1 - 1_1, -1_1, 0_1 \} (n = 1)$. Then, $\{ | \| \cdot \|_1, \gamma \} := \Phi_{\gamma,TV} = \mathcal{L}_1 (A - y)$ gives a parametric bridge between the rank of matrix and $\| \cdot \|_1$ [26].

e) (MEN) Let $\mathcal{L} := \mathbb{R}^{n \times m}$ and $\Psi := \| \cdot \|_1$. Then, $\{ | \| \cdot \|_1, \gamma \} := \Phi_{\gamma,MC} = \mathcal{L}_1 (A - y)$ gives an MEN loss, which has been studied in Section II-B for robust classification.

Example 2 (LiMES loss): We let $X := \mathbb{R}^n$ and $Z := \mathbb{R}^m$ in (a)–(c) below, and $A \in \mathbb{R}^{n \times m}$ and $y \in \mathbb{R}^m$ in (a), (c), and (d).

a) (MC loss) Let $\Psi := \| \cdot \|_1$, $\mathcal{L} := I_m$, and $\mathcal{A} := \mathbb{R}^m : z \mapsto Ax - y$. Then, $\{ | \| \cdot \|_1, \gamma \} := \Phi_{\gamma,MC} (A - y)$ gives an MC loss, which has been studied in Section II-B for robust classification.

b) (ME-hinge loss) Let $\Psi := \sigma_{[1,0]} : \mathbb{R} := \mathbb{R} \mapsto \sup_{x \in [1,0]} e^x = \max \{ 0, -z \}$, $\mathcal{L} := I_m = 1 (m = m)$, and $\mathcal{A} := \mathbb{R}^m : z \mapsto a^T x$ for some given $a \in \mathbb{R}^m$ such that $\| a \|_1 = 1$. Then, $\{ | \| \cdot \|_1, \gamma \} := \Phi_{\gamma,ME} = \mathcal{L}_1 (A - y)$ is the hinge loss function, and we call $\{ | \| \cdot \|_1, \gamma \} := \Phi_{\gamma,ME} (A - y)$ the Moreau-enhanced hinge (ME-hinge) loss function. See Proposition 7 for the second equality here. The proximity operator of $\Phi_{\gamma,ME}$ is given for instance in [42, Example 24.37]. The ME-hinge loss will be used in Section III-G for robust classification.

c) (MC-W loss) Let $\Psi := \| \cdot \|_1$, $\mathcal{L} := I_m$, and $\mathcal{A} := \mathbb{R}^m : z \mapsto \mathbb{R}^m : x \mapsto (Ax - y)$ gives an MC-W loss.
\[ \Leftrightarrow (\sigma^*_\gamma \circ D)(D\mathcal{L}\mathcal{A}_2x) = 0 \]
\[ \Leftrightarrow x \in K_C := \{ \hat{x} \in X \mid D^2\mathcal{L}\mathcal{A}_2\hat{x} \in C \}. \]  
(19)

ii) Assume that
\[ \text{int } K_C \neq \emptyset, \]  
(20)

where \( \text{int } K_C \) is the interior of \( K_C \). Then, \( F \in \Gamma_0(X) \) if and only if (\( \bigcirc \)) is satisfied.

Proof: (a) It is clear under (\( \bigcirc \)) that \( 0.5\|\mathcal{A}_1\|_2^2 - 0.5\mu\|D^2\mathcal{L}\mathcal{A}_2\|_2^2 \in \Gamma_0(X) \). It can also be verified that \( \Psi \in \Gamma_0(X) \Rightarrow \Psi^* \in \Gamma_0(X) \Rightarrow (\Psi^* \circ D) \circ D^2\mathcal{L}\mathcal{A}_2 \in \Gamma_0(X) \).

(b) For \( v \in Z \), it can be verified that
\[ 1(\sigma^*_\gamma \circ D)(D\mathcal{L}\mathcal{A}_2x) = \min_{z \in Z} \left[ \mathcal{C}(Dz) + 0.5\|v - z\|_2^2 \right] \]
\[ = \min_{z \in D^{-1}(C)} 0.5\|v - z\|_2^2 =: 0.5d^2(v, D^{-1}(C)). \]

It follows thus that \( 1(\sigma^*_\gamma \circ D)(D\mathcal{L}\mathcal{A}_2x) = 0.5d^2(D\mathcal{L}\mathcal{A}_2, D^{-1}(C)) \), and using this equality in (18) verifies that \( F(x) = 0.5\|\mathcal{A}_1\|_2^2 - 0.5\mu\|D^2\mathcal{L}\mathcal{A}_2\|_2^2 \Leftrightarrow 1(\sigma^*_\gamma \circ D)(D\mathcal{L}\mathcal{A}_2x) = 0 \Leftrightarrow D\mathcal{L}\mathcal{A}_2x \in D^{-1}(C) \Leftrightarrow D^2\mathcal{L}\mathcal{A}_2x \in C \Leftrightarrow x \in K_C \).

(b.ii) Since the third term of (18) vanishes over int \( K_C \neq \emptyset \) by Proposition 5(b.i), \( F \) is nonconvex if condition (\( \bigcirc \)) is unsatisfied. This implies the necessity of (\( \bigcirc \)). The sufficiency is verified already in Proposition 5(a).

Proposition 5 shows the situation under which (\( \bigcirc \)) is a necessary and sufficient condition. The necessity implies that the condition cannot be weaker, or, in other words, the parameter \( \mu \) cannot exceed the upper bound obtained from (\( \bigcirc \)). We remark that the diagonality and positive definiteness imposed implicitly on \( D \) in Proposition 5 can be relaxed straightforwardly by solely imposing bijectivity.

Lemma 1: Let \( \mathcal{L} : X \rightarrow Z \) be a bounded linear operator. Given a nonempty set \( \emptyset \neq C \subset Z \) and a point \( \hat{x} \in X \), it holds that \( \mathcal{L}\hat{x} \in \text{int } C \) implies \( \hat{x} \in \text{int } \mathcal{L}^{-1}(C) \). If \( \mathcal{L} \) is surjective, \( \mathcal{L}\hat{x} \in C \Leftrightarrow \hat{x} \in \text{int } \mathcal{L}^{-1}(C) \).

