Stable Robust Regression under Sparse Outlier and Gaussian Noise

Masahiro Yukawa Kyohei Suzuki
Dept. Electronics and Electrical Engineering
Keio University, JAPAN

Isao Yamada
Dept. Information and Communications Engineering
Tokyo Institute of Technology, JAPAN

***Abstract*—We propose an efficient regression method which is highly robust against outliers and stable even in the severely noisy situations. The robustness here comes from the adoption of the minimax concave loss, while the stability comes from separate treatments of the outlier and noise by an introduction of an auxiliary vector modeling the Gaussian noise. We present a necessary and sufficient condition for convexity of the smooth part of the entire cost under a certain assumption, where a general model is used with its potential use for other applications envisioned. We show that the proposed formulation can be solved via reformulation by the forward-backward-based primal-dual method under the convexity condition. The numerical examples show the remarkable robustness of the proposed estimator under highly noisy situations.**

***Index Terms*—weakly convex function, minimax concave loss, Moreau envelope, robust regression**

I. Introduction

Outlier robustness is an inevitable issue of paramount importance in modern signal processing including future wireless communication systems as well as in machine learning [2, 3]. Its popularity stems from the fact that the typical squared-error loss function has vulnerability in the presence of outliers. A simple convex loss for robust regression is the least absolute deviation (LAD), which is the ℓ_1 norm of the estimation residual [4]. The most prominent example of the convex loss functions for robust regression is Huber's loss. Despite its mathematical tractability due to its convexity, Huber's loss has a severe limitation on its outlier robustness because it increases linearly as the error increases above a certain range. Another prominent example for robust regression is Tukey's biweight loss [5]. In contrast to Huber's loss, Tukey's biweight loss is mathematically intractable due to its nonconvexity, whereas it is highly robust against outliers due to the so-called *redescending property* (see Section III-A). As such, Huber's and Tukey's loss functions have an intrinsic tradeoff between the robustness and mathematical tractability.

In this paper, we shed a new light on this tradeoff by proposing an efficient regression method which is robust against strong outliers and is stable under severe Gaussian noise. The first ingredient of the proposed method is the adoption of the minimax concave (MC) penalty [6, 7] to define a loss function. This is the key for resolving the tradeoff issue. While the MC loss is a nonconvex function, it is weakly convex so that the overall cost function may become convex by adding the quadratic penalty (which comes naturally from our assumption about Gaussianity of the unknown vector to be estimated). It is mathematically tractable owing to the overall convexity, and it is highly robust against catastrophic outliers because it saturates in analogy with Tukey's biweight loss.

We then highlight another property of the MC penalty that the derivative does not vanish at the origin. This property suggests that the MC loss increases sharply when deviating slightly from the origin, and this may cause sensitivity to small perturbations. As the second ingredient, we thus introduce an auxiliary variable vector modeling small perturbations (Gaussian noise) to accommodate prior information about the outlier and noise. In our formulation, the unknown target vector and the noise vector are evaluated by the squared Euclidean norms, while the outlier is evaluated by the MC function. We show that the proposed formulation can be solved via reformulation by the forward-backward-based primal-dual method [8], provided that the smooth part of the entire cost function is convex. A necessary and sufficient condition for the convexity is presented based on a certain general model. The numerical examples show remarkable robustness of the proposed method under highly noisy situations.

II. Preliminaries

We briefly give the notation and state the problem addressed in this work.

A. Notation and definitions

Let $\mathbb{R}$, $\mathbb{R}_{++}$, and $\mathbb{N}$ denote the sets of real numbers, strictly positive real numbers, and nonnegative integers, respectively. For any $n, m \in \mathbb{N}^* := \mathbb{N} \setminus \{0\}$, the $n \times n$ identity matrix is denoted by $\boldsymbol{I}_n$, and the $n \times m$ zero matrix is denoted by $\boldsymbol{O}_{n\times m}$. Matrix transpose is denoted by $(\cdot)^\mathsf{T}$. The largest eigenvalue of a symmetric matrix is denoted by $\lambda_{\max}(\cdot)$. The ℓ_1 and ℓ_2 norms of Euclidean vector $\boldsymbol{x} := [x_1, x_2, \cdots, x_n]^\mathsf{T} \in \mathbb{R}^n$ are defined respectively by $\|\boldsymbol{x}\|_1 := \sum_{i=1}^n |x_i|$ and $\|\boldsymbol{x}\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$.

Given any function $f : \mathcal{H} \to (-\infty, +\infty] := \mathbb{R} \cup \{+\infty\}$ defined on a real Hilbert space $\mathcal{H}$ equipped with the induced norm $\|\cdot\|$, the function

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

is the Moreau envelope of f of index $\gamma \in \mathbb{R}_{++}$ [9], and

$$ \mathrm{Prox}_{\gamma f} : \mathcal{H} \to \mathcal{H} : x \mapsto \operatorname*{argmin}_{y\in\mathcal{H}} \left(f(y) + \frac{1}{2\gamma} \|x-y\|^2 \right) \quad (2)$$

is the proximity operator of f of index γ.

