SPOQ ℓ_p-over-ℓ_q Regularization for Sparse Signal Recovery applied to Mass Spectrometry

Afef Cherni, *Member, IEEE,* Emilie Chouzenoux, *Senior member, IEEE,* Laurent Duval, *Member, IEEE,* and Jean-Christophe Pesquet, *Fellow, IEEE*

***Abstract*—Underdetermined or ill-posed inverse problems require additional information for sound solutions with tractable optimization algorithms. Sparsity yields consequent heuristics to that matter, with numerous applications in signal restoration, image recovery, or machine learning. Since the ℓ_0 count measure is barely tractable, many statistical or learning approaches have invested in computable proxies, such as the ℓ_1 norm. However, the latter does not exhibit the desirable property of scale invariance for sparse data. Extending the SOOT Euclidean/Taxicab ℓ_1-over-ℓ_2 norm-ratio initially introduced for blind deconvolution, we propose SPOQ, a family of smoothed (approximately) scale-invariant penalty functions. It consists of a Lipschitz-differentiable surrogate for ℓ_p-over-ℓ_q quasi-norm/norm ratios with $p \in]0, 2[$ and $q \geq 2$. This surrogate is embedded into a novel majorize-minimize trust-region approach, generalizing the variable metric forward-backward algorithm. For naturally sparse mass-spectrometry signals, we show that SPOQ significantly outperforms ℓ_0, ℓ_1, Cauchy, Welsch, SCAD and** $\mathrm{CEL0}$ **penalties on several performance measures. Guidelines on SPOQ hyperparameters tuning are also provided, suggesting simple data-driven choices.**

***Index Terms*—Inverse problems, majorize-minimize method, mass spectrometry, nonconvex optimization, nonsmooth optimization, norm ratio, quasinorm, sparsity.**

I. Introduction and background

A. Measuring and optimizing sparsity in data science

Sparsity classically refers to a heuristic principle [1], [2]: simplest representations or models may be better (for representation, interpretation, inference, etc.)[1]. In data sciences, it can qualify degrees of freedom for parametric models, reduce a search space, define stopping criteria, bound filter support, simplify signals or images with meaningful structures. Sparse signal recovery amounts to restoring data from inherently sparse information (e.g., spiking neurons, analytical chemistry) degraded by smoothing kernels and noise. From partial observations, it selects one solution, among all potential consistent model candidates. We review here precise formulations for sparsity measures, especially used in deconvolution, restoration or recovery. Starting with norm and quasinorm based regularizations, we will then address penalties made of ratios of those, that were recently proposed to enforce (near) scale-invariance and that will be at the core of our paper.

The natural playground for sparsity in discrete time series analysis is the space of almost-zero real sequences: $c_{00}(\mathbb{R}) = \{\mathbf{x} = (x_n)_{n\in\mathbb{N}} \mid \exists N \in \mathbb{N} \text{ such that } x_n = 0,^{\text{rld}}\ \forall n \geq N\}$. It is closed under finite addition and convolution [3, p. 597], [4, Chapter 2]. Unless stated otherwise, in the following, we consider sparse finite sequences (hence, in $c_{00}(\mathbb{R})$), each being associated with a vector $\mathbf{x} = (x_n)_{1\leq n\leq N} \in \mathbb{R}^N$. The cornerstone measure of parsimony is the count index, i.e. the number of non-zero terms in $\mathbf{x}$, denoted by $\ell_0(\mathbf{x})$. It is also called *cardinality function*, *numerosity measure*, *parsimony*, or *sparsity*. For $p \in]0, +\infty[$, we define $\ell_p^p(\mathbf{x}) = \sum_{n=1}^{N} |x_n|^p$, which is a norm for $p \geq 1$. For quasinorms ℓ_p, $p < 1$, a weaker triangle inequality holds: $\|\mathbf{x} + \mathbf{y}\| \leq K\,(\|\mathbf{x}\| + \|\mathbf{y}\|)$ with[2] $K \in [1, +\infty[$. The ℓ_p quasinorm ($p < 1$) is sometimes called p-norm-like [5]. ℓ_0 is piecewise constant, nonsmooth and nonconvex. It is often considered as unusable for data optimization in large linear systems,[3] since it leads to NP-hard problems [6], [7]. Under drastic conditions, an ℓ_0-problem can be solved exactly using a convex relation by surrogate penalties, like the ℓ_1-norm [8]. Such conditions are rarely met in practice. The use of the ℓ_1-norm yields approximate solutions and becomes more heuristic [9]. Mixed-integer programming reformulations using branch-and-bound methods [10] are possible, yet often restricted to relatively small-sized problems.

Norm- or quasinorm-based penalties have subsequently played an important role in sparse data processing or parsimonious modeling for high-dimensional regression. Squared Euclidean norm ℓ_2^2 possesses efficient implementations but often heavily degrades data sharpness. As a data fidelity term, the ℓ_2^2 cost function alone cannot address, at the same time, residual noise properties and additional data assumptions. It can be supplemented by various variational regularizations, *à la Tikhonov* [11]. Those act on the composition of data with a well-suited sparsifying operator, e.g. identity, gradients, higher-order derivatives, or wavelet frames [12]. The ℓ_2^2 penalty case corresponds to ridge regression [13]. In basis pursuit [14], feature selection [15], or *lasso* method (least

A. Cherni is with the Department of Aix-Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France.

E. Chouzenoux and J.-C. Pesquet are with Université Paris-Saclay, CentraleSupélec, Inria, Centre de Vision Numérique, 91190 Gif-sur-Yvette, France.

L. Duval is with ESIEE Paris, University Gustave Eiffel, LIGM, Noisy-le-Grand, and IFP Energies nouvelles, Rueil-Malmaison, France.

This work was partly supported by the Excellence Initiative of Aix-Marseille University — A*Midex, a French "Investissements d'Avenir" program, the Agence Nationale de la Recherche of France under MAJIC (ANR-17-CE40-0004-01) project, and the Institut Universitaire de France

[1] The law of parsimony is used in theology, history, social and empirical disciplines; Occam's razor after William of Ockham: "Entities should not be multiplied without necessity" .

[2] The lowest K, modulus of concavity of the quasinorm, saturates to 1 for norms. For $0 < p \leq 1$, $\ell_p(\mathbf{x}+\mathbf{y}) \leq 2^{\frac{1-p}{p}} \left(\ell_p(\mathbf{x}) + \ell_p(\mathbf{y})\right)$.

[3] Other denominations are *subset selection*, *minimum weight solution*, *sparse null-space*, or *minimum set cover*.

absolute shrinkage and selection operator [16]), the one-norm ℓ_1 or taxicab distance is preferred, as it promotes a form of sparsity. Solutions to the "ℓ_2^2 fidelity plus ℓ_1" regularization problem are related to total variation regularization in image restoration [17]. It can be combined with higher-order derivatives for trend filtering and source separation in analytical chemistry [18]. A convex combination of ridge and lasso regularizations yields the elastic net regularization [19], [20]. Other convex p-norms ($p \geq 1$) regularizations have been addressed, as in bridge regression which interpolates between lasso and ridge [21]. Expecting sparser solutions in practice, non-convex least-squares plus ℓ_p ($p < 1$) problems have been addressed [22]. Other non-convex metrics have also been used in sparse regression and variable selection [23], [24]. Although *a priori* appealing for sparse data restoration, such problems retain NP-hard complexities [25]. Another caveat of the above norm/quasinorm penalties, as proxies for reasonable approximations to ℓ_0, is their scale-variance: norms and quasinorms satisfy the absolute homogeneity axiom ($\ell_p(\lambda \mathbf{x}) = |\lambda| \ell_p(\mathbf{x})$, for $\lambda \in \mathbb{R}$).

B. Penalties with quasinorm and norm ratios

Better scale-invariance can be obtained by a proper normalization of quasinorms or norms penalties. We thus investigate a broader family of (quasi-)norm ratios ℓ_p-over-ℓ_q with couples $(p, q) \in]0, 2[\times [2, \infty[$, based on both their counting properties and probabilistic interpretation. First, we have equivalence relations in finite dimension:

$$\ell_q(\mathbf{x}) \leq \ell_p(\mathbf{x}) \leq \ell_0(\mathbf{x})^{\frac{1}{p} - \frac{1}{q}} \ell_q(\mathbf{x}) \tag{1}$$

with $p \leq q$, from the standard power-mean inequality [26] implying classical ℓ_p-space embeddings and generalized Rogers-Hölder's inequalities. The LHS in (1) is attained when $\mathbf{x}$ is a sparsest signal in $c_{00}(\mathbb{R})$ with a single non-zero component. The RHS is reached by a maximally non-sparse $\mathbf{x}$, with all samples set to a non-zero constant [27]. Thus, ℓ_p-over-ℓ_q quasinorm-ratios provide interesting proxies for a sparseness measure of $\mathbf{x}$, to quantify how much the “action” or “energy” of a discrete signal is concentrated into only a few of its components (*cf.* Figure 6). They are invariant under integer (circular) shift or sample shuffling in the sequence, and under non-zero scale change (or 0-degree homogeneity). Those ratios are sometimes termed *pq*-means.

For every $p \in]0, 2[$ and $q \in [2, +\infty[$, we thus define:

$$(\ell_p/\ell_q(\mathbf{x}))^p = \sum_{n=1}^{N} \left(\frac{|x_n|^q}{\sum_{n'=1}^{N} |x_{n'}|^q} \right)^{p/q} . \tag{2}$$

Interpreting the inner term

$$p_n = \frac{|x_n|^q}{\sum_{n'=1}^{N} |x_{n'}|^q} \tag{3}$$

as a unitless discrete probability distribution, ℓ_p-over-ℓ_q quasinorm-ratio rewrites as an increasing function ($u \rightarrow u^{1/p}$) of a sum of concave functions ($u \rightarrow u^{p/q}$ when $p \leq q$) of probabilities. The minimization of such an additive information cost function [28], [29], a special case of Schur-concave functionals [30], [31], is used for instance in best basis selection [32]. Thus, special cases of ℓ_p-over-ℓ_q quasinorm ratios have served as sparsity-inducing penalties in the long history of blind signal deconvolution or image deblurring, as stopping criteria (for instance in NMF, non-negative matrix factorization [33]), measures of sparsity satisfying a number of sound parsimony-prone axioms [27], estimates for time-frequency spread or concentration [34]. The ℓ_p quasinorm weighted by ℓ_2 is considered as a “possible sparseness criterion” in [35] when scaled with the normalizing factor $N^{\frac{1}{p} - \frac{1}{2}}$. It bears close connection with the kurtosis [36] for centered distributions and central moments, widely used in sparse sensory coding [37] or adaptive filtering [38].

The most frequent one with $(p, q) = (1, 2)$ is used [39], [40] as a surrogate to ℓ_0 [41], [42]. The work [40] (SOOT: Smoothed ℓ_1-over-ℓ_2 norm ratio) proposed an efficient optimization algorithm to address it with theoretically guaranteed convergence, with application to seismic signal retrieval. This ratio was also used to enhance lasso recovery on graphs [43]. Its early history includes the "minimum entropy deconvolution” proposed in [44], where the "varimax norm", akin to kurtosis $(\ell_4/\ell_2)^4$, is maximized to yield visually simpler (spikier) signals. It was inspired by simplicity measures proposed in factor analysis [45], and meant to improve one of the earliest mentioned ℓ_0 regularization [46] in seismic. The relationship with the concept of entropy was explained later [47]. It was generalized to the so-called "variable norm deconvolution" by maximizing $(\ell_q/\ell_2)^q$ [48]. Note that techniques in [44], [48] are relatively rudimentary. They aim at finding some inverse filter that maximizes a given contrast. They do not explicitly take into account noise statistics. Even more, the deconvolved estimate is linearly obtained from observations, see [49] for an overview. Recently, [50] uses ℓ_1-over-ℓ_∞ for sparse recovery, and [51] ℓ_∞-over-ℓ_0 for cardinality-penalized clustering. The family of entropy-based sparsity measures $(\ell_q/\ell_1)^{\frac{q}{1-q}}$ [52] (termed q-ratio sparsity level in [53]), extends a previous work on squared ℓ_1-over-ℓ_2 ratios [54] for compressed sensing. Finally, [55] proposes an extension of [39] to an ℓ_q-over-ℓ_2 ratio to discriminate between sharp and blurry images, and [56], [57] use a norm ratio for the purpose of impulsive signature enhancement in sparse filtering, still without rigorous convergence proofs.