Proof: We denote by \( B(x, \epsilon) := \{ u \in X \mid \|u - x\| < \epsilon \} \) an open ball centered at \( x \in X \) with radius \( \epsilon \in \mathbb{R}_+ \). Assume that \( \mathcal{L}\hat{x} \in C \). Then, there exists some \( \epsilon \in \mathbb{R}_+ \) such that \( B(\mathcal{L}\hat{x}, \epsilon) \subset C \). It can then be shown straightforwardly that \( \mathcal{L}(B(\hat{x}, \epsilon)) \subset \mathcal{L}^{-1}(C) \), and hence \( \hat{x} \in \text{int } \mathcal{L}^{-1}(C) \). The converse implication in the equivalence part is an implication of the well-known open mapping theorem [6]. To see this, assume that \( \mathcal{L} \) is surjective and that \( \hat{x} \in \text{int } \mathcal{L}^{-1}(C) \). Then, there exists an \( \epsilon \in \mathbb{R}_+ \) such that \( B(\hat{x}, \epsilon) \subset \mathcal{L}^{-1}(C) \), and the image \( \mathcal{L}(B(\hat{x}, \epsilon)) \) is an open set due to the open mapping theorem. The inclusion \( \mathcal{L}\hat{x} \in \mathcal{L}(B(\hat{x}, \epsilon)) \subset C \) due to definition of inverse mapping thus implies \( \hat{x} \in \text{int } \mathcal{L}(B(\hat{x}, \epsilon)) \subset C \).

The following lemma gives a way of checking the nonemptiness condition of int \( K_C \) for necessity in Proposition 5.

\textbf{D. Proximal Debiassing Algorithms}

We present iterative algorithms using the proximity operator to compute the LiMES model for the case of \( D := \gamma^{-1/2}I \) for simplicity, which covers many applications including the debiased sparse modeling (4), SORK (8), ORR (9), and robust classification (31) (see Section III-G). (An extension to a general diagonal positive-definite operator \( D \) is straightforward.) In this case, (16) reduces to
\[ J_{\Omega_\gamma} = 0.5\|\mathcal{A}_1\|^2 - \mu \gamma \Psi \circ \mathcal{L} \mathcal{A}_2 + \mu \Psi \circ \mathcal{A}_2 \]  
(21)
where \( \Omega_\gamma := (\mathcal{A}_1^\gamma \circ (\mathcal{L}^\gamma \mathcal{A}_2)^\gamma \circ \mathcal{A}_2) \). Here, the gradients of \( 0.5\|\mathcal{A}_1\|^2 \) and \( \gamma \Psi \circ \mathcal{L} \mathcal{A}_2 \) at \( x \in X \) are given, respectively, by \( \nabla(0.5\|\mathcal{A}_1\|^2)(x) = M_1^\gamma \mathcal{A}_1 \) and (see Section I-D)
\[ \nabla(\gamma \Psi \circ \mathcal{L} \mathcal{A}_2)(x) = M_2^\gamma \nabla \gamma \Psi(\mathcal{L} \mathcal{A}_2 x) \]
\[ = \gamma^{-1} M_2^\gamma \mathcal{L}^\gamma (I - \text{Prox}_{\gamma \Psi}(\mathcal{L} \mathcal{A}_2 x)). \]
\[ = \gamma^{-1} M_2^\gamma \mathcal{L}^\gamma (I - \text{Prox}_{\gamma \Psi}(\mathcal{L} \mathcal{A}_2 x)). \]
\[ \gamma^{-1} \|L\|^2 M_2^2, \]

Both gradient operators \( \nabla(0.5\|\mathcal{A}_1\|^2)(x) \) and \( \nabla(\gamma \Psi \circ \mathcal{L} \mathcal{A}_2) \) are Lipschitz continuous with constants \( \|M_1\|^2 \) and \( \gamma^{-1} \|L\|^2 M_2^2 \), respectively.

1) Proximal Debiassing-Gradient Algorithm for Typical Type-S Applications: Let \( \mathcal{A}_2 := I \) which is used in typical type-S applications. This allows to use an efficient algorithm requiring no auxiliary variable. Specifically, under condition (\( \bigcirc \)), (21) can be minimized by the proximal gradient method:
\[ x_{k+1} := \text{Prox}_{\beta_k \gamma \Psi}(x_k - \beta_k (M_1^\gamma \mathcal{A}_1 \mathcal{k} - \mu \gamma^{-1} \mathcal{L}^\gamma (I - \text{Prox}_{\gamma \Psi}(\mathcal{L} \mathcal{A}_2 x)) \), \( k \in \mathbb{N} \).
\[ \text{where } \beta_k \in (0, 2/\|M_1\|^2 + \mu \gamma^{-1} \|L\|^2). \] ISDA presented in Section II-A4 is reproduced by letting \( \Psi := \| \cdot \|_1 \) and \( \mathcal{L} := \| \cdot \|_1 \).
$P_M$ in (23), which makes $\text{Prox}_{\Psi\mu} \| \cdot \|_1 = \text{soft}_\delta$ for any $\delta \in \mathbb{R}_{++}$. The gradient term $\mu L \nabla \gamma \Psi(L \delta x_k)$ actually plays the same role as the “debiasing” term of ISDA. We therefore refer to the algorithm as the proximal debiasing-gradient algorithm.