B. Robust regression under Gaussian noise and sparse outlier

We consider the following linear model: $\boldsymbol{y} := \boldsymbol{A}\boldsymbol{x}_\star + \boldsymbol{\varepsilon}_\star + \boldsymbol{o}_\diamond$, where the output vector $\boldsymbol{y}$ is a linear transform of the unknown vector $\boldsymbol{x}_\star \in \mathbb{R}^n$ under the known matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ subject to the Gaussian noise $\boldsymbol{\varepsilon}_\star \in \mathbb{R}^m$ and the sparse outlier $\boldsymbol{o}_\diamond \in \mathbb{R}^m$ [4]. We assume that $\boldsymbol{x}_\star \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon}_\star \in \mathbb{R}^m$ are mutually uncorrelated random vectors both of which obey i.i.d. zero-mean normal distributions with variances $\sigma^2_{\boldsymbol{x}_\star} \in \mathbb{R}_{++}$ and $\sigma^2_{\boldsymbol{\varepsilon}_\star} \in \mathbb{R}_{++}$, respectively (see Remark 1 for discussions about these statistical assumptions). In this case, the random vector $\boldsymbol{\xi}_\star := [\boldsymbol{x}_\star^\mathsf{T}\ \boldsymbol{\varepsilon}_\star^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{n+m}$ obeys

This work was supported by JSPS KAKENHI Grant Number (22H01492). A full version of the present work is available in [1].

a zero-mean normal distribution with its (nonsingular) covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_\star} = \begin{bmatrix} \sigma_{\boldsymbol{x}_\star}^2 \boldsymbol{I}_n & \boldsymbol{O}_{n\times m} \\ \boldsymbol{O}_{m\times n} & \sigma_{\boldsymbol{\varepsilon}_\star}^2 \boldsymbol{I}_m \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$.

III. Stable Outlier-Robust Regression in the Presence of Outlier

We first present a comparison between Tukey's biweight and MC functions (both of which are nonconvex), which leads to the proposed formulation involving an auxiliary vector. We then present the convexity condition for the proposed formulation, and an optimization algorithm is finally presented.

A. Tukey's biweight and MC functions

Tukey's biweight is defined by $\Phi_\gamma^{\mathrm{TK}}(\boldsymbol{a}) := \sum_{i=1}^m \phi_\gamma^{\mathrm{TK}}(a_i)$, $\boldsymbol{a} := [a_1, a_2, \cdots, a_m]^\top \in \mathbb{R}^m$, with $\phi_\gamma^{\mathrm{TK}}(a) := (\gamma^2/6)\left[1-(1-(a/\gamma)^2)^3\right]$ if $|a| \le \gamma$; $\phi_\gamma^{\mathrm{TK}}(a) := \gamma^2/6$ if $|a| > \gamma$. Its derivative $(\phi_\gamma^{\mathrm{TK}})'$ has the following properties: (i) the derivative vanishes $(\phi_\gamma^{\mathrm{TK}})'(0) = 0$ at the origin, (ii) $(\phi_\gamma^{\mathrm{TK}})'$ increases up to some point, and (iii) it then decreases until vanishing completely; i.e., $(\phi_\gamma^{\mathrm{TK}})'(a) = 0$ if $|a| \ge \gamma$. This is the so-called strong redescending property [4] (see Fig. 1). Property (i) indicates insensitivity to small perturbations as well as to large outliers.

The MC function, on the other hand, of index $\gamma \in \mathbb{R}_{++}$ is defined by [6, 7]

$$\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{a}) := \|\boldsymbol{a}\|_1 - {}^{\gamma}\|\cdot\|_1(\boldsymbol{a}) = \sum_{i=1}^m \phi_\gamma^{\mathrm{MC}}(a_i),\ \boldsymbol{a} \in \mathbb{R}^m, \quad (3)$$

where

$$\phi_\gamma^{\mathrm{MC}} : \mathbb{R} \to \mathbb{R} : a \mapsto \begin{cases} |a| - a^2/2\gamma, & \text{if } |a| \le \gamma, \\ \gamma/2, & \text{if } |a| > \gamma. \end{cases} \quad (4)$$

We refer to γ as *the saturation parameter*, since it controls the points at which $\phi_\gamma^{\mathrm{MC}}$ saturates on each side of the real axis. The Moreau envelope ${}^{\gamma}\|\cdot\|_1$ coincides with the well-known Huber function. The MC function $\phi_\gamma^{\mathrm{MC}}$ is differentiable at all points but the origin, and its derivative is given by $(\phi_\gamma^{\mathrm{MC}})'(a) = \mathrm{sign}(a) - a/\gamma$ if $|a| \in (0,\gamma)$; $(\phi_\gamma^{\mathrm{MC}})'(a) = 0$ if $|a| \ge \gamma$ (see Fig. 1).[1] This implies that $(\phi_\gamma^{\mathrm{MC}})'(a)$ has property (iii) raised above, which leads to remarkable robustness against huge outliers [4], while it lacks property (i), i.e., the derivative does not vanish at the origin: $\lim_{a\downarrow 0}(\phi_\gamma^{\mathrm{MC}})'(a) = 1$ and $\lim_{a\uparrow 0}(\phi_\gamma^{\mathrm{MC}})'(a) = -1$. We therefore introduce an auxiliary vector in our formulation to model small (and nonsparse) perturbations so that the residual evaluated by the MC function is desired to be exactly sparse.

B. Proposed formulation

Relying on the sparsity of the outlier vector $\boldsymbol{o}_\diamond$ as well as the Gaussianity of $\boldsymbol{x}_\star$ and $\boldsymbol{\varepsilon}_\star$, we formulate the robust regression task as follows:[2]

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{\varepsilon}\in\mathbb{R}^m} \mu\Phi_\gamma^{\mathrm{MC}}(\underbrace{\boldsymbol{y}-(\boldsymbol{A}\boldsymbol{x}+\boldsymbol{\varepsilon})}_{\text{estimate of } \boldsymbol{o}_\diamond}) + \frac{1}{2\sigma_{\boldsymbol{x}}^2}\|\boldsymbol{x}\|_2^2 + \frac{1}{2\sigma_{\boldsymbol{\varepsilon}}^2}\|\boldsymbol{\varepsilon}\|_2^2, \quad (5)$$