C. Contribution and outline

Our main contribution resides in providing a new family of smooth-enough surrogates to ℓ_0 with sufficient Lipschitz regularity. The resulting penalties, called SPOQ, Smoothed ℓ_p-over-ℓ_q, extend the SOOT, smoothed ℓ_1-over-ℓ_2 ratio penalty from [40]. A novel trust-region algorithm generalizes and improves the variable metric forward-backward algorithm from [58] is then introduced, to minimize the sum of the SPOQ penalty and a non-necessarily smooth convex term. The convergence of our algorithm to a critical point of the problem is shown under mild technical assumptions. We then show the applicability of our method on the practical problem of sparse signal retrieval arising in mass spectrometry data analysis. In particular, we illustrate the great ability of SPOQ penalty to retrieve accurately the sought spike position and intensities, when compared to a bunch of state-of-the-art sparse recovery methods. Section I recalled the parsimony role and introduces sparsity measures. Section II describes the observation model and the proposed SPOQ quasinorm-norm ratio regularization. We then derive our trust-region minimization algorithm and analyze its convergence in Section III. Section IV illustrates the good performance of SPOQ regularization in recovering "naturally sparse" mass spectrometry signals, as compared to several sparsity penalties for restoration and recovery.

II. Proposed formulation

A. Sparse signal reconstruction

Let us consider the observation model

$$\mathbf{y} = \boldsymbol{D}\mathbf{x} + \mathbf{b} \tag{4}$$

where $\mathbf{y} = (y_m)_{1\leq m\leq M} \in \mathbb{R}^M$ represents the degraded measurements related to the original signal $\mathbf{x} = (x_n)_{1\leq n\leq N} \in \mathbb{R}^N$ through the observation matrix $\boldsymbol{D} \in \mathbb{R}^{M\times N}$. Hereabove, $\mathbf{b} \in \mathbb{R}^M$ models additive acquisition noise. In this work, we focus on the inverse problem aiming at recovering signal $\mathbf{x}$ from $\mathbf{y}$ and $\boldsymbol{D}$, under the assumption that the sought signal is sparse, i.e., has few non-zero entries. A direct (pseudo) inversion of $\boldsymbol{D}$ generally yields poor-quality solutions, because of noise and the ill-conditioning of $\boldsymbol{D}$. More suitable is a penalized approach, which defines an estimate $\hat{\mathbf{x}} \in \mathbb{R}^N$ of $\mathbf{x}$ as a solution of the constrained minimization problem

$$\underset{\mathbf{x}\in\mathcal{C}}{\text{minimize}} \;\; \Psi(\mathbf{x}) + \Theta(\mathbf{x}), \tag{5}$$

where $\mathcal{C}$ is a non-empty convex and compact subset of $\mathbb{R}^N$. Moreover, function $\Theta : \mathbb{R}^N \to]-\infty, +\infty]$ is a non-necessarily smooth convex data fidelity function measuring the discrepancy between the observation and the model. One can define Θ as the least-squares function $\zeta : \mathbf{x} \mapsto \xi\|\boldsymbol{D}\mathbf{x} - \mathbf{y}\|^2$, or adopt an interesting constrained formulation, by setting

$$(\forall \mathbf{x} \in \mathbb{R}^N) \quad \Theta(\mathbf{x}) = \iota_{\mathcal{B}_\xi^{\mathbf{y}}}(\boldsymbol{D}\mathbf{x}). \tag{6}$$

Hereabove, $\xi > 0$ is a parameter depending on noise characteristics, $\mathcal{B}_\xi^{\mathbf{y}}$ is the Euclidean ball centered at $\mathbf{y}$ with radius ξ, and $\iota_{\mathcal{S}}$ denotes the indicator function of a set $\mathcal{S}$, equal to zero for $\mathbf{x} \in \mathcal{S}$, and $+\infty$ otherwise. Furthermore, function Ψ defined from $\mathbb{R}^N \to]-\infty, +\infty]$ is a regularization function used to enforce desirable properties on the solution. The choice of Ψ is essential for reaching satisfying results by promoting desirable properties in the sought signal. When sparsity is expected, the ℓ_1 norm is probably the most used regularization function. It is a convex envelope proxy to ℓ_0, a key feature for deriving efficient minimizations to solve (5). However, because it is not scale invariant, the ℓ_1 penalty can lead to an under-estimation bias of signal amplitudes, which is detrimental to the quality of the solution. In this work, we propose a new regularization strategy, relying on the ℓ_p-over-ℓ_q norm ratio, aiming at limiting scale ambiguity in the estimation.

B. Proposed SPOQ penalty

Let $p \in]0, 2[$ and $q \in [2, +\infty[$. We first define two smoothed approximations to ℓ_p and ℓ_q parametrized by constants $(\alpha, \eta) \in]0, +\infty[^2$: for every $\mathbf{x} = (x_n)_{1\leq n\leq N} \in \mathbb{R}^N$,

$$\ell_{p,\alpha}(\mathbf{x}) = \left(\sum_{n=1}^{N}\left(\left(x_n^2 + \alpha^2\right)^{p/2} - \alpha^p\right)\right)^{1/p} \tag{7}$$

and

$$\ell_{q,\eta}(\mathbf{x}) = \left(\eta^q + \sum_{n=1}^{N}|x_n|^q\right)^{1/q}. \tag{8}$$

Remark that the traditional p and q (quasi-)norms are recovered for $\alpha = \eta = 0$. Our Smoothed ℓ_p-over-ℓ_q (SPOQ) penalty is then defined as the following function:

$$\Psi(\mathbf{x}) = \log\left(\frac{(\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p)^{1/p}}{\ell_{q,\eta}(\mathbf{x})}\right). \tag{9}$$

Parameter $\beta \in]0, +\infty[$ is introduced to account for the fact that the log function is not defined at 0. It is worth noting that the proposed function (9) combines both the sparsity promotion effect of the non-convex logarithmic loss [59], [60] and of the ℓ_p-over-ℓ_q ratio. It generalizes the Euclidean/Taxicab Smoothed ℓ_1-over-ℓ_2 (SOOT) penalty [40], recovered for $p = 1$ and $q = 2$. Figure 1 illustrates the shape of the SPOQ penalty in the case $N = 2$, for $p = 1$ and $q = 2$ (i.e., SOOT), and $p = 1/4$ and $q = 2$, in comparison with ℓ_0, ℓ_1, SCAD, and CEL0. It is worth noticing that, on the second row, the logarithm sharpens the ℓ_1-over-ℓ_2 behavior toward ℓ_0. By choosing $p = 1/4$, SPOQ further enhances the folds along the axes. As a result, the bottom-right picture best mimics the top-left ℓ_0 representation.

C. Mathematical properties

We present here several properties of the proposed SPOQ penalty, that will be essential for deriving an efficient optimization algorithm to solve Problem (5).

1) Gradient and Hessian: Let us express the gradient and Hessian matrices of function Ψ at $\mathbf{x} \in \mathbb{R}^N$:

$$\begin{cases} \nabla \ell_{q,\eta}^q(\mathbf{x}) = q\,\left(\operatorname{sign}(x_n)|x_n|^{q-1}\right)_{1\leq n\leq N} \\ \nabla^2 \ell_{q,\eta}^q(\mathbf{x}) = q(q-1)\ \operatorname{Diag}\left((|x_n|^{q-2})_{1\leq n\leq N}\right) \end{cases} \tag{10}$$

and

$$\begin{cases} \nabla \ell_{p,\alpha}^p(\mathbf{x}) = p\,\left(x_n(x_n^2+\alpha^2)^{\frac{p}{2}-1}\right)_{1\leq n\leq N} \\ \nabla^2 \ell_{p,\alpha}^p(\mathbf{x}) = p\ \operatorname{Diag}\Big((((p-1)x_n^2 + \alpha^2) \\ \qquad\qquad \times (x_n^2+\alpha^2)^{\frac{p}{2}-2})_{1\leq n\leq N}\Big), \end{cases} \tag{11}$$

with the notation $\operatorname{sign}(x) = 0$ for $x = 0$, -1 for $x < 0$ and $+1$ for $x > 0$. It is worth noting that one can decompose the SPOQ penalty under the form

$$\Psi(\mathbf{x}) = \Psi_1(\mathbf{x}) - \Psi_2(\mathbf{x}), \tag{12}$$

by setting

$$\Psi_1(\mathbf{x}) = \frac{1}{p}\log\left(\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p\right), \tag{13}$$

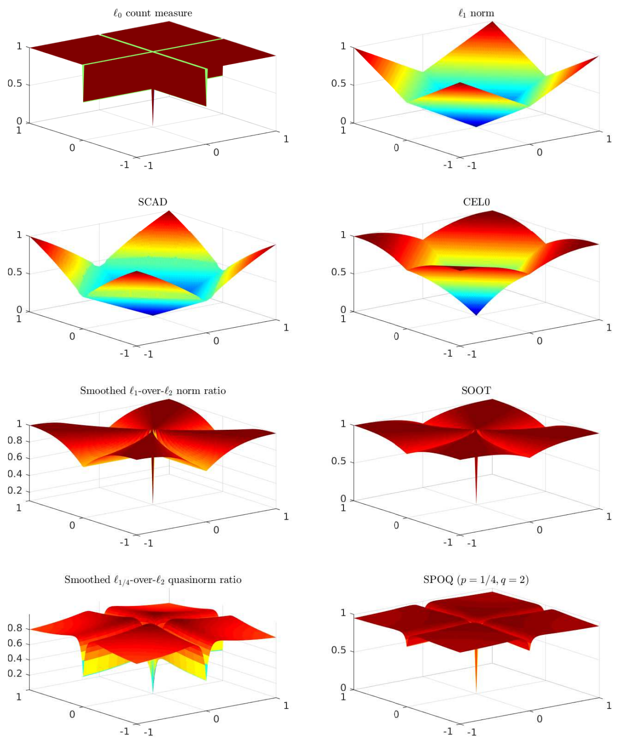

Figure 1. Sparsity-promoting penalties, scaled to [0, 1]. From top to bottom and from left to right: ℓ_0, ℓ_1, SCAD (with $\delta = 2.5, a = 1.1$), CEL0 (with $\delta = 1$), smoothed ℓ_1-over-ℓ_2, SOOT, smoothed ℓ_p-over-ℓ_q and SPOQ with $(\alpha, \beta, \eta) = (7 \times 10^{-7}, 3 \times 10^{-3}, 1 \times 10^{-1})$.

and

$$\Psi_2(\mathbf{x}) = \frac{1}{q} \log \left(\ell_{q,\eta}^q(\mathbf{x}) \right). \tag{14}$$

Hence, $\nabla\Psi = \nabla\Psi_1 - \nabla\Psi_2$, and $\nabla^2\Psi = \nabla^2\Psi_1 - \nabla^2\Psi_2$, with

$$\nabla\Psi_1(\mathbf{x}) = \frac{1}{p} \frac{\nabla \ell_{p,\alpha}^p(\mathbf{x})}{\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p}, \tag{15}$$

$$\nabla\Psi_2(\mathbf{x}) = \frac{1}{q} \frac{\nabla \ell_{q,\eta}^q(\mathbf{x})}{\ell_{q,\eta}^q(\mathbf{x})}, \tag{16}$$

$$p\nabla^2\Psi_1(\mathbf{x}) = \frac{\nabla^2 \ell_{p,\alpha}^p(\mathbf{x})}{\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p} - \frac{\nabla \ell_{p,\alpha}^p(\mathbf{x}) \left(\nabla \ell_{p,\alpha}^p(\mathbf{x})\right)^\top}{\left(\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p\right)^2}, \tag{17}$$

$$q\nabla^2\Psi_2(\mathbf{x}) = \frac{\nabla^2 \ell_{q,\eta}^q(\mathbf{x})}{\ell_{q,\eta}^q(\mathbf{x})} - \frac{\nabla \ell_{q,\eta}^q(\mathbf{x}) \left(\nabla \ell_{q,\eta}^q(\mathbf{x})\right)^\top}{\ell_{q,\eta}^{2q}(\mathbf{x})}. \tag{18}$$

From the above, we derive the following proposition, stating that for suitable parameter choices, function Ψ has $\mathbf{0}_N$, i.e. the zero vector of dimension N, as a minimizer, which is desirable for a sparsity promoting regularization function.

Proposition 1. Assume that either $q = 2$ and $\eta^2\alpha^{p-2} > \beta^p$, or $q > 2$. Then, $\nabla^2\Psi(\mathbf{0}_N)$ is a positive definite matrix and $\mathbf{0}_N$ is a local minimizer of Ψ. In addition, if

$$\eta^2 \geq \beta^2 \max \left\{ \frac{8\alpha^{2-p}}{p(2+p)\beta^{2-p}}, \frac{1}{(2^{p/2} - 1)^{2/p}} \right\} \tag{19}$$

then $\mathbf{0}_N$ is a global minimizer of Ψ.

Proof: See Appendix A. ■

2) *Majorization properties:* We now gather in the following proposition two properties that allow us to build quadratic surrogates for Function (9).