2) Primal-Dual Debiasing Algorithm for Type-R Applications: Let $\Psi(z) := \mu \Psi(z + c_2), z \in \mathbb{Z}$, so that $\Psi(M_2 x) = \mu \Psi(\delta_2 x)$. The problem in (21) can then be rewritten as

$$\min_{x \in \mathbb{R}} 0.5 \| \delta_2 x \|^2 - \mu \gamma \Psi(L \delta_2 x) + \tilde{\Psi}(M_2 x).$$

(24)

Since $\text{Prox}_{\Psi/\sigma}(z) = -c_2 + \text{Prox}_{\Psi/\sigma}(z + c_2)$ for $\sigma \in \mathbb{R}_{++}$, it follows that

$$\text{Prox}_{\Psi/\sigma}(z) = z - \sigma \text{Prox}_{\Psi/\sigma}(\sigma^{-1} z)$$

$$= z + \sigma c_2 - \sigma \text{Prox}_{\Psi/\sigma}(\sigma^{-1} z + c_2),$$

(25)

where the first equality is due to the well-known identity [42, Theorem 14.3]: $\text{Prox}_{f + g} = \text{Prox}_f \circ \text{Prox}_g \circ \gamma^{-1} I = I$ for any $f \in \Gamma_0(\mathbb{Z})$ and $\gamma \in \mathbb{R}_{++}$. Problem (24) can be solved by the existing operator splitting methods such as the forward-backward-based primal-dual method [63, 64, 65]; see Algorithm 1 below.11

Algorithm 1: (Primal-Dual Debiasing Algorithm).

Set: $x_0 \in \mathcal{X}, y_0 \in \mathbb{Z}, (\tau, \sigma) \in \mathbb{R}_{++}^{2n}, \beta_k \in \mathbb{R}_{++}$
For $k = 0, 1, 2, \ldots$, do:

$s_k = x_k - \tau M_2 \delta_2 x_k - \mu \gamma^{-1} M_2 L^\top (I - \text{Prox}_{\gamma \Psi}(L \delta_2 x_k))$

$u_k = s_k - \tau M_2 \delta_2 u_k$

$q_k = \text{Prox}_{\Psi/\sigma}(v_k + \sigma M_2 u_k)$

$p_k = s_k - \tau M_2 \delta_2 q_k$

$(x_{k+1}, v_{k+1}) = (x_k, v_k) + \beta_k (p_k, q_k) - (x_k, v_k)$

Convergence condition of Algorithm 1: (i) $\tau \sigma \| M_2 \|^2 \in (0, 1)$
and $\tau \in (0, 2/(\| M_2 \| + \mu \gamma^{-1} \| L \delta_2 \|^2))$, (ii) $\beta_k \in \mathbb{R}_{++}$
and $\text{inf}_{k \in \mathbb{N}} \beta_k \in \mathbb{R}_{++}$, (iii) the function $J_{\Omega, \gamma}$ in (12) has a minimizer, and (iv) $\text{int}(\text{dom} \Psi) \cap \text{range} M_2 \neq \emptyset$.

E. Stable Outlier-Robust Regression as a Special Case of LiMES Model

We consider a general situation when the augmented vector $\xi := [x^\top, \varepsilon^\top]^\top \in \mathbb{R}^{n+m}$ obeys a zero-mean normal distribution with its (nonsingular) covariance matrix $\Sigma_\xi \in \mathbb{R}^{(n+m) \times (n+m)}$. In this case, the standard statistical argument may suggest the use of $0.5 \Sigma_\xi^{-1/2} \| \cdot \|_2^2$, where $\xi := [x^\top, \varepsilon^\top]^\top \in \mathbb{R}^{n+m}$ and $\Sigma_\xi$ is an estimate of $\Sigma_\xi$. The estimate $y - (A x + \varepsilon) = y - [A I_m] \xi$ of the sparse outlier is encouraged to be sparse by employing $(\| \cdot \|_1)^{1/2} \| [A I_m] \xi - y \|$ as a fidelity function. The above arguments amount to the following minimization problem:

$$\min_{\xi \in \mathbb{R}^{n+m}} 0.5 \| \Sigma_\xi^{-1/2} \xi \|^2_2 + \mu \| \xi \|_1^{\gamma - 1/2} \| [A I_m] \xi - y \|.$$  

(26)

which is a special case of the LiMES model with $\mathcal{X} := \mathbb{R}^{n+m}, \mathcal{Z} := \mathbb{Z}^m, \Psi := \gamma \| \cdot \|_1, \mathcal{L} := I_m, D := \gamma^{-1/2} I_m$, and $\omega_2 : \xi \mapsto \| [A I_m] \xi - y \|$. The formulation in (26) is a general form of SORR. Under the statistical assumption stated in Section II-B, it follows that $\Sigma_\xi = \text{diag}(\sigma^2_{I_m}, \sigma^2_{I_m})$. We therefore let $\Sigma := \text{diag}(\sigma^2_{I_m}, \sigma^2_{I_m})$, with which (26) reduces to (8). Problem (26) can be solved by using Algorithm 1 under the convexity condition in (10). Note that, among the convergence conditions (i)–(iv) listed below Algorithm 1, only (i) and (ii) needs to be cared in this specific case. Indeed, conditions (iii) and (iv) is satisfied automatically, because (26) always has a solution due to the coercivity of the objective function12, and $\text{int}(\text{dom} \| \cdot \| - y \|_1)) \cap \text{range} [A I_m] \neq \emptyset$.

F. Stable Principal Component Pursuit: A Type-S Application

We consider the following model:

$$Y = L + S + W,$$

(27)

where $Y \in \mathbb{R}^{n \times m}$ is a noisy measurement of the superposition of the low-rank matrix $L \in \mathbb{R}^{n \times m}$ and the sparse matrix $S \in \mathbb{R}^{n \times m}$ with the additive white Gaussian noise $W \in \mathbb{R}^{n \times m}$. The problem of recovering $L$ and $S$ from the measurement $Y$ is called stable principal component pursuit (SPPC) [36], which we shall formulate as follows:

$$\min_{L, S \in \mathbb{R}^{n \times m}} 0.5 \| [I_n I_n] L \|_F^2 + \| \Psi_{M_1}^* \left( \begin{bmatrix} L \\ S \end{bmatrix} \right) \|_F^2,$$

(28)

Here, $\| \cdot \|_F$ denotes the Frobenius norm, $D := \text{diag}(\sqrt{\mu L}, \sqrt{\mu S}, \sqrt{\mu L}, \sqrt{\mu S})$, $\mathcal{L} := \mathcal{M}_1 \| \cdot \|_{\text{nu, } S} + \mu S \| \cdot \|_1$
and is a norm on $\mathbb{R}^{2n \times m}$ for any $\mu L, \mu S \in \mathbb{R}_{++}$ with $\| \cdot \|_1$ summing up the absolute values of the entries. It can be verified that

$$\Psi_D^* \left( \begin{bmatrix} L \\ S \end{bmatrix} \right) = \mu L \| L \|_{\text{nu, } S} + \mu S \| \cdot \|_1,$$