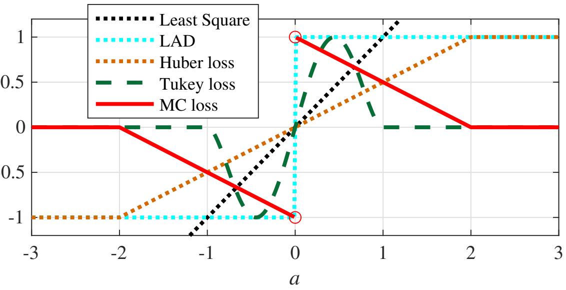


Fig. 1. The derivatives of $\eta\phi_\gamma^{\mathrm{TK}}$ and $\phi_\gamma^{\mathrm{MC}}$ for $\gamma := 1$ and $\eta := 25\sqrt{5}/16$.

where $\mu \in \mathbb{R}_{++}$ is the regularization parameter, $\sigma_{\boldsymbol{x}}^2 \in \mathbb{R}_{++}$ and $\sigma_{\boldsymbol{\varepsilon}}^2 \in \mathbb{R}_{++}$ are estimates of $\sigma_{\boldsymbol{x}_\star}^2$ and $\sigma_{\boldsymbol{\varepsilon}_\star}^2$, respectively. The first term $\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{y} - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}))$ of (5) is the MC loss encouraging sparsity of the estimation residual $\boldsymbol{y} - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon})$ which can be regarded as an estimate of the sparse outlier. The last two terms are derived directly from the Gaussianity assumptions (see Remark 1 below) of $\boldsymbol{x}_\star$ and $\boldsymbol{\varepsilon}_\star$, playing double roles of convexification and regularization (in the Tikhonov sense). Intuitively, when the noise power $\sigma_{\boldsymbol{\varepsilon}_\star}^2$ is large, the inverse $\sigma_{\boldsymbol{\varepsilon}}^{-2}$ of its estimate would be small, allowing $\|\boldsymbol{\varepsilon}\|_2^2$ to be large so that $\boldsymbol{\varepsilon}$ mimics $\boldsymbol{\varepsilon}_\star$ (of which the norm $\|\boldsymbol{\varepsilon}_\star\|_2$ is expected to be large due to the large noise power) well to mitigate the MC loss $\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{y} - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}))$ efficiently. This leads to the "stability" of our estimator in the spirit of [11]. We therefore call it *stable outlier-robust regression (SORR)*. We present below the SORR formulation in a slightly different shape:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{\varepsilon}\in\mathbb{R}^m} \widetilde{\mu}\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{y} - (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon})) + \frac{1}{2}\|\boldsymbol{x}\|_2^2 + \frac{\rho}{2}\|\boldsymbol{\varepsilon}\|_2^2, \quad (6)$$

where $\rho := \sigma_{\boldsymbol{x}}^2/\sigma_{\boldsymbol{\varepsilon}}^2 \in \mathbb{R}_{++}$ is an estimate of the signal to noise ratio (SNR), and $\widetilde{\mu} := \sigma_{\boldsymbol{x}}^2\mu \in \mathbb{R}_{++}$ (see Remarks 1 and 2). The performance of the SORR estimator is fairly insensitive to small fluctuations of ρ, as shown in Section V.

Remark 1 (Insensitivity to the Gaussianity assumptions) *To make the formulation in (5) perfectly matches the statistical properties of the signals, $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ have been assumed Gaussian. However, this does not imply that the proposed estimator breaks down immediately when the statistical assumptions are violated. This is in analogy with the fact that the Tikhonov regularization works well in a wide range of situations. We actually tested a case when $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ are uniformly distributed, and we observed that SORR performed equally well to the Gaussian case under appropriate tuning of ρ, which has no direct link to the SNR any more in such a non-Gaussian case. (The results will be reported elsewhere.) In our recent work, moreover, SORR has been extended to the case when $\boldsymbol{x}_\star$ is a sparse vector* [12]*. Finally, it is straightforward to extend SORR to anisotropic Gaussian distributions.*

C. Convexity condition

Let $\boldsymbol{\xi} := [\boldsymbol{x}^\top\ \boldsymbol{\varepsilon}^\top] \in \mathbb{R}^{n+m}$, $\boldsymbol{\Theta} := [\boldsymbol{A}\ \boldsymbol{I}_m] \in \mathbb{R}^{m\times(n+m)}$, and $\boldsymbol{Q} := \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{O}_{n\times m} \\ \boldsymbol{O}_{m\times n} & \rho^{1/2}\boldsymbol{I}_m \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$. The proposed formulation in (6) can then be rewritten as

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^{n+m}} J(\boldsymbol{\xi}) := \frac{1}{2}\|\boldsymbol{Q}\boldsymbol{\xi}\|_2^2 + \widetilde{\mu}\,\Phi_\gamma^{\mathrm{MC}}(\boldsymbol{\Theta}\boldsymbol{\xi} - \boldsymbol{y}). \quad (7)$$

[1] The figure may suggest a possible relation between the MC loss and Hampel's three part redescending function (the piecewise linear derivative) [4]. However, the MC loss is *not* a special case of Hampel's function (nor its limit).

[2] A related formulation has been presented in [10] in the context of robust recovery of jointly sparse signals, involving the MC loss but not discriminating noise and outlier explicitly unlike the present work.

Note here that $\Phi_{\gamma}^{\mathrm{MC}}(\boldsymbol{\Theta}\boldsymbol{\xi}-\boldsymbol{y})=\Phi_{\gamma}^{\mathrm{MC}}(\boldsymbol{y}-\boldsymbol{\Theta}\boldsymbol{\xi})$. By (3) and (7), we obtain

$$J(\boldsymbol{\xi})=\underbrace{\frac{1}{2}\left\|\boldsymbol{Q}\boldsymbol{\xi}\right\|_2^2-\widetilde{\mu}\ {}^{\gamma}\left\|\cdot\right\|_1(\boldsymbol{\Theta}\boldsymbol{\xi}-\boldsymbol{y})}_{=:F(\boldsymbol{\xi})}+\widetilde{\Psi}(\boldsymbol{\Theta}\boldsymbol{\xi}). \tag{8}$$

Here, $\widetilde{\Psi}(\boldsymbol{z}):=\widetilde{\mu}\left\|\boldsymbol{z}-\boldsymbol{y}\right\|_1$, $\boldsymbol{z}\in\mathbb{R}^m$, is clearly convex, while $F(\boldsymbol{\xi})$ is smooth and it is convex under the condition below.