Proposition 2. Let Ψ be defined by (9).
(i) Ψ is a L-Lipschitz differentiable function on $\mathbb{R}^N$, *i.e*, for every $(\mathbf{x}, \mathbf{x}') \in (\mathbb{R}^N)^2$,

$$\|\nabla\Psi(\mathbf{x}) - \nabla\Psi(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\| \tag{20}$$

where

$$L = p\frac{\alpha^{p-2}}{\beta^p} + \frac{p}{2\alpha^2} \max \left\{ 1, \left(\frac{N\alpha^p}{\beta^p} \right)^2 \right\} + \frac{q-1}{\eta^2}. \tag{21}$$

In particular,

$$\Psi(\mathbf{x}') \ \leq \ \Psi(\mathbf{x}) + (\mathbf{x}' - \mathbf{x})^\top \nabla\Psi(\mathbf{x}) + \frac{L}{2}\|\mathbf{x}' - \mathbf{x}\|^2. \tag{22}$$

(ii) For every $\rho \in [0, +\infty[$, define the ℓ_q-ball complement:

$$\overline{\mathcal{B}}_{q,\rho} = \{\mathbf{x} = (x_n)_{1 \leq n \leq N} \in \mathbb{R}^N \mid \sum_{n=1}^N |x_n|^q \geq \rho^q\}. \tag{23}$$

Ψ admits a quadratic tangent majorant at every $\mathbf{x} \in \overline{\mathcal{B}}_{q,\rho}$, i.e.

$$\begin{aligned}(\forall \mathbf{x}' \in \overline{\mathcal{B}}_{q,\rho}) \quad \Psi(\mathbf{x}') &\leq \Psi(\mathbf{x}) + (\mathbf{x}' - \mathbf{x})^\top \nabla\Psi(\mathbf{x}) \\ &+ \frac{1}{2}(\mathbf{x}' - \mathbf{x})^\top \boldsymbol{A}_{q,\rho}(\mathbf{x})(\mathbf{x}' - \mathbf{x}),\end{aligned} \tag{24}$$

where

$$\begin{aligned}\boldsymbol{A}_{q,\rho}(\mathbf{x}) &= \chi_{q,\rho} \boldsymbol{I}_N \\ &+ \frac{1}{\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p} \operatorname{Diag}\left((x_n^2 + \alpha^2)^{p/2-1} \right)_{1 \leq n \leq N},\end{aligned} \tag{25}$$

with

$$\chi_{q,\rho} = \frac{q-1}{(\eta^q + \rho^q)^{2/q}}. \tag{26}$$

Moreover, for every $\mathbf{x} \in \mathbb{R}^N$,

$$\chi_{q,\rho} \boldsymbol{I}_N \preceq \boldsymbol{A}_{q,\rho}(\mathbf{x}) \preceq (\chi_{q,\rho} + \beta^{-p}\alpha^{p-2}) \boldsymbol{I}_N. \tag{27}$$

Proof: See Appendix B. ■

Proposition 2(i) leads to a rather simple majorizing function for Ψ, valid on the whole Euclidean space $\mathbb{R}^N$. This extends our previous result established in [40] for the particular case when $p = 1$ and $q = 2$. The majorization property presented in Proposition 2(ii) only holds in the non-convex set $\overline{\mathcal{B}}_{q,\rho}$. By limiting the size of the region where majorization is imposed, one may expect more accurate approximations for Ψ. This observation motivates the trust-region minimization algorithm that we propose in the next section to solve Problem (5).

III. Minimization algorithm

A. Preliminaries

We first introduce some key notation and concepts. Problem (5) can be rewritten equivalently as:

$$\underset{\mathbf{x} \in \mathbb{R}^N}{\text{minimize}} \ \ \Omega(\mathbf{x}) \tag{28}$$

where $\Omega = \Psi + \Phi$ with Ψ defined in (9) and $\Phi = \Theta + \iota_{\mathcal{C}}$. We will assume that function Θ belongs to $\Gamma_0(\mathbb{R}^N)$, the class of convex lower semi-continuous functions. This is for instance valid for the least-squares term as well as for function (6). Note that the assumptions made on set $\mathcal{C}$ implies that Φ is coercive and belongs to $\Gamma_0(\mathbb{R}^N)$. The particular structure of Ω, summing a Lipschitz differentiable function Ψ and the non-necessarily smooth convex term Φ suits it well to the class of variable metric forward-backward (VMFB) optimization methods [61], [62], [58], [63]. In such methods, one alternates gradient steps on Ψ and proximity steps on Φ, preconditioned by a specific sequence of metric matrices. Let us recall that, for $\Phi \in \Gamma_0(\mathbb{R}^N)$, and for a symmetric positive definite (SPD) metric matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$, the proximity operator of Φ at $\mathbf{x} \in \mathbb{R}^N$ relative to the metric matrix $\boldsymbol{A}$ is defined as

$$\operatorname{prox}_{\boldsymbol{A},\Phi}(\mathbf{x}) = \underset{\mathbf{z}\in\mathbb{R}^N}{\operatorname{argmin}} \left(\frac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2_{\boldsymbol{A}} + \Phi(\mathbf{z})\right) \tag{29}$$

with notation $\|\mathbf{u}\|_{\boldsymbol{A}} = \sqrt{\mathbf{u}^\top \boldsymbol{A}\mathbf{u}}$ for $\mathbf{u} \in \mathbb{R}^N$. Then, the VMFB method for solving Problem (28) reads, for every $k \in \mathbb{N}$,

$$\mathbf{x}_{k+1} = \operatorname{prox}_{\gamma_k^{-1}\boldsymbol{A}_k,\Phi}\left(\mathbf{x}_k - \gamma_k(\boldsymbol{A}_k)^{-1}\nabla\Psi(\mathbf{x}_k)\right), \tag{30}$$

where $\mathbf{x}_0 \in \mathbb{R}^N$, and $(\gamma_k)_{k\in\mathbb{N}}$ and $(\boldsymbol{A}_k)_{k\in\mathbb{N}}$ are sequences of positive stepsizes and SPD metric matrix, respectively, chosen in such a way to guarantee the convergence of VMFB iterates to a solution to Problem (5) [58]. Two main challenges arise, when implementing the VMFB algorithm, namely (i) the choice for the preconditioning matrices $(\boldsymbol{A}_k)_{k\in\mathbb{N}}$, and (ii) the evaluation of the proximity operator involved in the update (30). In [58], a novel methodology was proposed based on the choice of preconditioning matrices satisfying a majorization condition for Ψ. This methodology provides a practically efficient algorithm. Furthermore, it allows to establish convergence in the case of a non necessarily convex function Ψ, as soon as it satisfies the so-called Kurdyka-Łojasewiecz inequality [64]. Convergence also holds when the proximity update is subject to numerical errors. These advantages are particularly beneficial in our context, as our SPOQ penalty Ψ is non-convex, and the data fidelity term Φ may have a non closed form for its proximity operator (for instance, in the case of (6)). As shown in Proposition 2, function Ψ is Lipschitz differentiable and thus a constant metric could be used in our implementation of VMFB, then reduced to the standard forward-backward (FB) scheme. However, FB algorithm sometimes exhibit poor convergence speed performance. In particular, it is clear that L — the Lipschitz constant defined in (21), albeit an upper bound — can become very high for small parameters (α, β, η), which is actually the case of interest as they only act as smoothing constants for ensuring differentiability. As shown in Proposition 2(ii), it is possible to build a more accurate quadratic majorizing approximation on Ψ, whose curvature depends on the point it is calculated. However, the majorization in that case holds only on a subset of $\mathbb{R}^N$. We extend [58] with a trust-region scheme in order to use local majorizing metrics. Without deteriorating the convergence guarantees of the original method, this gives rise to a novel preconditioned proximal gradient scheme adapted at each iteration.

B. Proposed algorithm

In this section, we present our Algorithm 1 called trust-region (TR) VMFB, for solving Problem (28). At each iteration $k \in \mathbb{N}$, we will make $B \geq 1$ trials of values $(\rho_{k,i})_{k\in\mathbb{N},1\leq i\leq B}$ for the trust-region radius. For each tested radius $\rho_{k,i} \geq 0$, a VMFB update $\mathbf{z}_{k,i}$ is computed within the majorizing metric matrix $\boldsymbol{A}_{q,\rho_{k,i}}$ at $\mathbf{x}_k$, defined in Proposition 2(ii). Then, a test is performed for checking whether the update does belong to the region $\overline{\mathcal{B}}_{q,\rho_{k,i}}$. If not, the region size is reduced with a factor $\theta \in]0,1[$, and a new VMFB step is performed. The trust-region loop stops as soon as $\mathbf{z}_{k,i} \in \overline{\mathcal{B}}_{q,\rho_{k,i}}$. Note that, for the last trial, i.e. $i = B$, a radius equal to 0 is tested, which allows us to guarantee that our method is well-defined. More precisely, this leads to the following sequence, for the radius values:

$$\rho_{k,i} = \begin{cases} \sum_{n=1}^N |x_{n,k}|^q & \text{if} \quad i = 1 \\ \theta\rho_{k,i-1} & \text{if} \quad 2 \leq i \leq B-1 \\ 0 & \text{if} \quad i = B. \end{cases} \tag{31}$$

Let us remark that $\mathbf{x}_k \in \overline{\mathcal{B}}_{q,\rho_{k,1}}$ and the following inclusion holds by construction:

$$\overline{\mathcal{B}}_{q,\rho_{k,1}} \subset \overline{\mathcal{B}}_{q,\rho_{k,2}} \cdots \subset \overline{\mathcal{B}}_{q,\rho_{k,B}} = \mathbb{R}^N. \tag{32}$$

Thus, for every $i \in \{1,\ldots,B\}$, $\mathbf{x}_k \in \overline{\mathcal{B}}_{q,\rho_{k,i}}$ and the Proposition 2(ii), ensuring the validity of the local majorizing quadratic, can be applied.

Algorithm 1 TR-VMFB algorithm

```
Initialize:  $\mathbf{x}_0 \in \mathrm{dom}\Phi,\ B \in \mathbb{N}^*,\ \theta \in ]0,1[,\ (\gamma_k)_{k\in\mathbb{N}} \in ]0,+\infty[$
For $k = 0,1,\ldots$ :
    For $i = 1,\ldots,B$ :
        Set trust-region radius  $\rho_{k,i}$  using (31)
        Construct  $\boldsymbol{A}_{k,i} = \boldsymbol{A}_{q,\rho_{k,i}}(\mathbf{x}_k)$  using (25)
        $\mathbf{z}_{k,i} = \operatorname{prox}_{\gamma_k^{-1}\boldsymbol{A}_{k,i},\Phi}\left(\mathbf{x}_k - \gamma_k(\boldsymbol{A}_{k,i})^{-1}\nabla\Psi(\mathbf{x}_k)\right)$
        If  $\mathbf{z}_{k,i} \in \overline{\mathcal{B}}_{q,\rho_{k,i}}$ : Stop loop
    $\mathbf{x}_{k+1} = \mathbf{z}_{k,i}$
```

As already mentioned, the computation of the proximity operator of Φ within a general SPD metric matrix cannot usually be performed in a closed form, and an inner solver is required. In order to encompass this situation, we propose in Algorithm 2 an inexact form of our TR-VMFB method. The precision for the computation of the proximity update is measured by means of two inequalities, Alg. 2(a) and Alg. 2(b). In practice, an inner solver such as [65], [66] can be employed to compute this update. Moreover, the outer loop of TR-VMFB would be ran until a given stopping criterion is met, as we will explain in Section IV-B.

C. Convergence analysis

In this section, we show that Algorithm 1 can be viewed as a special instance of Algorithm 2 provided that κ is chosen large enough. Moreover, we establish a descent lemma for Algorithm 2, that allows us to deduce its convergence

Algorithm 2 TR-VMFB algorithm — Inexact form

```
Initialize:  $\mathbf{x}_0 \in \mathrm{dom}\Phi,\ B \in \mathbb{N}^*,\ \theta \in ]0,1[,\ (\gamma_k)_{k\in\mathbb{N}} \in ]0,+\infty[$
For $k = 0, 1, \dots$ :
    For $i = 1, \dots, B$ :
        Set trust-region radius $\rho_{k,i}$ using (31)
        Construct $\boldsymbol{A}_{k,i} = \boldsymbol{A}_{q,\rho_{k,i}}(\mathbf{x}_k)$ using (25)
        Find $\mathbf{z}_{k,i} \in \mathbb{R}^N$ such that
        (a) $\Phi(\mathbf{z}_{k,i}) + (\mathbf{z}_{k,i} - \mathbf{x}_k)^\top \nabla\Psi(\mathbf{x}_k)$
                $+\gamma_k^{-1}\|\mathbf{z}_{k,i} - \mathbf{x}_k\|^2_{\boldsymbol{A}_{k,i}} \le \Phi(\mathbf{x}_k)$
        (b) $\|\nabla\Psi(\mathbf{x}_k) + \mathbf{r}_{k,i}\| \le \kappa\|\mathbf{z}_{k,i} - \mathbf{x}_k\|_{\boldsymbol{A}_{k,i}}$
                with $\mathbf{r}_{k,i} \in \partial\Phi(\mathbf{z}_{k,i})$ and $\kappa > 0$
        If $\mathbf{z}_{k,i} \in \overline{\mathcal{B}}_{q,\rho_{k,i}}$ : Stop loop
    $\mathbf{x}_{k+1} = \mathbf{z}_{k,i}$
```

to a solution to Problem (28). We start with the following assumptions on the sequences $(\gamma_k)_{k\in\mathbb{N}}$ and $(\boldsymbol{A}_{k,i})_{k\in\mathbb{N},1\le i\le B}$, that are necessary for our convergence analysis:

Assumption 1.
(i) There exists $(\underline{\gamma},\overline{\gamma}) \in]0,+\infty[^2$ such that for every $k \in \mathbb{N}$, $\underline{\gamma} \le \gamma_k \le 2 - \overline{\gamma}$.
(ii) There exists $(\underline{\nu},\overline{\nu}) \in]0,+\infty[^2$, such that, for every $k \in \mathbb{N}$ and for every $i \in \{1,\dots,B\}$, $\underline{\nu}\boldsymbol{I}_N \preceq \boldsymbol{A}_{k,i} \preceq \overline{\nu}\boldsymbol{I}_N$.