(29)

The SPPC formulation given in (28) is a special case of LiMES for $\mathcal{X} := \mathbb{R}^{2n \times m}, \mathcal{Y} := \mathbb{R}^{n \times m}, \mathcal{A} := [I_n I_n] \cdot \mathcal{Y}$, and $\omega_2 := I_{2n} (M_2 := I_{2n})$. We emphasize here that $(\mathcal{L} := \mathcal{M}_1)$ plays a key role for convexity preservation as in Section II-A,

11Due to the presence of the Moreau envelope in the smooth part $0.5 \| \delta_2 x \|^2 - \gamma \Psi \circ \delta_2$, the popular ADMM and Chambolle-Pock algorithms [66] are not suitable to the present case, because the former requires a minimizer of some function involving $0.5 \| \delta_2 x \|^2 - \gamma \Psi \circ \delta_2$ (and thus requires an inner loop), and the latter requires the proximity operator of $0.5 \| \delta_2 x \|^2 - \gamma \Psi \circ \delta_2$ which cannot be written in a closed form in general. Some other algorithms such as Condat’s primal dual splitting method [67] may also be used.

12A function $f \in \Gamma_0(\mathcal{H})$ is coercive if $f(x) \rightarrow +\infty$ as $\| x \| \rightarrow +\infty$.  

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
although the condition is given in terms of the parameters contained in $D$ as shown in the following proposition.

Proposition 6 (Convexity condition for (28)): Given $\mu_L, \mu_S, \gamma \in \mathbb{R}^{++}$, and $Y \in \mathbb{R}^{n \times m}$, the smooth part $0.5\left\| I_n I_n^T - Y Y^T \right\|_F^2 = (\Psi \circ D^2 \Theta) \circ D^2 P_{\mathcal{M}_1}$ is convex if and only if $\mu_L + \mu_S \leq 4\gamma$.

Proof: Since $c_2 := 0$ for SCP, the smooth part of (28) is convex if and only if (\ref{a}) is satisfied by Corollary 1. It can be verified that (\ref{a}) $\iff M_1^T M_1 - \mu M_2^T \mu \mathcal{L}^2 \mathcal{L}^2 M_2 = [I_n I_n]^T [I_n I_n] - P_{\mathcal{M}_1} D^2 P_{\mathcal{M}_1} = (1 - \frac{\mu_L + \mu_S}{4\gamma}) [I_n I_n]^T [I_n I_n] \geq 0 \iff 4\gamma \geq \mu_L + \mu_S$.

As the proximity operator of $\Psi$ can be computed directly by those of the individual functions $\mu_L \| \cdot \|_{\text{nuuc}}$ and $\mu_S \| \cdot \|_1$, the problem in (28) can be solved efficiently by the proximal gradient method (23). We remark that the formulation in (28) for $\mathcal{L} := I$ has been studied in the framework of GMC in [68], where the problem is solved by a convex optimization algorithm involving dual variables. In sharp contrast, no auxiliary variable is required in our case, because $D$ is diagonal (cf. Remark 1). An $\ell_0$-based approach can also be found in the literature [69].

G. Robust Classification: A Type-R Application

We consider a standard (supervised) classification task where the pairs $(a_i, y_i) \in \mathbb{R}^n \times \{+1, -1\}, i \in \{1, 2, \ldots, m\}$, of input vector and its label are available. We assume here that the input vectors $a_i$ are normalized such that $\|a_i\|_2 = 1$; it is implicitly assumed that $a_i \neq 0$. We then consider the following problem formulation:

$$\min_{x \in \mathbb{R}^n} 0.5\|x\|_2^2 + \mu \sum_{i=1}^m (\sigma_{[-1,0]} \circ (y_i a_i^T \cdot -1))_{\gamma, -1/2} f(x).$$

Here, $\sigma_{[-1,0]} \circ (y_i a_i^T \cdot -1)$ is the popular hinge loss, and thus each summand is the ME-hinge loss (see Example 2(b)). To show that (31) is a special case of LiMES, the following lemma will be used.

Lemma 3: Let $X'$ and $K'$ be finite dimensional Hilbert spaces. Let $\mathfrak{A} : X' \to K : x \mapsto \mathcal{L} x + b$, where $b \in K$ and $\mathcal{L} : X \to K$ is a bounded linear operator such that range $\mathcal{L} = K$ and $\mathfrak{A} \mathfrak{A}^\dagger = P_{\mathcal{L}}$ with $\mathcal{L} : \mathcal{L} \subset X'$.

Then, for any $\psi \in \Gamma_0(K)$ and $\gamma \in \mathbb{R}^+$, it holds that

$$\gamma (\psi \circ \mathfrak{A}) = \gamma \psi \circ \mathfrak{A}, \quad (\psi \circ \mathfrak{A})_{\gamma, -1/2} f = \psi_{\gamma, -1/2} f \circ \mathfrak{A}.$$ 

Proof: See Appendix C.

Proposition 7 ((31) as a special case of LiMES model): Let $\Psi : \mathbb{R}^m \to \mathbb{R} : z := [z_1, z_2, \ldots, z_m] \mapsto \sigma_{[-1,0]}(z) = \sum_{i=1}^m \sigma_{[-1,0]}(z_i) \circ \sigma_{\mathcal{L}_{\sigma}} : \mathbb{R}^m \to \mathbb{R}^m : x \mapsto M_2 x - m_1$ with $M_2 := [y_1 a_1 y_2 a_2 \cdots y_m a_m]^T \in \mathbb{R}^{m \times m}$ and $m_1 := [1, 1, \ldots, 1]^T \in \mathbb{R}^m$. Then, the second term in (31) can be expressed as

$$\Psi_{\gamma, -1/2} f \circ \sigma_{\mathcal{L}_{\sigma}} = \sum_{i=1}^m (\sigma_{[-1,0]} \circ (y_i a_i^T \cdot -1))_{\gamma, -1/2} f.$$ 