Proposition 1 (Convexity condition for SORR (8)**)** *The smooth part $F(\boldsymbol{\xi})$ is convex if and only if $\widetilde{\mu}\in[0,\gamma/(\rho^{-1}+\lambda_{\max}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A}))]$.*

Remark 2 (Parameter design) *The SORR formulation in* (6) *has the three parameters $\widetilde{\mu}$, γ, and ρ. Regarding the regularization parameter $\widetilde{\mu}$, the particular choice $\widetilde{\mu}=\gamma/[\rho^{-1}+\lambda_{\max}(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{A})]$ is recommended, as it usually gives the best performance in the range given in Proposition 1. This is because a smaller $\widetilde{\mu}$ makes the effect of the Tikhonov regularization stronger, causing undesirable estimation bias.*[3] *The saturation parameter γ can be designed depending on the minimal value of "disturbances" that could be regarded as outliers. The SNR estimate ρ can be designed based on the prior knowledge about noise conditions of environments, because it is usually possible to measure only noises in advance in typical communication systems. If such prior knowledge is unavailable, it is recommended to set ρ to a reasonably large value, such as $\rho=0.1$. The reason is the following. A too small ρ diminishes the upper bound of $\widetilde{\mu}$. As a result, the regularization effects become overwhelming, and this causes unacceptably large errors, as shown in Section V.*

D. Optimization algorithm

An application of the forward-backward-based primal-dual method [8] to (7) yields the following algorithm.

Algorithm 1 (Primal-dual debiasing algorithm)

```
Set: $\boldsymbol{\xi}_0\in\mathbb{R}^{n+m}$, $\boldsymbol{v}_0\in\mathbb{R}^m$, $(\tau,\sigma)\in\mathbb{R}_{++}^2$, $\beta_k\in\mathbb{R}_{++}$
For $k=0,1,2,\cdots$, do:
    $\boldsymbol{s}_k=\boldsymbol{\xi}_k-\tau\nabla F(\boldsymbol{\xi}_k)\in\mathbb{R}^{n+m}$
    $\boldsymbol{u}_k=\boldsymbol{s}_k-\tau\boldsymbol{\Theta}^{\mathsf{T}}\boldsymbol{v}_k\in\mathbb{R}^{n+m}$
    $\boldsymbol{q}_k=\mathrm{Prox}_{\sigma\widetilde{\Psi}^*}(\boldsymbol{v}_k+\sigma\boldsymbol{\Theta}\boldsymbol{u}_k)\in\mathbb{R}^m$
    $\boldsymbol{p}_k=\boldsymbol{s}_k-\tau\boldsymbol{\Theta}^{\mathsf{T}}\boldsymbol{q}_k\in\mathbb{R}^{n+m}$
    $(\boldsymbol{\xi}_{k+1},\boldsymbol{v}_{k+1})=(\boldsymbol{\xi}_k,\boldsymbol{v}_k)+\beta_k\left[(\boldsymbol{p}_k,\boldsymbol{q}_k)-(\boldsymbol{\xi}_k,\boldsymbol{v}_k)\right]$
End
```

As the gradient of the Moreau envelope in (1) is given by $\nabla\,{}^{\gamma}f(x)=\gamma^{-1}(x-\mathrm{Prox}_{\gamma f}(x))$ which is γ^{-1}-Lipschitz continuous [9, 13], the gradient in the algorithm is given by

$$\nabla F(\boldsymbol{\xi})=\boldsymbol{Q}^2\boldsymbol{\xi}-\widetilde{\mu}\boldsymbol{\Theta}^{\mathsf{T}}[\boldsymbol{\Theta}\boldsymbol{\xi}-\boldsymbol{y}-\mathrm{Soft}_{\gamma}(\boldsymbol{\Theta}\boldsymbol{\xi}-\boldsymbol{y})]/\gamma,$$

where $\mathrm{Soft}_{\gamma}:=\mathrm{Prox}_{\gamma\|\cdot\|_1}:\mathbb{R}^m\to\mathbb{R}^m:[z_1,z_2,\cdots,z_m]^{\mathsf{T}}\mapsto[\mathrm{soft}(z_1),\mathrm{soft}(z_2),\cdots,\mathrm{soft}(z_m)]^{\mathsf{T}}$ is the shrinkage (soft thresholding) operator. Here, $\mathrm{soft}:\mathbb{R}\to\mathbb{R}:a\mapsto\mathrm{sign}(a)\max\{0,|a|-\gamma\}$ with $\mathrm{sign}(a):=1$ if $a\geq 0$; $\mathrm{sign}(a):=-1$ otherwise. By virtue of the identity $\mathrm{Prox}_{\gamma f}+\gamma\mathrm{Prox}_{f^*/\gamma}\circ\gamma^{-1}I=I$, where I denotes the identity operator, the proximity operator can be computed as

$$\mathrm{Prox}_{\sigma\widetilde{\Psi}^*}(\boldsymbol{z})=\boldsymbol{z}-\sigma\left[\boldsymbol{y}+\mathrm{Soft}_{\widetilde{\mu}\sigma^{-1}}(\sigma^{-1}\boldsymbol{z}-\boldsymbol{y})\right],\ \boldsymbol{z}\in\mathbb{R}^m.$$

The following result is immediate from [8].