Remark 1. By construction, iterates $(\mathbf{x}_k)_{k\in\mathbb{N}}$ produced by Algorithms 1 and 2, belong to the domain of Φ and therefore to the set $\mathcal{C}$. This implies that sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is bounded, so that there exists $\rho_{\max} \ge 0$ such that, for every $k \in \mathbb{N}$ and $i \in \{1,\dots,B\}$, we have $\rho_{k,i} \le \rho_{\max}$. Assumption 1(ii) thus holds as a consequence of (27), by setting $\underline{\nu} = \chi_{q,\rho_{\max}}$ and $\overline{\nu} = \chi_{q,0} + \beta^{-p}\alpha^{p-2}$.

The following lemma establishes the link between Algorithm 1 and its inexact form, Algorithm 2.

Lemma 1.
Under Assumption 1, for every $i \in \{1,\dots,B\}$, there exist $\mathbf{r}_{k,i} \in \partial\Psi(\mathbf{z}_{k,i})$ such that conditions Alg. 2(a) and Alg. 2(b) are fulfilled, with

$$\mathbf{z}_{k,i} = \mathrm{prox}_{\gamma_k^{-1}\boldsymbol{A}_{k,i},\Phi}\left(\mathbf{x}_k - \gamma_k(\boldsymbol{A}_{k,i})^{-1}\nabla\Psi(\mathbf{x}_k)\right) \tag{33}$$

and $\kappa \ge \underline{\gamma}^{-1}\sqrt{\overline{\nu}}$.

Proof: Let $k \in \mathbb{N}$ and $i \in \{1,\dots,B\}$, and set $\mathbf{z}_{k,i}$ as in (33). Due to the variational definition of the proximity operator, and the convexity of Φ, there exists $\mathbf{r}_{k,i} \in \partial\Phi(\mathbf{z}_{k,i})$ such that

$$\begin{cases}\mathbf{r}_{k,i} = -\nabla\Psi(\mathbf{x}_k) + \gamma_k^{-1}\boldsymbol{A}_{k,i}(\mathbf{x}_k - \mathbf{z}_{k,i})\\ (\mathbf{z}_{k,i} - \mathbf{x}_k)^\top\mathbf{r}_{k,i} \ge \Phi(\mathbf{z}_{k,i}) - \Phi(\mathbf{x}_k).\end{cases} \tag{34}$$

Thus, $\mathbf{z}_{k,i}$ satisfies:

$$\Phi(\mathbf{z}_{k,i}) + (\mathbf{z}_{k,i} - \mathbf{x}_k)^\top\nabla\Psi(\mathbf{x}_k) + \gamma_k^{-1}\|\mathbf{z}_{k,i} - \mathbf{x}_k\|^2_{\boldsymbol{A}_{k,i}} \le \Phi(\mathbf{x}_k) \tag{35}$$

Therefore, condition Alg. 2(a) holds. Moreover, using (34) and Assumption 1,

$$\begin{aligned}\|\mathbf{r}_{k,i} + \nabla\Psi(\mathbf{x}_k)\| &= \gamma_k^{-1}\|\boldsymbol{A}_{k,i}(\mathbf{x}_k - \mathbf{z}_{k,i})\|\\ &\le \underline{\gamma}^{-1}\sqrt{\overline{\nu}}\|\mathbf{x}_k - \mathbf{z}_{k,i}\|_{\boldsymbol{A}_{k,i}}.\end{aligned} \tag{36}$$

Hence the condition Alg. 2(b) holds for $\kappa \ge \underline{\gamma}^{-1}\sqrt{\overline{\nu}}$. ■

We now establish a descent property on the sequence generated by our method.

Lemma 2.
Under Assumption 1, there exists $\mu \in]0,+\infty[$ such that, for every $k \in \mathbb{N}$,

$$\Omega(\mathbf{x}_{k+1}) \le \Omega(\mathbf{x}_k) - \frac{\mu}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \tag{37}$$

with $(\mathbf{x}_k)_{k\in\mathbb{N}}$ defined in Algorithm 2.

Proof: We have

$$(\forall k \in \mathbb{N}) \quad \Omega(\mathbf{x}_{k+1}) = \Psi(\mathbf{x}_{k+1}) + \Phi(\mathbf{x}_{k+1}) \tag{38}$$

Under condition Alg. 2(a),

$$\Phi(\mathbf{x}_{k+1}) + (\mathbf{x}_{k+1} - \mathbf{x}_k)^\top\nabla\Psi(\mathbf{x}_k) + \gamma_k^{-1}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2_{\boldsymbol{A}_{k,i}} \le \Phi(\mathbf{x}_k). \tag{39}$$

By construction, $\mathbf{x}_{k+1} \in \overline{\mathcal{B}}_{q,\rho_{k,i}}$ for some $i \in \{1,\dots,B\}$. Moreover, $\mathbf{x}_k \in \overline{\mathcal{B}}_{q,\rho_{k,1}} \subset \overline{\mathcal{B}}_{q,\rho_{k,i}}$. Therefore, by Proposition 2,

$$\Psi(\mathbf{x}_{k+1}) \le \Psi(\mathbf{x}_k) + (\mathbf{x}_{k+1} - \mathbf{x}_k)^\top\nabla\Psi(\mathbf{x}_k) + \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2_{\boldsymbol{A}_{k,i}}. \tag{40}$$

Thus,

$$\begin{aligned}\Omega(\mathbf{x}_{k+1}) \le \Psi(\mathbf{x}_k) + \Phi(\mathbf{x}_k) + \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2_{\boldsymbol{A}_{k,i}}\\ - \gamma_k^{-1}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2_{\boldsymbol{A}_{k,i}},\\ \le \Omega(\mathbf{x}_k) - (\gamma_k^{-1} - \frac{1}{2})\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2_{\boldsymbol{A}_{k,i}}.\end{aligned} \tag{41}$$

Consequently, using Assumption 1, we deduce (37) by taking $\mu = \frac{\underline{\nu}\overline{\gamma}}{2(2-\overline{\gamma})}$. ■

Theorem 1.
If Φ is a semi-algebraic function on $\mathbb{R}^N$ and Assumption 1 holds, then the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by Algorithm 1 converges to a critical point $\hat{\mathbf{x}}$ of Ω.

Proof: Since $\mathcal{C}$ is compact, function Ω is coercive. Moreover, it belongs to an o-minimal structure including semi-algebraic functions and logarithmic function, so that it satisfies Kurdyka-Łojasiewicz inequality [64], [67]. Therefore, by using Lemma 2, and [58, Theorem 4.1], we deduce that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges to a critical point of Ω. ■

IV. Application to mass spectrometry processing

A. Problem statement

In this section, we illustrate the usefulness of the proposed SPOQ regularizer in the context of mass spectrometry (MS) data processing. MS is a fundamental technology of analytical chemistry to identify, quantify, and extract important information on molecules from pure samples and complex chemical mixtures. Thanks to its high performance and capabilities, MS is applied as a routine experimental procedure in several fields, including clinical research [68], anti-doping and proteomics

[69], metabolomics [70], biomedical and biological analyses [71], [72], diagnosis process, cancer and tumors profiling [73], food contamination detection [74].

In an MS experiment, the raw signal arising from the molecule ionization in an ion beam is measured as a function of time via Fourier Transform. A spectral analysis step is then performed leading to the so-called MS spectrum signal. It presents a set of positive-valued peaks distributed according to the charge state and the isotopic distribution of the studied molecule, generating typical patterns. The observed signal entails the determination of the most probable sample chemical composition, through the determination of the monoisotopic mass, charge state, and abundance of each present isotope.

In the particular context of proteomic analysis, the studied chemical compound contains only molecules involving carbon, hydrogen, oxygen, nitrogen, and sulfur. Thus, its isotopic pattern at a given mass and charge state can be easily synthesized, by making use of the so-called "averagine"[4] model [75], [76]. Assuming that the charge state is known and mono-valued (see [77] for the multi-charged case), we propose to express the measured MS spectrum $\mathbf{y} \in \mathbb{R}^M$ as the sparse combination of individual isotopic patterns, i.e.

$$\mathbf{y} = \sum_{n=1}^{N} x_n \mathbf{d}(m_n^{\text{iso}}, z) + \mathbf{b} \tag{42}$$

where $\mathbf{d}(m_n^{\text{iso}}, z) \in [0, +\infty[^M$ represents the mass distribution built with the "averagine" model at isotopic mass m_n^{iso} and charge z, discretized on a grid of size M, and $x_n \geq 0$ the associated weight. A non-zero value for entry x_n corresponds to the presence of monoisotope with mass m_n^{iso}. Moreover, $\mathbf{b} \in \mathbb{R}^M$ models the acquisition noise and some possible errors arising from the spectral analysis step. Let us form a dictionary matrix $\boldsymbol{D} \in \mathbb{R}^{M \times N}$ whose n-th column reads $\mathbf{d}(m_n^{\text{iso}}, z)$. Then, the above observation model (42) reads as (4), and the problem becomes the restoration of the sparse positive-valued signal $\mathbf{x}$, given $\mathbf{y}$ and $\boldsymbol{D}$. We proposed in [77] a restoration based on a penalized least squares problems with ℓ_1 prior and a primal-dual splitting minimization. In this section, we show by means of several experiments the benefits obtained by considering instead the proposed SPOQ penalty. We also perform comparisons between SPOQ and various other non-convex penalties.

B. Simulated datasets and settings

Two synthetic signals A and B, with size $N = 1000$, are used for the sought vector $\mathbf{x}$, containing P randomly selected nonzero components ($P = 48$ and $P = 94$, respectively). In both examples, the mass axis contains N regularly spaced values between $m_{\min} = 1000$ Daltons and $m_{\max} = 1100$ Daltons, and we set $M = N$. This allows us to generate the associated dictionary $\boldsymbol{D}$. The condition number of this matrix is equal to 4×10^4. The observed vector $\mathbf{y}$ is then deduced using Model (4), where the noise is assumed to be zero-mean Gaussian, i.i.d with known standard deviation σ (chosen as a given percentage of the MS spectrum maximal amplitude).

Figure 2 presents the sought isotopic distributions $\mathbf{x}$ and an example of associated MS spectra, for dataset A and B. In order to retrieve the original sparse signals, we will solve Problem (28) using Θ defined in (6) and $\mathcal{C} = [0, x_{\max}]^N$ with $x_{\max} = 10^5$. Concerning the regularization function Ψ, we will make comparisons between the ℓ_1 norm, ℓ_0, the Smoothly Clipped Absolute Deviation (SCAD) penalty $\Psi(\mathbf{x}) = \sum_{n=1}^N \psi(x_n)$ with, for every $x \in \mathbb{R}$, $\psi(x) = \delta|x|$ if $|x| < \delta$, $(2a\delta|x| - x^2 - \delta^2)/2(a-1)$ if $\delta \leq |x| < a\delta$, and $(a+1)\delta^2/2$ if $|x| \geq a\delta$ defined for $\delta > 0$ and $a > 2$ [78], the SPOQ penalty for $p \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1, 1.25, 1.5\}$ and $q \in \{2, 3, 5, 10\}$, the Cauchy penalty $\Psi(\mathbf{x}) = \sum_{n=1}^N \log(1 + x_n^2/\delta^2)$ with $\delta > 0$ [79, p. 111–112], the Welsch penalty $\Psi(\mathbf{x}) = \sum_{n=1}^N (1 - \exp(-x_n^2/\delta^2))$ with $\delta > 0$ [80], and the Continuous Exact ℓ_0 penalty (CEL0) $\Psi(\mathbf{x}) = \sum_{n=1}^N \left(\delta - \frac{\|\mathbf{d}_n\|^2}{2}\left(|x_n| - \frac{\sqrt{2\delta}}{\|\mathbf{d}_n\|}\right)^2 \mathbf{1}_{\{|x_n| < \frac{\sqrt{2\delta}}{\|\mathbf{d}_n\|}\}}\right)$ where $\delta > 0$ and $\mathbf{d}_n$ is the n-th column of $\boldsymbol{D}$ [81], [82],[5]

The resolution of (28) is performed by using the primal-dual splitting algorithm in [83], [84] in the case of ℓ_1 norm, ℓ_0, SCAD and CEL0 penalties. For Cauchy and Welsch penalties, we use the VMFB strategy, using the majorizing metrics described in [58]. Finally, in the case of SPOQ, we run our trust-region VMFB method, where we set $\theta = 0.5$, $B = 10$, and $\gamma_k \equiv 1.9$. The proximity operator of Φ within the metric matrix is computed by using the parallel proximal splitting algorithm from [65], with a maximum number of $5 \cdot 10^3$ iterations. With the exception of ℓ_1, all the tested penalization potentials are non-convex and only convergence to a local minimum can be guaranteed. In order to limit the sensitivity to spurious local minima, we initialize the optimization method using 10 iterations of primal-dual splitting algorithm with ℓ_1 penalty. All algorithms were run until the stopping criterion defined as $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leq \epsilon \|\mathbf{x}_k\|$ with $\epsilon = 10^{-4}$ is satisfied, and a maximum of 10^3 iterations. The most difficult task in this application is to estimate the support of the signal. Each of the iterative approaches presented has been evaluated with this regard. In order to avoid any bias in the estimation of the signal values, the support estimation process has been followed by a basic least squares step.