Proof: Let $(O \neq) M_{2,i} : \mathbb{R}^n \to \mathbb{R} : x \mapsto \frac{i}{x_i} a_i^T x, \quad i = 1, 2, \ldots, m$. It then holds that $M_{2,i} M_{2,i}^T = P_{\text{range} M_{2,i}}^2$, as $\|M_{2,i}\| = 1$, and range $M_{2,i} = \mathbb{R}$ as $M_{2,i} \neq O$. For each $i \in \{1, 2, \ldots, m\}$, letting $K := \mathbb{R}, \psi := \sigma_{[-1,0]}$, $\mathcal{L} := M_{2,i}$, and $b := -1$ in Lemma 3 yields

$$(\sigma_{[-1,0]} \circ (y_i a_i^T \cdot -1))_{\gamma, -1/2} f(x) = \sigma_{[-1,0]} \circ (y_i a_i^T \cdot -1) \circ (x_i).$$

from which together with the separability of $\Psi$ it follows that $\Psi_{\gamma, -1/2} f(x) = \sum_{i=1}^m (\sigma_{[-1,0]} \circ (y_i a_i^T \cdot -1))_{\gamma, -1/2} f(x) = \sum_{i=1}^m (\sigma_{[-1,0]} \circ (y_i a_i^T \cdot -1))_{\gamma, -1/2} f(x)$.

In light of Proposition 7, the formulation in (31) is a special case of LiMES for $\mathcal{X} := \mathbb{R}^n, \mathcal{Z} := \mathbb{R}^m, \mathcal{A}_{\sigma} := [I_n I_n]^T, \mathcal{L} := I_m$, and $\mathcal{L} := \gamma^{-1} I_m$; Table III summarizes the applications of LiMES. The convexity condition is given as below.

Proposition 8 (Convexity condition for (31)): The smooth part of (31) is convex if $\mu_{\max}(M_2 M_2^T) \leq \gamma$. Suppose, in particular, that (i) range $M_2 = \mathbb{R}^m$, or (ii) $\gamma \in (1, +\infty)$. Then, the smooth part of (31) is convex if and only if $\mu_{\max}(M_2 M_2^T) \leq \gamma$.

Proof: Assume that range $M_2 = \mathbb{R}^m$. In this case, since $D$ is a positive definite operator and $\mathcal{L} = I_m$, we have range $(D^2 \mathcal{L} M_2) = \mathcal{Z}$, and hence (iii) of Lemma 2 is clearly satisfied. Assume on the other hand that $\gamma \in (1, +\infty)$. It then holds that $D^2 \mathcal{L} \mathcal{A}_{\sigma} \mathcal{L}^2 \mathcal{A}_{\sigma} = \gamma^2 (M_2 \mathcal{O}_m - 1_m) = \gamma^{-1} 1_m \mathcal{L} \subset \mathcal{X}$, and thus (iii) of Lemma 2 is satisfied again.

Thus, it follows that int $K \neq \emptyset$ under any of conditions (i) and (ii) of the proposition. Hence, in light of Propositions 5 and 7, the smooth part of (31) is convex if and only if (\ref{a}) is satisfied. Finally, (\ref{a}) $\iff M_1^T M_1 - \mu \mathcal{L}^2 \mathcal{L}^2 M_2 = [I_n I_n]^T [I_n I_n] - \mu \gamma^{-1} M_2 M_2^T \geq 0 \iff 1 - \mu \gamma^{-1} \mu_{\max}(M_2 M_2^T) \geq 0$. This verifies the assertion.

IV. NUMERICAL EXAMPLES

We show the efficacy of the LiMES model in two applications: sparse modeling in the underdetermined case and robust regression.

A. Experiment A: Sparse Modeling in Underdetermined Case

We compare the performance of the PMC penalty (see Section II-A) for sparse modeling with those of the following penalties: $\ell_1$ (lasso) implemented by the iterative shrinkage-thresholding algorithm (ISTA) [47], and GMC with the linear operator $B := (\alpha_{\text{GMC}}/\mu)^{1/2} A$ for $\alpha_{\text{GMC}} \in [0, 1]$. The standard linear model $y = Ax + \varepsilon$, is considered with the i.i.d. standard Gaussian input matrix $A \in \mathbb{R}^{m \times n}$ for $m := 64$ and $n := 128$. Here, $x_0 \in \mathbb{R}^n$ is the sparse unknown vector with $s$ nonzero components, and $\varepsilon_0 \in \mathbb{R}^m$ is the i.i.d. zero-mean Gaussian noise vector with signal-to-noise ratio (SNR) 20 dB and 30 dB, where SNR := $\|Ax_0\|_2^2/\|\varepsilon_0\|_2^2$. The regularization parameter is tuned so that all the methods share the same sparseness as the true $x_0$ with respect to the sparseness measure [70] $\|n/(n - \sqrt{n})\|[1 - \|x\|_1/\|\sqrt{n}\|\|\|\|\} \in [0, 1]$. For PMC, $\gamma := \mu/\alpha_{\text{PMC}}\lambda_{\min}(A^T A)$ for $\alpha_{\text{PMC}} \in (0, 1]$ is used (see
II-A shows a particular instance (SNR 20 dB, 2(a) plots the average estimate of each method and $\alpha$ (b) $\in$ and the ridge regression are plotted. Due to the nonconvexity show the system mismatch $= 21 \in$ for different sparsity levels. It can be seen that PMC $\parallel$ are tuned manually ($M_S := \text{range } [0, \infty \cdot \text{AND } \cdots \cdot M_{\text{FOR }} := \{M_Y, \ldots \cdot \text{SPCP}, \text{I and } \in \sigma = \text{subject to } \sigma \in \mu = \{\gamma \in \mathbb{R}\}$ for its potentially better performances. The MC function is employed in our $\parallel$ and noise $\min \in [5]$ based on Proposition $\parallel \text{by } \in \text{and } \mu \in \text{and } \sigma_3 \in \text{or } = 1$ can be unacceptably large sometimes, although it may perform better than PMC on average. This clearly suggests the efficacy of the PMC penalty.