Theorem 1 *Suppose that $F(\boldsymbol{\xi})$ is convex according to the condition given in Proposition 1. Then, the sequence $(\boldsymbol{\xi}_k)_{k\in\mathbb{N}}$ generated by Algorithm 1 converges to a solution of* (7) *under the following conditions: (i) $\tau\sigma\left\|\boldsymbol{\Theta}\right\|^2\in(0,1)$ and $\tau\in(0,2/(\kappa+\widetilde{\mu}\gamma^{-1}\left\|\boldsymbol{\Theta}\right\|^2))$, and (ii) $(\beta_k)_{k\in\mathbb{N}}\subset(0,1]$ and $\inf_{k\in\mathbb{N}}\beta_k\in\mathbb{R}_{++}$. Here, $\kappa:=\lambda_{\max}(\boldsymbol{Q}^2)=\max\{1,\rho\}$, and $\left\|\boldsymbol{\Theta}\right\|$ is the spectral norm which can be computed by $\left\|\boldsymbol{\Theta}\right\|^2=\lambda_{\max}(\boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathsf{T}})=\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}}+\boldsymbol{I}_m)=\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})+1$.*

Note that the other conditions given in [8] are satisfied automatically in the present case, because the function $J(\boldsymbol{\xi})$ in (7) has a minimizer as it is coercive[4], and $\mathrm{int}(\mathrm{dom}\,\widetilde{\Psi})\cap\mathrm{range}\,\boldsymbol{\Theta}=\mathbb{R}^m\neq\emptyset$. We finally mention that the computational complexity per iteration is $O(mn)$.

IV. Convexity analysis in a general form

We present our convexity analysis in a general form, which can potentially be used for different formulations and which will avoid unnecessary repetitions of deriving convexity conditions when one considers such different formulations.

A. A general model including SORR as a specific example

Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be finite-dimensional Hilbert spaces. In any of those spaces, we denote the zero vector by 0, the inner product by $\langle\cdot,\cdot\rangle$, its induced norm by $\|\cdot\|$, and the zero operator by O. The set of all proper lower-semicontinuous convex functions defined over a Hilbert space $\mathcal{H}$ is denoted by $\Gamma_0(\mathcal{H})$.[5] Define a couple of affine operators $\mathscr{A}_1:\mathcal{X}\to\mathcal{Y}:x\mapsto M_1x+c_1$ and $\mathscr{A}_2:\mathcal{X}\to\mathcal{Z}:x\mapsto M_2x+c_2$, where $(O\neq)M_1:\mathcal{X}\to\mathcal{Y}$ and $(O\neq)M_2:\mathcal{X}\to\mathcal{Z}$ are bounded linear operators, and $c_1\in\mathcal{Y}$ and $c_2\in\mathcal{Z}$ are constant vectors. Let $D:\mathcal{Z}\to\mathcal{Z}$ be a diagonal positive-definite operator; i.e., $\langle Dz,z\rangle>0$ for all $z\in\mathcal{Z}\setminus\{0\}$. Given a function $\Psi\in\Gamma_0(\mathcal{Z})$, we consider the following minimization problem:

$$\min_{x\in\mathcal{X}}\ \frac{1}{2}\left\|\mathscr{A}_1x\right\|^2+\mu\Psi_D(\mathscr{A}_2x), \tag{9}$$

where $\mu\in\mathbb{R}_{++}$ is the regularization parameter and $\Psi_D(z):=\Psi(z)-\min_{v\in\mathcal{Z}}\left(\Psi(v)+\frac{1}{2}\left\|D(z-v)\right\|^2\right)$, $z\in\mathcal{Z}$. The proposed formulation presented in (7) is reproduced by letting in (9) $\mathcal{X}:=\mathcal{Y}:=\mathbb{R}^{n+m}$, $\mathcal{Z}:=\mathbb{R}^m$, $\mu:=\widetilde{\mu}$, $\Psi:=\|\cdot\|_1$, $D:=\gamma^{-1/2}\boldsymbol{I}_m$, $\mathscr{A}_1:=M_1:=\boldsymbol{Q}$, and $\mathscr{A}_2:\boldsymbol{\xi}\mapsto\boldsymbol{\Theta}\boldsymbol{\xi}-\boldsymbol{y}$ (i.e., $M_2:=\boldsymbol{\Theta}$ and $c_2:=-\boldsymbol{y}$). Note that $\widetilde{\mu}\Psi(\mathscr{A}_2\boldsymbol{\xi})=\widetilde{\Psi}(\boldsymbol{\Theta}\boldsymbol{\xi})$.

Remark 3 (Relation to the existing model) *The model in* (9) *can be regarded as a particular example of the linearly involved generalized Moreau enhanced (LiGME) model* [14] *by expressing $\Psi_D(\mathscr{A}_2x)=\widetilde{\Psi}_D(M_2x)$ with $\widetilde{\Psi}:=\Psi(\cdot+c_2)$. Note however that we present our model in the form of* (9) *because it is suitable for robust regression. We emphasize that the convexity results to be presented in Section IV-C cannot be obtained straightforwardly from the results of* [14].

[3] In case that $\widetilde{\mu}$ exceeds the range given in Proposition 1, the regularization effect becomes weaker but the global optimality is no longer guaranteed due to nonconvexity.

[4] A function $f\in\Gamma_0(\mathcal{H})$ is *coercive* if $f(x)\to+\infty$ as $\|x\|\to+\infty$.