The considered non-convex regularizations depend on smoothing parameters, namely δ for Cauchy, Welsch and CEL0, and (α, β, η) for SPOQ. When not precised, hyperparameters were optimized with grid search to maximize the signal-to-noise ratio (SNR) defined as

$$\text{SNR}(\mathbf{x}, \hat{\mathbf{x}}) = 20 \log_{10}\left(\frac{\|\mathbf{x}\|_2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2}\right) \tag{43}$$

where $\hat{\mathbf{x}}$ is the estimated signal and $\mathbf{x}$ the original one. Moreover, the bound ξ in (6) is set to $\sqrt{N}\sigma$. A sensitivity analysis is performed to assess the influence of these parameters on the solution quality. For quantitative comparisons, we use the SNR defined above, the thresholded SNR metric denoted TSNR, defined as the SNR computed only on the support of the sought sparse signal, and the sparsity degree given as the number of

[4] Determining an *average amino acid* from a statistical distribution.

[5] The characteristic function is defined as $\mathbf{1}_\chi = 1$ if χ holds, 0 otherwise.

entries of the restored signal greater (in absolute value) than a given threshold (here we take 10^{-4}). Figure 2 shows the difficulty to distinguish the monoisotopic masses $\mathbf{x}$ from the MS spectrum $\mathbf{y}$, especially when different isotopic peaks are present with different intensities in the same mass region.

C. Numerical results

1) Comparison of sparse penalties: Tables I and II show the quality reconstruction of signals A and B for different regularization functions and two relative noise levels of the MS spectrum maximal amplitude ($0.1\,\%$ and $0.2\,\%$), when the SNR, TSNR and sparsity degree are averaged on 10 noise realizations. It appears that the SPOQ approach yields very good performance for a suitable choice of p and q. One can notice that the quality degrades for $p > 1$, especially for small values of q. A good compromise seems reached for $p \in \{0.75, 1\}$ and $q \in \{2, 3\}$. Though ℓ_0, ℓ_1, SCAD and CEL0 regularization functions ensure a good TSNR, SPOQ shows its clear superiority in all the experiments on this metric. In terms of SNR, it is again superior to all competitors, except on one configuration when SCAD is slightly better. However, in that case, SPOQ is able to provide a better TSNR, showing a better estimation of the sought support. This is confirmed by noticing that the estimated sparsity degree with SPOQ is, in all four experiments, the closest to the reference. Finally, Cauchy and Welsch have slightly lower performance than the other regularization methods, possibly due to the smoothing induced by parameter δ. These results prove that SPOQ can be the most efficient sparse penalty for an appropriate choice of p and q.

2) Advantage of trust-regions: Figure 3 shows the convergence profile, in terms of SNR evolution, of the trust-region VMFB algorithm (1), the VMFB algorithm and the FB algorithm, to recover datasets A and B when $p = 0.75$ and $q = 2$, for a given noise realization. Let us remind that VMFB algorithm corresponds to (30). Here, we set $\boldsymbol{A}_k = \boldsymbol{A}_{q,0}(\mathbf{x}_k)$ and $\gamma_k = 1.9$ for every $k \in \mathbb{N}$. FB algorithm is obtained by setting $\boldsymbol{A}_k = L\boldsymbol{I}_N$ and $\gamma_k = 1.9$ in (30), where L is the Lipschitz constant given by (21). It is worth noting that our trust-region VMFB algorithm converges much faster than the two other variants, which themselves behave here quite similarly. As already stated, the numerical value of the Lipschitz constant L can reach high values, when the smoothing parameters are small. For instance, $L = 1.016 \times 10^{12}$ in the example presented in Figure 3. In practice, one can notice that the potential negative effect of such large constant on the convergence speed is avoided thanks to our proposed trust-region approach. This illustrates the advantage of our local preconditioning scheme.

3) Setting SPOQ parameters: In all our tests, the smoothing parameters (δ, a) for Cauchy, Welsch, SCAD and CEL0 penalties were chosen empirically, so as to maximize the final SNR. Aside, we provide a pairwise sensitivity analysis for the α, β and η parameters (9) of SPOQ in Figure 4. We consider dataset A, for the noise level $0.1\,\%$, and the setting $p = 0.75$ and $q = 2$ as it was observed to lead to the best results in this case. One parameter being fixed, we cover a large span of orders of magnitude for the two others ($\alpha \in [10^{-7}, 10^2]$, $\beta \in [10^{-7}, 10^2]$ and $\eta \in [10^{-7}, 10^2]$). The first observation is the layered structure of both figures. This is interpreted as the notably weak interdependence of hyperparameters, which is advantageous. Secondly, the horizontal red/dark red strip, where the best SNR performance is attained, is relatively large, spanning about one order in magnitude in the tuned parameter. This suggests robustness, with tenuous performance variation through mild parameter imprecision. Thirdly, α seems to have little impact, especially when β and η are optimized. Note that we did not display the variations for fixed α as we observed that the SNR exhibits non-noticeable value variations. Parameter α essentially controls the L-Lipschitz value (21) and the derivability of $\ell_{p,\alpha}$ at 0 (see (7)).

4) Noise level influence: Using different penalties (ℓ_0, ℓ_1, Cauchy, Welsch, SCAD, CEL0 and two instances of SPOQ), we present in Figure 5 SNR values obtained from datasets A and B reconstruction at different noise levels. As expected, SNR for all methods decreases as noise intensity increases. Let us remind that the standard deviation σ in our case is expressed as a percentage of the MS spectrum maximal amplitude. A noise level greater than $0.1\,\%$ corresponds here to a quite high noise level for our datasets, and obviously leads to a deterioration of reconstruction quality. SPOQ proves its capability to ensure the best quality reconstruction in comparison with others penalties. The choice $p = 0.75$ and $q = 2$ shows its superiority over SOOT (i.e. $p = 1$ and $q = 2$) for all tested noise levels. Moreover, it is worthy to notice that, on this example, the superiority of SPOQ versus SCAD becomes more significant, as the noise level increases.

5) Sparsity level influence: Our final test consists in evaluating the performance of SPOQ penalty for various sparsity degrees. To do so, we tried out different datasets with a fixed size $N = 1000$ and different sparsity degrees $P \in \{10, 20, 48, 94, 182, 256, 323, 388\}$, generated in a similar fashion as in our datasets A and B. We make use of the SPOQ penalty with $p \in \{0.25, 0.75, 1\}$ and $q = 2$. Figure 6 presents the evolution of estimated sparsity degree. As we can see, the latter is well estimated when the signal presents a high sparsity level (Figure 6, case of $p = 0.25$ and $q = 2$, $p = 0.75$ and $q = 2$). However as P increases, the reconstruction quality of SPOQ where $p = 1$ and $q = 2$ (i.e., SOOT) tends to worsen. This confirms the interesting flexibility in setting parameter p.

V. Conclusion

SPOQ offers scale-invariant penalties, based on ratios of smoothed quasinorms and norms. These surrogates to the ℓ_0 count index are non-convex, yet possess Lipschitz regularity, that permits efficient optimization algorithms based on the majorize-minimize methodology. In particular, we propose a novel trust-region approach, that extends the variable metric forward-backward algorithm. On sparse mass-spectrometry peak signals, SPOQ outperforms other sparsity penalties for various quality metrics. Moreover, once the norm exponents are chosen, smoothing hyperparameters are easy to set. Further works include algorithmic acceleration and application to other types of sparse data processing, such as image deconvolution.

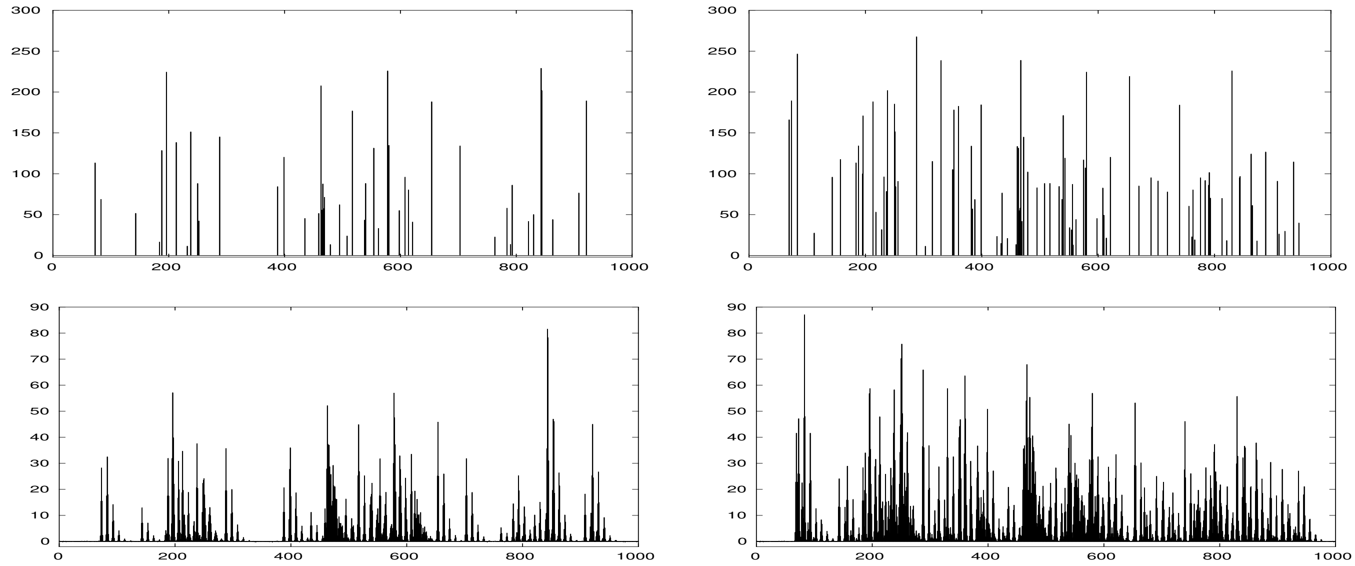

Figure 2. Original sparse signals and associated MS spectra for dataset A (left, $N = 1000$, $P = 48$) and dataset B (right, $N = 1000$, $P = 94$), top: synthetic data, bottom: noisy MS spectra ($\sigma = 0.1\%$ of the MS spectrum maximal amplitude).