II-B shows the system mismatch $\parallel \in$ for different sparsity levels. It can be seen that PMC $\parallel$ are tuned manually ($M_S := \text{range } [0, \infty \cdot \text{AND } \cdots \cdot M_{\text{FOR }} := \{M_Y, \ldots \cdot \text{SPCP}, \text{I and } \in \sigma = \text{subject to } \sigma \in \mu = \{\gamma \in \mathbb{R}\}$ for its potentially better performances. The MC function is employed in our $\parallel$ and noise $\min \in [5]$ based on Proposition $\parallel \text{by } \in \text{and } \mu \in \text{and } \sigma_3 \in \text{or } = 1$ can be unacceptably large sometimes, although it may perform better than PMC on average. This clearly suggests the efficacy of the PMC penalty.

Section II-A). The parameters $\alpha_{\text{GMC}}$ and $\alpha_{\text{PMC}}$ are tuned manually to attain the lowest system mismatch for each method. The results are averaged over 300 trials.

Figs. 2(a) and (b) show the system mismatch $\parallel ||x_0 - x||^2 ||x_0||^2$ for different sparsity levels. It can be seen that PMC outperforms the other methods particularly when the sparsity level is middle, $s \in [18, 24]$ more specifically. Note here that the proposed approach requires no auxiliary vector unlike GMC (see Remark 1). Fig. 2(c) plots the average estimate of each method over the 300 trials for SNR 20 dB with sparsity level $s := 21$. It can be seen that PMC estimates $x_0$ with high accuracy, indicating that the estimation bias is reduced successfully.

Finally, Fig. 3 shows a particular instance (SNR 20 dB, $s := 21$) to show that a direct application of the original MC penalty to an underdetermined system may fail. The MC penalty is implemented by ISDA in (7). For reference, the performance of the ordinary least square (OLS) estimate $A^\dagger y \in \text{argmin}_{x \in \mathbb{R}^n} ||Ax - y||^2$ and the ridge regression are plotted. Due to the nonconvexity of the objective function involving the original MC penalty in the present underdetermined case, the system mismatch of MC could be unacceptably large sometimes, although it may perform better than PMC on average. This clearly suggests the efficacy of the PMC penalty.

B. Experiment B: Robust Regression in the Presence of Outlier

We compare the performances of SORR and OLS (see Section II-B) for robust regression with those of LAD [5], LAD-ridge ($\ell_1$-loss + Tikhonov regularization), Huber’s loss $\gamma || \cdot ||_1$ [2], [5], and the state-of-the-art method called the robust projected generalized gradient (RPGG) algorithm [35] which is based on the following formulation

$$\min_{x \in \mathbb{R}^n, e \in \mathbb{R}^m} \mu (|| \cdot ||_1 - \gamma || \cdot ||_1(e) \text{ subject to } y = Ax + e \text{ for } \gamma_1, \gamma_2 \in [0, \infty]) \text{.}$$

The sparse outlier model $y = Ax + \varepsilon + o_\sigma$ is used, where the input matrix $A \in \mathbb{R}^{m \times n}$ and noise $\varepsilon \in \mathbb{R}^m$ are generated randomly with $m := 128$ and $n := 64$ in the same way as in Experiment A. To show that SORR is stable under large Gaussian noise, we consider the cases of SNR 10 dB and 20 dB. The nonsparse vector $x_s \in \mathbb{R}^m$ is generated randomly from the i.i.d. standard Gaussian distribution (i.e., $\sigma_x^2 := 1$). The outlier vector $o_\sigma$ is sparse with nonzero positions chosen randomly and with nonzero components generated from an i.i.d. zero-mean Gaussian distribution with variance determined by the signal-to-outlier ratio (SOR) $\gamma || \cdot ||_1(Ax_s + o_\sigma/\text{supp}(o_\sigma))^{-1}$. Here, supp$(x) := \{i \in \{1, 2, \ldots, m\} \mid x_i \neq 0\}$ is the support of a vector $x \in \mathbb{R}^m$. SORR, $\sigma_x^2 := \sigma_x^2$, and $\sigma_o^2 := \sigma_o^2$ are used to show the potential performance. For the primal–dual debiasing algorithm, the parameters are chosen as follows. The parameters $\tau$ and $\sigma$ are set to slightly smaller values than the upper bounds, respectively, shown under Algorithm 1. We simply let $\beta_k := 1$ for all $k \in \mathbb{N}$, and tune $\gamma$ and $\mu$ based on Proposition 3 by grid search to attain the best performance. For RPGG, we let $\gamma_1 := +\infty$ (i.e., $|| \cdot ||_1 - \gamma || \cdot ||_1(e) \text{ subject to } y = Ax + e \text{ for } \gamma_1, \gamma_2 \in [0, \infty])$ as $x_s$ is nonspars, and tune $\mu$ and $\gamma_2$ as well as the step size by grid search. For the other methods involving regularizers, the regularization parameters are tuned by grid search to attain the best performance. For

13Although RPGG is a method for robust sparse recovery, it could be used in the present nonspars case by letting $\mu := 0$. We instead tune $\mu$ to seek for its potentially better performances. The MC function is employed in our simulations for both data fidelity and penalty, as in the simulations of [35].
Huber’s loss, $\gamma$ is chosen to attain the best performance. The results are averaged over 300 trials.

Fig. 4 plots the results across outlier density $\text{supp}(o_c)/m$. The proposed SORR method exhibits highly accurate and stable performances, and it outperforms all the other methods significantly. To be specific, the difference from ORR is notable when the outlier density is low to middle. It should be mentioned that LAD performed poorly due to the presence of heavy noise as well as strong outliers. Fig. 5 plots the results across SOR to show the impacts of the change of the outlier power on the performance. Remarkably, the performances of SORR and ORR even improve as SOR decreases below $-12$ dB. This is because the influence of huge outliers on the MC loss vanishes above a certain range due to the same reason as for Tukey’s loss [5] and because such huge outliers will be easier to detect at the same time. The results clearly indicate the remarkable robustness of SORR (and ORR) against huge outliers. We mention that RPGG also exhibits a similar tendency over a reasonable range, although its performance degrades for SOR below $-40$ dB.  