[5] A function $f:\mathcal{H}\to(-\infty,+\infty]$ is convex on $\mathcal{H}$ if $f(\alpha x+(1-\alpha)y)\leq\alpha f(x)+(1-\alpha)f(y)$ for all $(x,y,\alpha)\in\mathrm{dom}\,f\times\mathrm{dom}\,f\times[0,1]$, where $\mathrm{dom}\,f:=\{x\in\mathcal{H}\mid f(x)<+\infty\}$. If in addition $\mathrm{dom}\,f\neq\emptyset$, f is a *proper convex* function. A convex function $f:\mathcal{H}\to(-\infty,+\infty]$ is *lower semicontinuous* (or *closed*) on $\mathcal{H}$ if the level set $\mathrm{lev}_{\leq a}f:=\{x\in\mathcal{H}:f(x)\leq a\}$ is closed for every $a\in\mathbb{R}$.

B. Decomposition into smooth and nonsmooth parts

Due to the nonsingularity of D, it can be verified that [1]

$$\begin{aligned} \Psi_D(z) &= \Psi(z) - {}^{1}(\Psi \circ D^{-1})(Dz) && (10) \\ &= \Psi(z) - \frac{1}{2}\|Dz\|^2 + {}^{1}(\Psi^* \circ D)(Dz), && (11) \end{aligned}$$

where $\Psi^* \in \Gamma_0(\mathcal{Z}) : z \mapsto \sup_{v\in\mathcal{Z}} \langle z, v\rangle - \Psi(v)$ is *the Fenchel conjugate of* Ψ. The equality in (11) relies on the following facts: for any $f \in \Gamma_0(\mathcal{Z})$, (i) ${}^{\gamma}f + {}^{1/\gamma}(f^*) \circ \gamma^{-1} I = \frac{1}{2\gamma}\|\cdot\|^2$ for any $\gamma \in \mathbb{R}_{++}$ [9, Theorem 14.3], and (ii) $(f \circ \mathfrak{L})^* = f^* \circ (\mathfrak{L}^*)^{-1}$ for any bijective bounded linear operator $\mathfrak{L} : \mathcal{Z} \to \mathcal{Z}$ with its adjoint operator $\mathfrak{L}^*$. From (10), the optimization problem in (9) can be rewritten as

$$\min_{x\in\mathcal{X}} \underbrace{\frac{1}{2}\|\mathscr{A}_1 x\|^2 - \mu\, {}^{1}(\Psi \circ D^{-1})(D\mathscr{A}_2 x)}_{=:F(x)\ (\text{smooth})} + \underbrace{\mu\Psi(\mathscr{A}_2 x).}_{(\text{nonsmooth})} \quad (12)$$

It is clear that $\mu\Psi(\mathscr{A}_2 x)$ is convex due to the convexity of Ψ as well as the fact that convexity is preserved under affine transform [9, Proposition 8.20]. The following subsection therefore concerns convexity of the smooth part $F(x)$.

C. Convexity analysis for the smooth part

Comparing (9) and (12) under (11) yields

$$F(x) = \frac{1}{2}\|\mathscr{A}_1 x\|^2 - \frac{\mu}{2}\|D\mathscr{A}_2 x\|^2 + \mu\, {}^{1}(\Psi^* \circ D)(D\mathscr{A}_2 x). \quad (13)$$

The convexity condition of $F(x)$ is given in the following proposition, in which the sufficiency is immediate by considering convexity of the quadratic part of the expression in (13).

Proposition 2 (Convexity condition for smooth part of (12)**)**

(a) $F \in \Gamma_0(\mathcal{X})$ *if*

$$(\spadesuit) \quad M_1^* M_1 - \mu M_2^* D^2 M_2 \succeq O.$$

Here, we denote by $\mathfrak{L} \succeq O$ *to mean that a given bounded linear operator* $\mathfrak{L} : \mathcal{X} \to \mathcal{X}$ *is positive semidefinite; i.e.,* $\langle \mathfrak{L}x, x\rangle \geq 0$ *for all* $x \in \mathcal{X}$.

(b) *Let* $\Psi := \sigma_C : \mathcal{Z} \to (-\infty, +\infty] : z \mapsto \sup_{v\in C} \langle z, v\rangle$ *is the support function of a nonempty closed convex set*[6] $C \subset \mathcal{Z}$. *Then, the following statements hold.*

(i) *Given any* $x \in \mathcal{X}$, *the following equivalence holds:*

$$\begin{aligned} & F(x) = \frac{1}{2}\|\mathscr{A}_1 x\|^2 - \frac{\mu}{2}\|D\mathscr{A}_2 x\|^2 \\ \Leftrightarrow\ & {}^{1}(\sigma_C^* \circ D)(D\mathscr{A}_2 x) = 0 \\ \Leftrightarrow\ & x \in K := \{\hat{x} \in \mathcal{X} \mid D^2 \mathscr{A}_2 \hat{x} \in C\}. \end{aligned}$$

(ii) *Assume that* $\operatorname{int} K \neq \emptyset$. *Then,* $F \in \Gamma_0(\mathcal{X})$ *if and only if* ($\spadesuit$) *is satisfied.*

The following result follows immediately by Proposition 2.

Corollary 1 *Let* $\Psi := |||\cdot|||$ *be an arbitrary norm defined on* $\mathcal{Z}$, *which can be expressed alternatively as* $\Psi = \sigma_{\mathrm{lev}_{\leq 1}|||\cdot|||_*}$ *with the* dual norm $|||\cdot|||_* := \sigma_{\mathrm{lev}_{\leq 1}|||\cdot|||}$ [15, 16]. *Assume that one of the following conditions are satisfied: (i)* $c_2 = 0$, *(ii)* $\operatorname{range} M_2 = \mathcal{Z}$, *or (iii)* $\mathscr{A}_2 \hat{x} = 0$ *for some* $\hat{x} \in \mathcal{X}$. *Then,* $F \in \Gamma_0(\mathcal{X})$ *if and only if condition* ($\spadesuit$) *is satisfied.*

Proposition 2 can be verified by Corollary 1 with $\operatorname{range} \boldsymbol{\Theta} = \mathbb{R}^m \neq \emptyset$. The convexity analysis for a more general model than given in (9) is presented in [1]. Corollary 1 under the correspondences presented in the end of Section IV-A yields Proposition 1.