		$\sigma \approx 0.081$ (0.1% of the dataset A maximum amplitude)									
		SPOQ				ℓ_0	ℓ_1	SCAD	Cauchy	Welsch	CEL0
	$p \backslash q$	2	3	5	10			$a = 2.25,\ \delta = 1$	$\delta = 100$	$\delta = 2$	$\delta = 0.5$
SNR	0.05	**52.20**	**52.20**	**52.20**	**52.20**	42.81	47.85	50.45	41.20	30.16	37.68
	0.1	**52.20**	**52.20**	**52.20**	**52.20**						
	0.15	**52.20**	**52.20**	**52.20**	**52.20**						
	0.2	**52.20**	**52.20**	**52.20**	**52.20**						
	0.25	**52.20**	**52.20**	**52.20**	**52.20**						
	0.5	**52.20**	**52.20**	**52.20**	**52.20**						
	0.75	**52.20**	**52.20**	**52.20**	**52.20**						
	1	50.40	47.63	46.15	44.87						
	1.25	29.88	32.14	34.84	38.43						
	1.5	10.05	34.14	38.19	41.27						
TSNR	0.05	**52.20**	**52.20**	**52.20**	**52.20**	47.82	49.87	50.45	44.95	42.13	46.57
	0.1	**52.20**	**52.20**	**52.20**	**52.20**						
	0.15	**52.20**	**52.20**	**52.20**	**52.20**						
	0.2	**52.20**	**52.20**	**52.20**	**52.20**						
	0.25	**52.20**	**52.20**	**52.20**	**52.20**						
	0.5	**52.20**	**52.20**	**52.20**	**52.20**						
	0.75	**52.20**	**52.20**	**52.20**	**52.20**						
	1	51.18	49.70	48.90	47.85						
	1.25	38.85	40.81	42.73	44.16						
	1.5	35.81	41.22	43.47	45.40						
Sparsity	0.05	**48**	**48**	**48**	**48**	236	77	**48**	248	545	435
	0.1	**48**	**48**	**48**	**48**						
	0.15	**48**	**48**	**48**	**48**						
	0.2	**48**	**48**	**48**	**48**						
	0.25	**48**	**48**	**48**	**48**						
	0.5	**48**	**48**	**48**	**48**						
	0.75	**48**	**48**	**48**	**48**						
	1	49	59	76	109						
	1.25	502	457	381	318						
	1.5	904	396	309	267						

		$\sigma \approx 0.163$ (0.2% of the dataset A maximum amplitude)									
		SPOQ				ℓ_0	ℓ_1	SCAD	Cauchy	Welsch	CEL0
	$p \backslash q$	2	3	5	10			$a = 2.75,\ \delta = 1.25$	$\delta = 2$	$\delta = 2$	$\delta = 0.5$
SNR	0.05	31.83	31.18	32.78	37.23	36.78	41.84	**43.20**	32.33	25.19	26.98
	0.1	31.83	31.18	34.38	38.64						
	0.15	31.83	31.18	37.23	38.66						
	0.2	31.84	31.18	38.64	38.66						
	0.25	31.85	31.18	38.64	38.66						
	0.5	33.45	38.66	39.99	39.99						
	0.75	39.66	39.68	40.02	40.02						
	1	42.72	40.32	40.06	38.70						
	1.25	25.78	24.18	27.99	31.82						
	1.5	−4.78	23.04	30.39	34.95						
TSNR	0.05	34.48	33.32	34.69	38.51	41.79	43.83	43.80	37.15	37.34	41.13
	0.1	34.48	33.32	36.08	39.74						
	0.15	34.48	33.32	38.51	39.75						
	0.2	34.44	33.32	39.74	39.75						
	0.25	34.44	33.32	39.74	39.75						
	0.5	35.83	39.75	40.86	40.86						
	0.75	40.65	40.65	40.86	40.85						
	1	**43.91**	42.68	42.82	41.76						
	1.25	34.18	33.44	36.40	38.10						
	1.5	28.95	33.96	36.65	39.32						
Sparsity	0.05	59	49	49	**48**	236	75	49	310	496	513
	0.1	59	49	**48**	**48**						
	0.15	59	49	**48**	**48**						
	0.2	59	49	**48**	**48**						
	0.25	59	49	**48**	**48**						
	0.5	59	**48**	**48**	**48**						
	0.75	**48**	**48**	**48**	49						
	1	51	64	75	118						
	1.25	386	511	410	330						
	1.5	957	480	345	273						

Table I

DATASET A ($N = 1000$, $P = 48$): COMPARISON OF SNR, TSNR AND SPARSITY DEGREE VALUES AVERAGED ON 10 NOISE REALIZATIONS USING SPOQ WITH DIFFERENT $p \in]0, 2[$ AND $q \in [2, +\infty[$ AND SOME OTHER REGULARIZATION FUNCTIONS.

SPOQ penalties are finally promising for sparsity estimation, sparse regression or feature selection.

APPENDIX A
PROOF OF PROPOSITION 1

First, we have $\nabla \ell_{p,\alpha}^{p}(\mathbf{0}_N) = \nabla \ell_{q,\eta}^{q}(\mathbf{0}_N) = \mathbf{0}_N$, so that $\mathbf{0}_N$ is a critical point of Ψ. By using (17),

$$\nabla^2 \Psi_1(\mathbf{0}_N) = \frac{\alpha^{p-2}}{\beta^p} \boldsymbol{I}_N, \tag{44}$$

with $\boldsymbol{I}_N$ identity matrix of $\mathbb{R}^N$. If $q > 2$, it follows from (18) that

$$\begin{aligned}\nabla^2 \Psi_2(\mathbf{0}_N) &= \frac{1}{q}\left(\frac{q(q-1)\,\mathrm{Diag}\left((0^{q-2})_{1 \le n \le N}\right)}{\eta^q} - \frac{q}{\eta^{2q}} \mathbf{0}_{N \times N}\right) \\ &= \mathbf{0}_{N \times N},\end{aligned} \tag{45}$$

otherwise, if $q = 2$,

$$\begin{aligned}\nabla^2 \Psi_2(\mathbf{0}_N) &= \frac{1}{2}\left(\frac{2(2-1)\,\mathrm{Diag}\left((0^0)_{1 \le n \le N}\right)}{\eta^2} - \frac{2}{\eta^{2q}} \mathbf{0}_{N \times N}\right) \\ &= \frac{1}{\eta^2} \boldsymbol{I}_N\end{aligned} \tag{46}$$

since $0^0 = 1$ by convention. Consequently

$$\nabla^2 \Psi_2(\mathbf{0}_N) = \begin{cases} \dfrac{1}{\eta^2} \boldsymbol{I}_N & \text{if } q = 2, \\ \mathbf{0}_{N \times N} & \text{elsewhere.} \end{cases} \tag{47}$$

According to these results, we deduce that $\nabla^2 \Psi(\mathbf{0}_N)$ is a positive definite matrix if ($q = 2$ and $\eta^2 \alpha^{p-2} > \beta^p$) or if $q > 2$. When these conditions are fulfilled, $\mathbf{0}_N$ is a local minimizer of Ψ.

		$\sigma \approx 0.087$ (0.1% of the dataset B maximum amplitude)									
		SPOQ				ℓ_0	ℓ_1	SCAD	Cauchy	Welsch	CEL0
	p \ q	2	3	5	10			$a = 3.75,\ \delta = 2.75$	$\delta = 100$	$\delta = 5$	$\delta = 0.5$
SNR	0.05	**51.78**	**51.78**	**51.78**	**51.78**	42.70	45.39	51.48	39.96	37.22	39.17
	0.1	**51.78**	**51.78**	**51.78**	**51.78**						
	0.15	**51.78**	**51.78**	**51.78**	**51.78**						
	0.2	**51.78**	**51.78**	**51.78**	**51.78**						
	0.25	**51.78**	**51.78**	**51.78**	**51.78**						
	0.5	**51.78**	**51.78**	**51.78**	**51.78**						
	0.75	**51.78**	**51.78**	**51.78**	**51.78**						
	1	47.82	46.08	44.13	44.35						
	1.25	32.21	31.50	34.90	38.64						
	1.5	10.25	16.71	36.46	40.36						
TSNR	0.05	**51.78**	**51.78**	**51.78**	**51.78**	46.52	47.89	51.48	43.09	44.79	46.69
	0.1	**51.78**	**51.78**	**51.78**	**51.78**						
	0.15	**51.78**	**51.78**	**51.78**	**51.78**						
	0.2	**51.78**	**51.78**	**51.78**	**51.78**						
	0.25	**51.78**	**51.78**	**51.78**	**51.78**						
	0.5	**51.78**	**51.78**	**51.78**	**51.78**						
	0.75	**51.78**	**51.78**	**51.78**	**51.78**						
	1	49.17	48.25	46.61	46.82						
	1.25	38.68	38.87	40.61	43.09						
	1.5	35.29	35.98	41.61	43.83						
Sparsity	0.05	**94**	**94**	**94**	**94**	297	157	**94**	330	413	484
	0.1	**94**	**94**	**94**	**94**						
	0.15	**94**	**94**	**94**	**94**						
	0.2	**94**	**94**	**94**	**94**						
	0.25	**94**	**94**	**94**	**94**						
	0.5	**94**	**94**	**94**	**94**						
	0.75	**94**	**94**	**94**	**94**						
	1	100	121	169	206						
	1.25	454	567	471	385						
	1.5	904	834	433	345						

		$\sigma \approx 0.174$ (0.2% of the dataset B maximum amplitude)									
		SPOQ				ℓ_0	ℓ_1	SCAD	Cauchy	Welsch	CEL0
	p \ q	2	3	5	10			$a = 2.75,\ \delta = 1.25$	$\delta = 90$	$\delta = 5$	$\delta = 0.5$
SNR	0.05	33.68	33.68	31.92	30.91	45.46	39.1028	42.51	31.64	29.1765	30.79
	0.1	33.68	33.68	32.42	31.50						
	0.15	33.87	33.87	32.54	31.56						
	0.2	33.41	33.40	32.93	32.99						
	0.25	33.58	33.93	34.46	33.39						
	0.5	40.97	40.24	42.44	36.77						
	0.75	**45.76**	**45.76**	**45.76**	41.93						
	1	41.99	39.62	38.68	38.52						
	1.25	25.87	25.16	26.48	32.23						
	1.5	0.03	3.23	27.59	33.94						
TSNR	0.05	35.39	35.39	33.49	32.62	**45.46**	41.55	43.34	34.82	37.68	41.54
	0.1	35.39	35.39	33.58	32.87						
	0.15	35.65	35.65	33.48	32.93						
	0.2	35.01	35.00	33.98	34.18						
	0.25	35.27	35.51	35.13	34.54						
	0.5	41.81	40.64	42.65	38.04						
	0.75	**45.76**	**45.76**	**45.76**	42.66						
	1	43.09	41.76	41.29	41.01						
	1.25	32.35	32.66	33.63	36.90						
	1.5	28.86	29.60	33.90	37.67						
Sparsity	0.05	95	95	**94**	**94**	**94**	155	96	330	472	531
	0.1	95	95	**94**	**94**						
	0.15	95	95	**94**	**94**						
	0.2	**94**	**94**	**94**	93						
	0.25	**94**	**94**	93	93						
	0.5	**94**	**94**	**94**	**94**						
	0.75	**94**	**94**	**94**	**94**						
	1	97	124	176	208						
	1.25	439	559	534	397						
	1.5	955	893	513	355						

Table II
DATASET B ($N = 1000$, $P = 94$): COMPARISON OF SNR, TSNR AND SPARSITY DEGREE VALUES AVERAGED ON 10 NOISE REALIZATIONS USING SPOQ WITH DIFFERENT $p \in]0, 2[$ AND $q \in [2, +\infty[$ AND SOME OTHER REGULARIZATION FUNCTIONS.

Let us now show that, under suitable assumptions,

$$(\forall \mathbf{x} \in \mathbb{R}^N) \quad \Psi(\mathbf{x}) \geq \Psi(\mathbf{0}_N) \quad \Leftrightarrow \quad \frac{(\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p)^{1/p}}{\ell_{q,\eta}(\mathbf{x})} \geq \frac{\beta}{\eta} \tag{48}$$

that is

$$\left(1 + \sum_{n=1}^{N} \frac{\alpha^p}{\beta^p}\Big(\big(\frac{z_n}{\alpha^2} + 1\big)^{p/2} - 1\Big)\right)^{2/p} \geq \left(1 + \sum_{n=1}^{N} \frac{z_n^{q/2}}{\eta^q}\right)^{2/q} \tag{49}$$

by setting, for every $n \in \{1, \ldots, N\}$, $z_n = x_n^2$. Let $\epsilon \in]0, +\infty[$. According to the second-order mean value theorem,

$$(\forall \upsilon \in [0, \epsilon]) \quad (\upsilon + 1)^{p/2} - 1 \geq \frac{p\upsilon}{2}\Big(1 - \frac{2-p}{4}\epsilon\Big). \tag{50}$$

On the other hand, since $\upsilon \mapsto \big((\upsilon + 1)^{p/2} - 1\big)/\upsilon^{p/2}$ is an increasing function on $]0, +\infty[$,

$$(\forall \upsilon \in [\epsilon, +\infty[) \quad (\upsilon + 1)^{p/2} - 1 \geq \frac{(\epsilon + 1)^{p/2} - 1}{\epsilon^{p/2}} \upsilon^{p/2}. \tag{51}$$

In the following, we will assume that $\epsilon < 4/(2-p)$. Let

$$I = \{n \in \{1, \ldots, N\} \mid z_n < \epsilon\alpha^2\} \tag{52}$$

and let $\overline{I} = \{1, \ldots, N\} \setminus I$. Since $2/p > 1$,

$$\begin{aligned}
&\left(1 + \sum_{n=1}^{N} \frac{\alpha^p}{\beta^p}\Big(\big(\frac{z_n}{\alpha^2} + 1\big)^{p/2} - 1\Big)\right)^{2/p} \\
&\geq \left(1 + \sum_{n \in I} \frac{\alpha^p}{\beta^p}\Big(\big(\frac{z_n}{\alpha^2} + 1\big)^{p/2} - 1\Big)\right)^{2/p} \\
&\quad + \left(\sum_{n \in \overline{I}} \frac{\alpha^p}{\beta^p}\Big(\big(\frac{z_n}{\alpha^2} + 1\big)^{p/2} - 1\Big)\right)^{2/p} \\
&\geq 1 + \sum_{n \in I} \frac{p\alpha^{p-2}}{2\beta^p}\Big(1 - \frac{2-p}{4}\epsilon\Big) z_n \\
&\quad + \left(\sum_{n \in \overline{I}} \frac{(\epsilon + 1)^{p/2} - 1}{\epsilon^{p/2}\beta^p} z_n^{p/2}\right)^{2/p}.
\end{aligned} \tag{53}$$