V. CONCLUDING REMARKS

We presented the efficient framework based on the LiMES model. The PMC penalty composes the Moreau envelope contained in the standard MC penalty with the projection operator onto the input subspace, thereby restricting the Moreau-enhancement effect to the subspace for preserving the overall convexity even in the underdetermined case. SORR distinguishes Gaussian noise and sparse outlier explicitly to attain stable performances in highly noisy situations. The convexity conditions for those specific instances were discussed in a unified fashion with the LiMES model. While the LiMES function is “nonseparable,” the objective function involved in the Moreau envelope is “separable”. This mixed nature of separability and nonseparability allows an application of the LiMES model to the case when the fidelity term is not strongly convex (as in the underdetermined case of linear regression) with an efficient implementation using the proximal gradient method. The operators $\mathcal{L}$ and $\mathcal{L}_2$ play key roles in the model: $\mathcal{L}$ corresponds to the projection mentioned above and $\mathcal{L}_2$ takes care of robust regression. The proximal debiasing algorithms to compute the LiMES model require convexity of the smooth part of the objective function, for which a sufficient condition was presented. The condition was shown to be a necessary condition as well under the nonempty-interior assumption when the seed function is a support function. This is the case for instance when the seed function is a norm and the range of $\mathcal{L}_2$ contains the zero vector. Applications of the LiMES model to SPCP and robust classification were also presented. The hinge loss function widely used for robust classification was shown to be expressed as a composition of the support function of a closed interval $[-1, 0]$ and an affine operator. Numerical examples showed that (i) the PMC penalty achieved debiased sparse modeling for underdetermined systems as well as outperforming GMC, and that (ii) SORR achieved stable and remarkably robust performances in the presence of both heavy Gaussian noise and sparse outlier as well as outperforming the existing robust methods including LAD, Huber’s loss, and RPGG.

The LiMES model will serve as a powerful tool to enhance performances with respect to a variety of penalty/loss functions based on the solid foundation of convex analysis, and there are plenty of opportunities to explore its further applications. In particular, it is our future works to investigate the efficacy of the LiMES model in SPCP and robust classification.

APPENDIX A

PROOF OF PROPOSITION 2

Since $c := 0$ for the debiased sparse modeling, the smooth part of (4) is convex if and only if (♀) is satisfied by Corollary 1. By definition of $M := \text{range } A^\top$, moreover, it holds that $P_M A^\top = A^\top$, from which together with $P_M = P_M \circ P_M$ it follows that (♀) $\iff M_1^\top M_1 - \mu M_2^\top D^2 L^2 M_2 A^\top A - \mu \gamma^{-1} P_M = P_M (A^\top A - \mu \gamma^{-1} I) P_M \succeq 0 \iff \lambda_{\min}^+ (A^\top A) \geq \mu \gamma^{-1}$.

APPENDIX B

PROOF OF PROPOSITION 3

According to the discussions in Section III-E, (8) is equivalent to (26). Since range $M_2 = \text{range } \lfloor A I_n \rfloor = \mathcal{Z}$ for SORR, the smooth part of (26) is convex if and only if (♀) is satisfied by Corollary 1. We prove the equivalence (♀) $\iff (\ast)$ below.
For \( \Sigma_\xi^{-1} = \text{diag}(\sigma_\xi^{-2}I_n, \sigma_\xi^{-2}I_m) \), it holds that (\( \triangle \)) \( \iff \)
\[ M_1^T M_1 - \mu M_1^T \mathcal{L}^T D^2 \mathcal{L} M_2 = \Sigma_\xi^{-1} - \mu \gamma^{-1}[A I_m]^T[A I_m] \geq 0 \]
which can be expressed equivalently as follows:
\[ \begin{bmatrix} \mu^{-1}(\gamma^{-1} \xi - A^T A) & -A^T \\ -A & (\mu^{-1}(\gamma^{-1} \xi - 1))I_m \end{bmatrix} \succeq 0. \tag{B.1} \]
By [60, Theorem 7.7.9], (B.1) holds if and only if all of the following conditions are satisfied:
\begin{enumerate}
\item \( \mu^{-1}(\gamma^{-1} \xi - 1) I_m \succeq 0 \Rightarrow \mu \max(A^T A)/\gamma \sigma_\xi^{-2} \)
\item \( \mu^{-1}(\gamma^{-1} \xi - 1) I_m \succeq 0 \Rightarrow \mu \leq \gamma \sigma_\xi^{-2} \)
\item \( A^T = (\mu^{-1}(\gamma^{-1} \xi - 1) I_m = A^T A)^{1/2} \gamma_\xi^{-1}(\mu^{-1}(\gamma^{-1} \xi - 1) I_m)^{1/2} \)
\end{enumerate}
for some \( \gamma_\xi \in \mathbb{R}^{n \times m} \) with its largest singular value at most one.