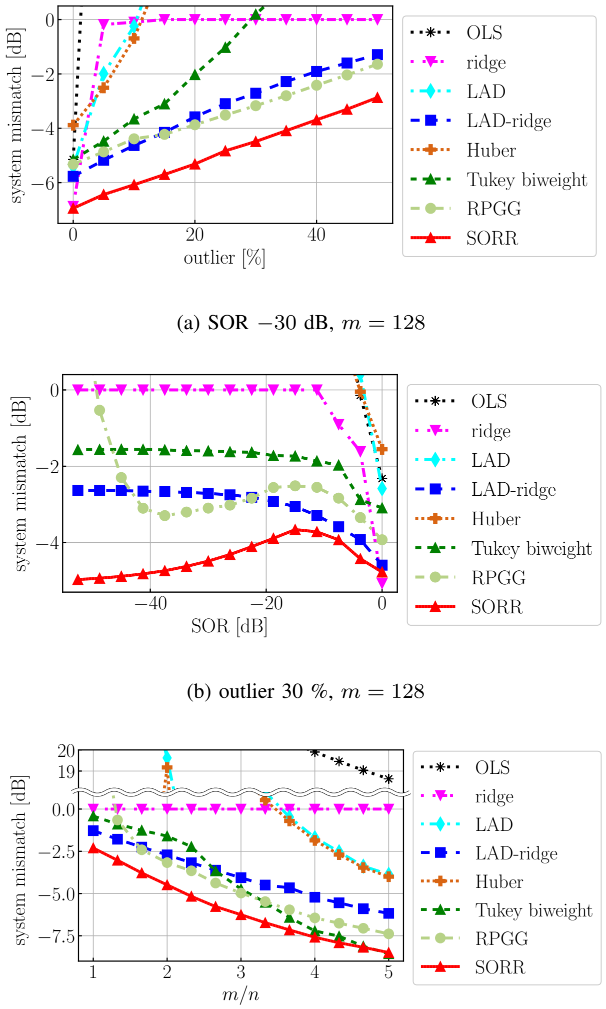
V. Numerical Examples

We compare the performance of SORR for robust regression with those of ridge regression, LAD [4], LAD-ridge (ℓ_1-loss + Tikhonov regularization), Huber's loss ${}^{\gamma}\|\cdot\|_1$ [3, 4], Tukey's biweight loss [5] for which the implementation follows [3] with LAD-ridge adopted as a "strong" initializer, and the state-of-the-art method called the robust projected generalized gradient (RPGG) algorithm [17] which is based on the following formulation[7]: $\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{e}\in\mathbb{R}^m} \mu\Phi_{\gamma_1}^{\mathrm{MC}}(\boldsymbol{x}) + \Phi_{\gamma_2}^{\mathrm{MC}}(\boldsymbol{e})$ subject to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$ for $\gamma_1, \gamma_2 \in (0, +\infty]$. We also show the performance of the ordinary least square (OLS) solution $\boldsymbol{A}^\dagger \boldsymbol{y}$ (which is an unbiased estimate) as a benchmark. The unknown vector $\boldsymbol{x}_\star \in \mathbb{R}^n$ is generated randomly from the i.i.d. standard Gaussian distribution (i.e., $\sigma_{\boldsymbol{x}}^2 := 1$) for $n := 64$, the input matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is also i.i.d. standard Gaussian, and the noise vector $\boldsymbol{\varepsilon}_\star \in \mathbb{R}^m$ is i.i.d. zero-mean Gaussian with SNR 5 dB, where $\mathrm{SNR} := \|\boldsymbol{A}\boldsymbol{x}_\star\|_2^2 / \|\boldsymbol{\varepsilon}_\star\|_2^2$. The outlier vector $\boldsymbol{o}_\diamond$ is sparse with nonzero positions chosen randomly and with nonzero components generated from an i.i.d. zero-mean Gaussian distribution. For SORR, the algorithm parameters τ and σ are set to slightly smaller values than the upper bounds, respectively, given under Algorithm 1. We simply let $\beta_k := 1$ for all $k \in \mathbb{N}$, and tune γ by grid search with $\widetilde{\mu}$ set to the upper bound given in Proposition 1. For RPGG, we let $\gamma_1 := +\infty$ (i.e., $\Phi_{\gamma_1}^{\mathrm{MC}} = \|\cdot\|_1$) as $\boldsymbol{x}_\star$ is nonsparse, and tune μ and γ_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 Huber's loss, γ is chosen to attain the best performance. The results are averaged over 300 trials.

Figure 2 plots the system mismatch $\|\boldsymbol{x}_\star - \boldsymbol{x}\|_2^2 / \|\boldsymbol{x}_\star\|_2^2$ across (a) outlier density $\operatorname{supp}(\boldsymbol{o}_\diamond)/m$, (b) signal-to-outlier ratio, $\mathrm{SOR} := (\|\boldsymbol{A}\boldsymbol{x}_\star\|_2^2/m)/[\|\boldsymbol{o}_\diamond\|_2^2/\operatorname{supp}(\boldsymbol{o}_\diamond)]$, and (c) m/n, where $\operatorname{supp}(\boldsymbol{x}) := |\{i \in \{1, 2, \cdots, m\} \mid x_i \neq 0\}|$. Here, $\rho := \sigma_{\boldsymbol{x}_\star}^2 / \sigma_{\boldsymbol{\varepsilon}_\star}^2$ is used for SORR to show its potential performance. The proposed SORR method exhibits highly accurate and stable performances, and it outperforms all the other methods significantly. Remarkably, the performances of SORR in Fig. 2(b) even improve as SOR decreases below -15 dB. This is because the influence of huge outliers on the MC loss vanishes above a certain range and because such huge outliers will be easier to detect at the same time. The results clearly indicate the remarkable robustness of SORR against huge outliers. Remarkably, Tukey's loss performs even worse than LAD-ridge (the initializer of Tukey's loss) in the low m/n (small sample) regime, while its performance is comparable to that of SORR in the high m/n (large sample) regime. It should be mentioned that LAD and Huber's loss perform poorly due to the presence of heavy noise as well as strong outliers, which make the norms of their corresponding estimates considerably large (LAD-ridge is free from such an issue owing to the Tikhonov regularization). We mention that RPGG also exhibits a similar tendency over some range, although its performance degrades below the range as noise and outlier are not explicitly discriminated in its formulation.