If

$$\frac{p\alpha^{p-2}}{2\beta^p}\Big(1 - \frac{2-p}{4}\epsilon\Big)\eta^2 \geq 1 \tag{54}$$

$$\frac{\big((\epsilon + 1)^{p/2} - 1\big)^{2/p}}{\epsilon\beta^2}\eta^2 \geq 1, \tag{55}$$

then

$$\left(1 + \sum_{n=1}^{N} \frac{\alpha^p}{\beta^p}\Big(\big(\frac{z_n}{\alpha^2} + 1\big)^{p/2} - 1\Big)\right)^{2/p} \geq 1 + \sum_{n \in I} \frac{z_n}{\eta^2} + \left(\sum_{n \in \overline{I}} \frac{z_n^{p/2}}{\eta^p}\right)^{2/p}. \tag{56}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSP.2020.3025731, IEEE Transactions on Signal Processing

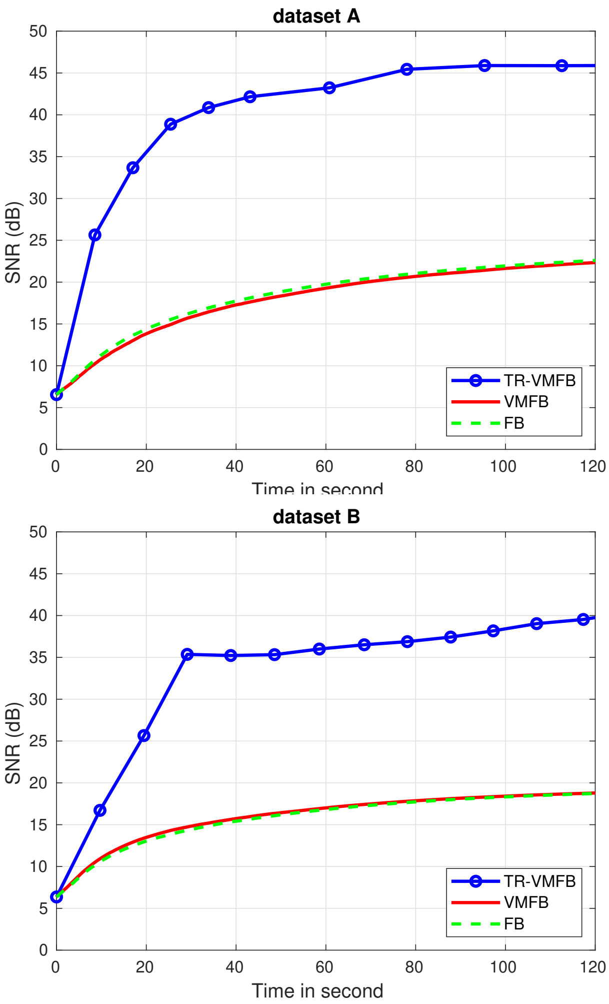

Figure 3. SNR evolution along time, for the proposed trust-region VMFB algorithm 1, VMFB algorithm [58] and FB algorithm with $L = 1.016 \times 10^{12}$, to process datasets A and B on a given noise realization (relative noise level: $0.1\,\%$).

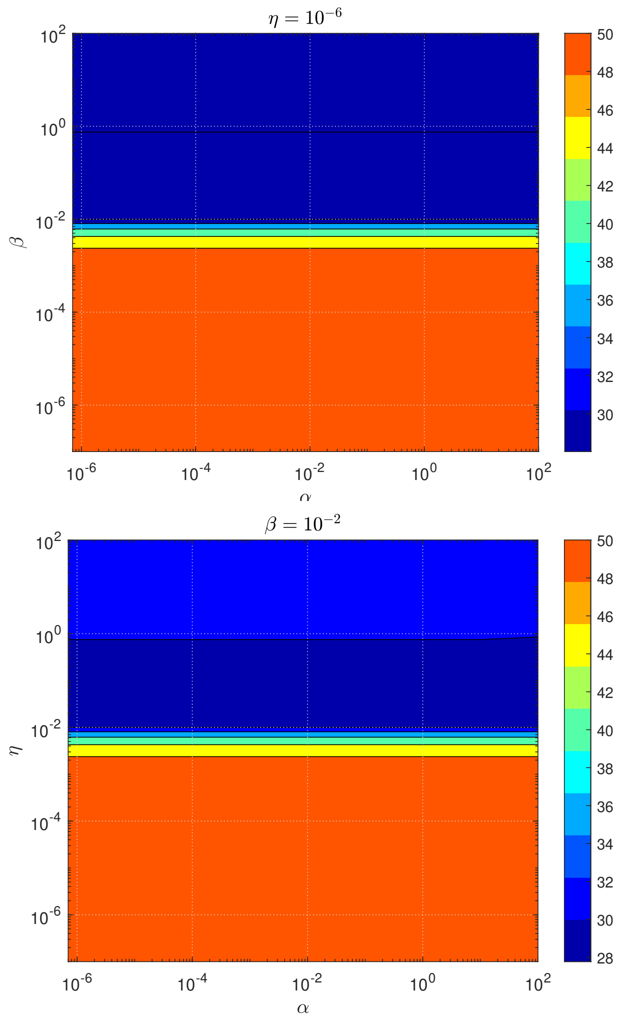

Figure 4. SNR computed for dataset A ($N = 1000$, $P = 48$) using SPOQ regularization with different α, β and η parameters where $p = 0.75$ and $q = 2$ (relative noise: $0.1\,\%$).

In addition, as $p/2 < 1 \leq q/2$,

$$\left(1+\sum_{n\in I}\frac{z_n}{\eta^2}\right)^{q/2} \geq 1+\sum_{n\in I}\frac{z_n^{q/2}}{\eta^q} \tag{57}$$

$$\left(\sum_{n\in \bar{I}}\frac{z_n^{p/2}}{\eta^p}\right)^{2/p} \geq \left(\sum_{n\in \bar{I}}\frac{z_n^{q/2}}{\eta^q}\right)^{2/q}, \tag{58}$$

where the last inequality follows from (1). This yields

$$\left(1+\sum_{n=1}^{N}\frac{\alpha^p}{\beta^p}\left(\left(\frac{z_n}{\alpha^2}+1\right)^{p/2}-1\right)\right)^{2/p} \geq \left(1+\sum_{n\in I}\frac{z_n^{q/2}}{\eta^q}\right)^{2/q} + \left(\sum_{n\in \bar{I}}\frac{z_n^{q/2}}{\eta^q}\right)^{2/q}. \tag{59}$$

We deduce Inequality (49) by applying the triangle inequality for the $\ell^{q/2}$ norm to the right-hand side of (59). The provided condition in (19) corresponds to the choice $\epsilon = 1$ in (54)-(55).

Appendix B
Proof of Proposition 2

(i) Let us first show that Ψ is Lipschitz-differentiableby investigating the properties of $\nabla^2\Psi$. We start by studying the behavior of $|||\nabla^2\Psi_1(\mathbf{x})|||$, where $\mathbf{x} \in \mathbb{R}^N$ and the spectral norm is denoted by $|||.|||$. Using (11) and (17), we obtain

$$|||\nabla^2\Psi_1(\mathbf{x})||| \leq \frac{|||\,\mathrm{Diag}\,(\mathbf{Z})|||}{\ell^p_{p,\alpha}(\mathbf{x})+\beta^p} + \frac{p\|\mathbf{Y}\|^2}{(\ell^p_{p,\alpha}(\mathbf{x})+\beta^p)^2} \tag{60}$$

where we make use of the shorter notation:

$$\begin{cases} \mathbf{Y} = \left(x_n(x_n^2+\alpha^2)^{\frac{p}{2}-1}\right)_{1\leq n\leq N} \\ \mathbf{Z} = \left(\left((p-1)x_n^2+\alpha^2\right)(x_n^2+\alpha^2)^{\frac{p}{2}-2}\right)_{1\leq n\leq N} \end{cases} \tag{61}$$

First, we have

$$\begin{aligned} &\frac{|||\,\mathrm{Diag}(\mathbf{Z})|||}{\ell^p_{p,\alpha}(\mathbf{x})+\beta^p} \\ &\quad= \frac{1}{\ell^p_{p,\alpha}(\mathbf{x})+\beta^p} \\ &\qquad\times \sup_{1\leq n\leq N}\left(\left|(p-1)x_n^2+\alpha^2\right|(x_n^2+\alpha^2)^{\frac{p}{2}-2}\right). \end{aligned} \tag{62}$$

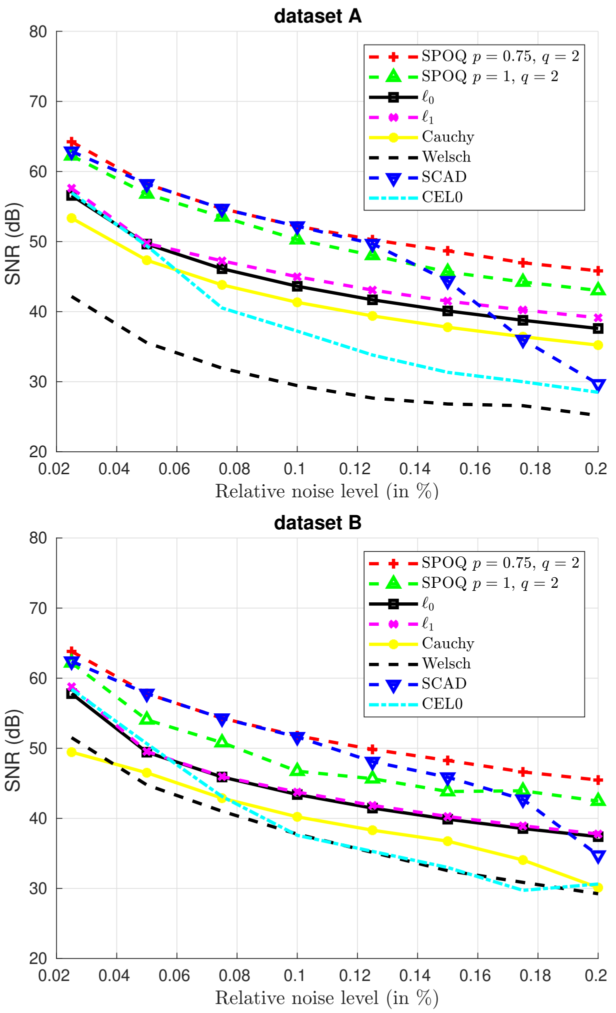

Figure 5. Influence of noise level (σ expressed as a percentage of the MS spectrum maximal amplitude) on quality reconstruction of datasets A (top) and dataset B (bottom) using various penalties: SPOQ with $p = 0.75$ and $q = 2$, SPOQ with $p = 1$ and $q = 2$ (or SOOT), ℓ_0, ℓ_1, Cauchy, Welsch, SCAD and CEL0 (SNR values averaged over 10 noise realizations).

Since $p < 2$ and, for every $n \in \{1, \dots, N\}$,

$$|(p-1)x_n^2 + \alpha^2| \leq x_n^2 + \alpha^2, \tag{63}$$

we deduce that

$$\begin{aligned} &\frac{|||\,\mathrm{Diag}(\mathbf{Z})\,|||}{\ell^p_{p,\alpha}(\mathbf{x}) + \beta^p} \\ &\quad\leq \frac{1}{\beta^p} \sup_{1\leq n\leq N} \left((p-1)(x_n^2+\alpha^2)^{\frac{p}{2}-1} + \alpha^2 (x_n^2+\alpha^2)^{\frac{p}{2}-2} \right) \\ &\quad\leq p\frac{\alpha^{p-2}}{\beta^p}. \end{aligned} \tag{64}$$

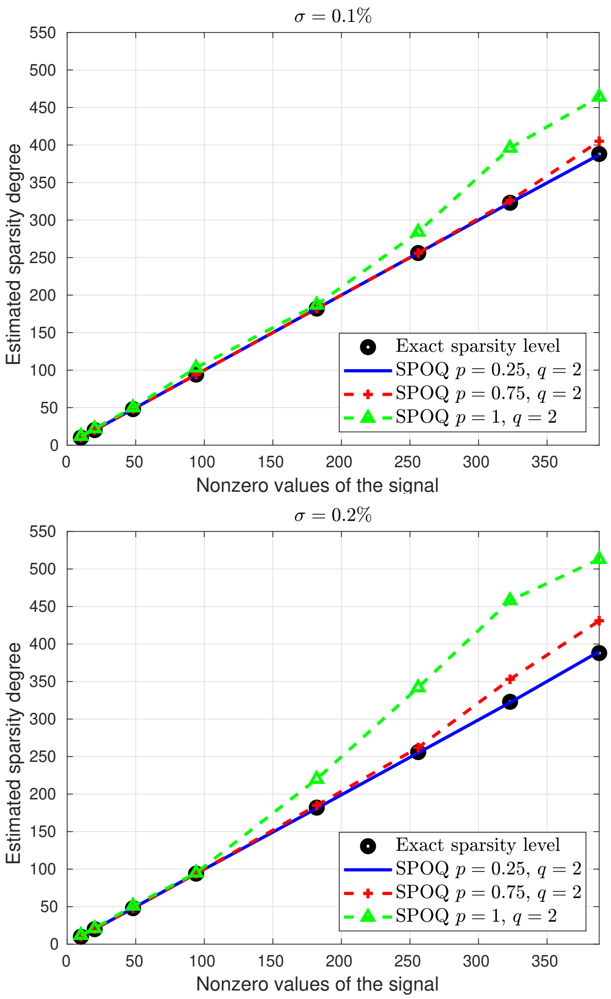

Figure 6. Estimated sparsity degree for different sparse signals using SPOQ on a single noise realization (relative noise: 0.1 % (up) and 0.2 % (bottom)).