If \( A = 0 \), then conditions (i) and (ii) hold trivially, and condition (iii) coincides with (10). Assume that \( A \neq 0 \) in the following. We shall show below that (i)–(iii) \( \iff \) (10). Suppose that conditions (i)–(iii) are satisfied. Condition (ii) under \( A \neq 0 \) implies that \( \mu^{-1}(\gamma^{-1} \xi - 1) \neq 0 \), and hence \( \mu^{-1}(\gamma^{-1} \xi - 1) > 0 \) by condition (ii). The equality in condition (iii) above can be rewritten as
\[ \nu_{\xi} A^T = \left( \nu_{\xi} I_n - A^T A \right)^{1/2} \gamma_\xi^\top, \tag{B.2} \]
where \( \nu_{\xi} := (\mu^{-1}(\gamma^{-1} \xi - 1) - 1)^{1/2} > 0 \), \( \nu_{\xi} : = \mu^{-1}(\gamma^{-1} \xi - 1) > 0 \), and \( \gamma_\xi : = \left( \begin{array} {c} \gamma_\xi \\ \xi \end{array} \right) \). Let \( A = V \Sigma U^\top \) be a singular value decomposition of \( A \), where \( U \in \mathbb{R}^{n \times n} \) and \( V \in \mathbb{R}^{m \times m} \) are orthogonal matrices, and \( \Sigma = \text{diag}(\xi_1, \xi_2, \ldots, \xi_{\min(n,m)}) \in \mathbb{R}^{n \times m} \) having \( \xi_1 \geq \xi_2 \geq \cdots \geq \xi_{\min(n,m)} \geq 0 \) for the diagonal entries and zeros for the off-diagonal entries. Then, (B.2) can be rewritten as
\[ U(\nu_{\xi} \Sigma)V^\top = U(\nu_{\xi} I_n - (\Sigma^\top \Sigma)^{1/2} U^\top \gamma_\xi V) \]
\[ \iff \nu_{\xi} \Sigma = (\nu_{\xi} I_n - (\Sigma^\top \Sigma)^{1/2}) U^\top \gamma_\xi \tag{B.3} \]
Let \( \gamma_\xi = -\gamma_\xi U^\top \) for some matrix \( \gamma_\xi \in \mathbb{R}^{n \times m} \). Then, (B.3) reads
\[ \nu_{\xi} \Sigma = (\nu_{\xi} I_n - (\Sigma^\top \Sigma)^{1/2}) \gamma_\xi. \tag{B.4} \]
Noting that \( \xi_1 > 0 \) due to the assumption \( A \neq 0 \), one can verify from (B.4) that \( \gamma_\xi \) must be written in the following form:
\[ \gamma_\xi = \text{diag}(\xi_1, \ldots, \xi_{\min(n,m)}) \in \mathbb{R}^{n \times m}, \tag{B.5} \]
of which the (1,1) entry is \( \xi_1 > 0 \), the lower-right submatrix is \( \xi_{2,2} \in \mathbb{R}^{(n-1) \times (m-1)} \), and the entries of the off-diagonal blocks are zeros. By (B.4) and (B.5), we obtain
\[ \nu_{\xi} \xi_1 = (\nu_{\xi} - \xi_1^{2/3})^{1/2} \xi_1, \quad \xi_1 > 0. \tag{B.6} \]
where \( \nu_{\xi} := \xi_1^{2/3} > 0 \) as \( \nu_{\xi} \xi_1 > 0 \). To see that \( \xi_1 > 0 \), is a singular value of \( \gamma_\xi \) (or that of \( \gamma_\xi U^\top \) equivalently), \( \xi_{2,2} \in \mathbb{R}^{(n-1) \times (m-1)} \) and \( U_{\xi_{2,2}} \in \mathbb{R}^{(m-1) \times (n-1)} \) are orthogonal matrices, and \( \xi_{2,2} : = \text{diag}(\xi_2, \ldots, \xi_{\min(n,m)}) \in \mathbb{R}^{(n-1) \times (m-1)} \) for singular values \( \xi_{1} \geq \cdots \geq \xi_{\min(n,m)} \geq 0 \). It then follows that \( \gamma_\xi = \text{diag}(\xi_1, \xi_2, \ldots, \xi_{\min(n,m)}) \in \mathbb{R}^{n \times m} \). Hence, \( \gamma_\xi = \text{diag}(\xi_1, \xi_{2,2}) \) agrees.

Here, the second equality is due to the Pythagorean theorem, the third equality holds because \( \psi(2u) = \psi^2 \) is independent of \( P_{\mathcal{L}}u \), the fourth equality is due to \( \mathcal{L}^2 = P_{\mathcal{L}} \) \( \iff \) \( P_{\mathcal{L}} = P_{\mathcal{L}}^\top \) \( \iff \) \( P_{\mathcal{L}} \). Finally, the fifth equality is due to range \( \mathcal{L} = \mathcal{K} \). By (C.1), it follows that \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \) \( \iff \) \( \psi(x) = \psi(2x) \).
Masahiro Yukawa (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees from the Tokyo Institute of Technology, Tokyo, Japan, in 2002, 2004, and 2006, respectively. He is currently a Professor with the Department of Electronics and Electrical Engineering, Keio University, Yokohama, Japan. He is currently the Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2015 to 2019, Springer Journal of Multidimensional Systems and Signal Processing from 2012 to 2016, and IEEE Transactions on Fundamentals of Electronics, Communications and Computer Sciences from 2009 to 2013. His research interests include mathematical adaptive signal processing, convex/sparse optimization, and machine learning.

Dr. Yukawa was the recipient of the JSPS Prize in 2021, Young Scientists’ Prize, Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology in 2014, Excellent Paper Award from the IEICE in 2006, among many others. He is a Member of the IEICE.

Kyohei Suzuki (Graduate Student Member, IEEE) received the B.E. and M.E. degrees in 2020 and 2022, respectively, in electronics and electrical engineering from Keio University, Yokohama, Japan, where he is currently working toward the Ph.D. degree in electronics and electrical engineering. His research interests include mathematical signal processing, sparse optimization, and robust statistics.

Isao Yamada (Fellow, IEEE) received the B.E. degree in computer science from the University of Tsukuba, Tsukuba, Japan, in 1985, and the M.E. and Ph.D. degrees in electrical and electronic engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1987 and 1990, respectively. He is currently a Professor with the Department of Information and Communications Engineering, Tokyo Institute of Technology. His research interests include mathematical signal processing, nonlinear inverse problems, and optimization theory. Since 2015, he has been the IEICE Fellow. He was the recipient of the MEXT Minister Award (Research Category), IEEE Signal Processing Magazine Best Paper Award in 2015, IEICE Excellent Paper Awards (in 1991, 1995, 2006, 2009, 2014 and 2022), IEICE Achievement Award in 2009, ICFR Research Award in 2004, Docomo Mobile Science Award (Fundamental Science Division) in 2005, and Fujino Prize in 2008. He was a Member of the IEEE Signal Processing Society Awards Board in 2022.

Hiroyuki Kaneko (Member, IEEE) received the B.E. and M.E. degrees in electronics and electrical engineering from Keio University, Yokohama, Japan, in 2020 and 2022, respectively. He is currently a Researcher with NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan. His research interests include sparse signal processing, convex optimization, and audio signal processing.