[6] A set $C \subset \mathcal{H}$ is said to be convex if $\alpha x + (1-\alpha)y \in C$ for all $(x, y, \alpha) \in C \times C \times [0, 1]$.

[7] Although RPGG is a method for robust sparse recovery, it could be used in the present nonsparse case by letting $\mu := 0$. We instead tune the μ 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 [17].



(a) SOR -30 dB, $m = 128$

(b) outlier 30 %, $m = 128$

(c) SOR -30 dB, outlier 30 %

Fig. 2. System mismatch for $n = 64$ under SNR 5 dB.

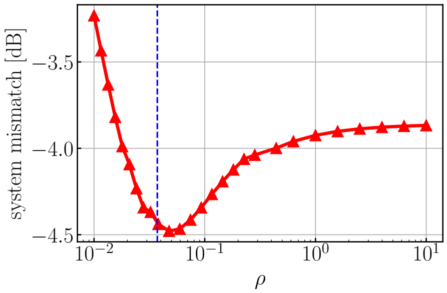


Fig. 3. Insensitivity of SORR to small fluctuations of ρ. Outlier 30%, SOR -30 dB, and $m = 128$. The blue dashed line depicts $\rho := \sigma^2_{\boldsymbol{x}_\star}/\sigma^2_{\boldsymbol{\epsilon}_\star}$.

Figure 3 shows the performance of the SORR estimator across the parameter ρ under outlier density 30%, SOR -30 dB, and $m = 128$. The results indicate reasonable insensitivity of SORR to small fluctuations of ρ. It can also be seen that the performance changes slower when ρ increases, than when it decreases, from the optimal value (see Remark 2).

VI. Concluding Remarks

We presented the SORR estimator which is robust against strong outliers and is stable under severe Gaussian noises. For stability in noisy environments, the auxiliary vector modeling the Gaussian noise was introduced to reflect the outlier sparsity and the noise Gaussianity properly in the formulation. The unknown target vector and the noise vector are thus evaluated by the squared Euclidean norms, while the outlier (corresponding to the estimation residual) is evaluated by the MC function. The rigorous analysis of convexity was presented in the general form. We showed that the proposed formulation was able to be solved via reformulation by the forward-backward-based primal-dual method under the convexity condition. The numerical examples showed that the SORR estimator was remarkably robust against large outliers and stable in the highly noisy situations at the same time. In particular, SORR significantly outperformed Tukey's biweight loss (as well as Huber's loss) in the small sample regime. This is quite advantageous particularly in high dimensional data analysis.

The SORR estimator successfully resolves the tradeoff between robustness and mathematical tractability, which was highlighted in Section I. The general result on convexity will be useful for a wide range of problems including robust classification (see [1]). It will be our interesting future work to extend the SORR estimator to the case of heavy tailed distributions.

References

[1] M. Yukawa, H. Kaneko, K. Suzuki, and I. Yamada, "Linearly-involved Moreau-enhanced-over-subspace model: debiased sparse modeling and stable outlier-robust regression," 2021, [Online]. Available: https://arxiv.org/abs/2201.03235.

[2] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 2nd ed. London: Academic Press, 2020.

[3] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust Statistics for Signal Processing*. Cambridge: Cambridge University Press, 2018.

[4] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed. Wiley, 2009.

[5] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," *Technometrics*, vol. 16, no. 2, pp. 147–185, May 1974.

[6] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, Apr. 2010.

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

[8] N. Komodakis and J.-C. Pesquet, "Playing with duality: An overview of recent primal–dual approaches for solving large-scale optimization problems," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 31–54, Nov. 2015.

[9] H. H. Bauschke and P. L. Combettes, *Convex Analysis And Monotone Operator Theory in Hilbert Spaces*, 2nd ed. New York: NY: Springer, 2017.

[10] K. Suzuki and M. Yukawa, "Robust recovery of jointly-sparse signals using minimax concave loss function," *IEEE Trans. Signal Process.*, vol. 69, pp. 669–681, 2021.

[11] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *Proc. IEEE Int. Symp. Inf. Theory*, 2010, pp. 1518–1522.

[12] K. Suzuki and M. Yukawa, "Sparse stable outlier-robust regression with minimax concave function," in *IEEE Int. Workshop on MLSP*, 2022, submitted.

[13] I. Yamada, M. Yukawa, and M. Yamagishi, *Minimizing Moreau envelope of nonsmooth convex function over the fixed point set of certain quasi-nonexpansive mappings*. in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 345–390.

[14] J. Abe, M. Yamagishi, and I. Yamada, "Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition," *Inverse Problems*, vol. 36, no. 3, pp. 1–36, Feb. 2020.

[15] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York: Cambridge University Press, 2013.

[16] S. Boyd and L.Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.

[17] C. Yang, X. Shen, H. Ma, B. Chen, Y. Gu, and H. C. So, "Weakly convex regularized robust sparse recovery methods with theoretical guarantees," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 5046–5061, Oct. 2019.