Besides, by setting $\nu = \sum_{n=1}^{N} \left(x_n^2 + \alpha^2\right)^{p/2}$,

$$\begin{aligned} \frac{\|\mathbf{Y}\|^2}{\left(\ell^p_{p,\alpha}(\mathbf{x}) + \beta^p\right)^2} &= \frac{1}{\left(\ell^p_{p,\alpha}(\mathbf{x}) + \beta^p\right)^2} \sum_{n=1}^{N} x_n^2 (x_n^2+\alpha^2)^{p-2} \\ &\leq \frac{1}{\left(\ell^p_{p,\alpha}(\mathbf{x}) + \beta^p\right)^2} \sum_{n=1}^{N} \frac{x_n^2}{(x_n^2+\alpha^2)^2} (x_n^2+\alpha^2)^{p} \\ &\leq \frac{1}{2\alpha^2 \left(\ell^p_{p,\alpha}(\mathbf{x}) + \beta^p\right)^2} \sum_{n=1}^{N} (x_n^2+\alpha^2)^{p} \\ &\leq \frac{1}{2\alpha^2 \left(\ell^p_{p,\alpha}(\mathbf{x}) + \beta^p\right)^2} \left(\sum_{n=1}^{N} (x_n^2+\alpha^2)^{p/2} \right)^2 \\ &= \frac{\nu^2}{2\alpha^2 (\nu - N\alpha^p + \beta^p)^2} \\ &= \frac{1}{2\alpha^2} \left(1 + \frac{N\alpha^p - \beta^p}{\nu - N\alpha^p + \beta^p} \right)^2 \\ &\leq \frac{1}{2\alpha^2} \max\left\{ 1, \left(\frac{N\alpha^p}{\beta^p} \right)^2 \right\}. \end{aligned} \tag{65}$$

These results prove that $\nabla^2\Psi_1$ is bounded

Let us now study the Hessian of Ψ_2 at $\mathbf{x} \in \mathbb{R}^N$. Let $\epsilon \in$

$]0,+\infty[$, let

$$\mathbf{\Lambda}_\epsilon = \frac{\nabla^2 \ell_{q,\eta}^q(\mathbf{x}) + q(q-1)\epsilon \boldsymbol{I}_N}{\ell_{q,\eta}^q(\mathbf{x})}, \tag{66}$$

and let

$$\nabla_\epsilon^2 \Psi_2(\mathbf{x}) = \frac{1}{q}\left(\mathbf{\Lambda}_\epsilon - \frac{\nabla \ell_{q,\eta}^q(\mathbf{x})\left(\nabla \ell_{q,\eta}^q(\mathbf{x})\right)^\top}{\ell_{q,\eta}^{2q}(\mathbf{x})}\right) \tag{67}$$

By continuity,

$$\lim_{\epsilon\to 0} |||\nabla_\epsilon^2 \Psi_2(\mathbf{x})||| = |||\nabla^2 \Psi_2(\mathbf{x})||| \tag{68}$$

On the other hand, since $\mathbf{\Lambda}_\epsilon$ is a positive definite matrix,

$$\nabla_\epsilon^2 \Psi_2(\mathbf{x}) = \frac{1}{q}\mathbf{\Lambda}_\epsilon^{1/2}(\boldsymbol{I}_N - \mathbf{v}_\epsilon \mathbf{v}_\epsilon^\top)\mathbf{\Lambda}_\epsilon^{1/2} \tag{69}$$

where, by using (10),

$$\begin{aligned}&\mathbf{v}_\epsilon\\ &= \mathbf{\Lambda}_\epsilon^{-1/2}\frac{\nabla \ell_{q,\eta}^q(\mathbf{x})}{\ell_{q,\eta}^q(\mathbf{x})}\\ &= \sqrt{\frac{q}{q-1}}\frac{1}{\ell_{q,\eta}^{q/2}(\mathbf{x})}\Big[\frac{\operatorname{sign}(x_1)|x_1|^{q-1}}{\sqrt{|x_1|^{q-2}+\epsilon}},\ldots,\frac{\operatorname{sign}(x_N)|x_N|^{q-1}}{\sqrt{|x_N|^{q-2}+\epsilon}}\Big]^\top.\end{aligned} \tag{70}$$

Therefore,

$$\begin{aligned}|||\nabla_\epsilon^2 \Psi_2(\mathbf{x})||| &= |||\frac{1}{q}\mathbf{\Lambda}_\epsilon^{1/2}(\boldsymbol{I}_N - \mathbf{v}_\epsilon \mathbf{v}_\epsilon^\top)\mathbf{\Lambda}_\epsilon^{1/2}|||\\ &\le \frac{1}{q}|||\mathbf{\Lambda}_\epsilon|||\;|||\boldsymbol{I}_N - \mathbf{v}_\epsilon \mathbf{v}_\epsilon^\top|||.\end{aligned} \tag{71}$$

According to (66),

$$\mathbf{\Lambda}_\epsilon = \frac{q(q-1)}{\ell_{q,\eta}^q(\mathbf{x})}\ \mathrm{Diag}\left((|x_n|^{q-2}+\epsilon)_{1\le n\le N}\right). \tag{72}$$

Consequently,

$$|||\mathbf{\Lambda}_\epsilon||| = \frac{q(q-1)}{\ell_{q,\eta}^q(\mathbf{x})}\sup_{1\le n\le N}(|x_n|^{q-2}+\epsilon). \tag{73}$$

We thus derive from (71) that

$$\begin{aligned}|||\nabla_\epsilon^2 \Psi_2(\mathbf{x})||| = &\frac{q-1}{\ell_{q,\eta}^q(\mathbf{x})}\sup_{1\le n\le N}(|x_n|^{q-2}+\epsilon)\\ &\times\max\{1, \|\mathbf{v}_\epsilon\|^2 - 1\}\end{aligned} \tag{74}$$

As $\epsilon \to 0$, (68) yields

$$|||\nabla^2 \Psi_2(\mathbf{x})||| \le \frac{q-1}{\ell_{q,\eta}^q(\mathbf{x})}\sup_{1\le n\le N}|x_n|^{q-2}\ \max\{1, \|\mathbf{v}\|^2 - 1\} \tag{75}$$

where, according to (70),

$$\|\mathbf{v}\|^2 = \lim_{\epsilon\to 0}\|\mathbf{v}_\epsilon\|^2 = \frac{q}{q-1}\frac{\sum_{n=1}^N |x_n|^q}{\ell_{q,\eta}^q(\mathbf{x})} \tag{76}$$

which is equivalent to

$$\|\mathbf{v}\|^2 - 1 = \frac{1}{q-1}\Big(1 - \frac{q\eta^q}{\ell_{q,\eta}^q(\mathbf{x})}\Big) \tag{77}$$

Since $\left(1 - \frac{q\eta^q}{\ell_{q,\eta}^q(\mathbf{x})}\right) < 1$ and $\frac{1}{q-1} < 1$ for all $q \in [2,+\infty[$, we deduce that $\|\mathbf{v}\|^2 - 1 < 1$. Resultingly,

$$\begin{aligned}|||\nabla^2 \Psi_2(\mathbf{x})||| &\le \frac{q-1}{\ell_{q,\eta}^q(\mathbf{x})}\sup_{1\le n\le N}|x_n|^{q-2}\\ &= \frac{q-1}{(\eta^q + \sum_{n=1}^N |x_n|^q)^{2/q}}\left(\frac{\sup_{1\le n\le N}|x_n|^q}{\eta^q + \sum_{n=1}^N |x_n|^q}\right)^{\frac{q-2}{q}}\\ &\le \frac{q-1}{(\eta^q + \sum_{n=1}^N |x_n|^q)^{2/q}} \qquad (78)\\ &\le \frac{q-1}{\eta^2}. \qquad (79)\end{aligned}$$

From the boundedness of $\nabla^2\Psi_1$ and $\nabla^2\Psi_2$, $\nabla^2\Psi = \nabla^2\Psi_1 - \nabla^2\Psi_2$ is bounded, hence Ψ is a Lipschitz-differentiable and

$$\begin{aligned}|||\nabla^2 \Psi(\mathbf{x})||| &= |||\nabla^2\Psi_1(\mathbf{x}) - \nabla^2\Psi_2(\mathbf{x})|||\\ &\le |||\nabla^2\Psi_1(\mathbf{x})||| + |||\nabla^2\Psi_2(\mathbf{x})|||.\end{aligned} \tag{80}$$

Using (64), (65), and (79), we conclude that

$$|||\nabla^2 \Psi(\mathbf{x})||| \le p\frac{\alpha^{p-2}}{\beta^p} + \frac{p}{2\alpha^2}\max\Big\{1, \Big(\frac{N\alpha^p}{\beta^p}\Big)^2\Big\} + \frac{q-1}{\eta^2},$$

hence Ψ is Lipschitz differentiable with constant L as in (21).

(ii) Let us now prove that Ψ satisfies the majorization inequality (24). By noticing that $\xi \mapsto (\xi+\alpha^2)^{p/2}$ is a concave function, it follows from standard majorization properties [85] that for every $(\mathbf{x}', ex) \in (\mathbb{R}^N)^2$, and $n \in \{1,\ldots,N\}$:

$$\begin{aligned}(x_n'^2 + \alpha^2)^{p/2} \le &(x_n^2+\alpha^2)^{p/2} + p x_n (x_n^2+\alpha^2)^{p/2-1}(x_n' - x_n)\\ &+ \frac{p}{2}(x_n^2+\alpha^2)^{p/2-1}(x_n' - x_n)^2.\end{aligned} \tag{81}$$

As a consequence,

$$\ell_{p,\alpha}^p(\mathbf{x}') \le \ell_{p,\alpha}^p(\mathbf{x}) + (\mathbf{x}'-\mathbf{x})^\top \nabla \ell_{p,\alpha}^p(\mathbf{x}) + \frac{p}{2}(\mathbf{x}'-\mathbf{x})^\top \boldsymbol{A}_1(\mathbf{x})(\mathbf{x}'-\mathbf{x})$$

where $\boldsymbol{A}_1(\mathbf{x}) = \mathrm{Diag}\left(\left(x_n^2+\alpha^2\right)^{p/2-1}\right)_{1\le n\le N}$. By using the Napier inequality expressed as

$$\big(\forall (u,v) \in]0,+\infty[^2\big) \qquad \log u \le \log v + \frac{u-v}{v}, \tag{82}$$

we get

$$\begin{aligned}\Psi_1(\mathbf{x}') \le &\Psi_1(\mathbf{x}) + (\mathbf{x}'-\mathbf{x})^\top \nabla\Psi_1(\mathbf{x})\\ &+ \frac{1}{2(\ell_{p,\alpha}^p(\mathbf{x}) + \beta^p)}(\mathbf{x}'-\mathbf{x})^\top \boldsymbol{A}_1(\mathbf{x})(\mathbf{x}'-\mathbf{x}).\end{aligned} \tag{83}$$

By applying the descent lemma to function $-\Psi_2$, and using (78) we obtain

$$\begin{aligned}\big(\forall(\mathbf{x},\mathbf{x}') \in \overline{\mathcal{B}}_{q,\rho}^2\big) \qquad -\Psi_2(\mathbf{x}') \le &-\Psi_2(\mathbf{x}) - (\mathbf{x}'-\mathbf{x})^\top\nabla\Psi_2(\mathbf{x})\\ &+ \frac{\chi_{q,\rho}}{2}\|\mathbf{x}'-\mathbf{x}\|^2.\end{aligned} \tag{84}$$

The majorization property is then derived from (83) and (84). The inequality (27) can be deduced in a straightforward manner, by noticing that both $\ell_{p,\alpha}^p(\mathbf{x})$ and $(x_n^2+\alpha^2)^{p/2-1}$ are minimal for $\mathbf{x} = \mathbf{0}_N$.

Acknowledgment

The authors thank Prof. Audrey Repetti (Heriot-Watt University, UK) for initial motivations and thoughtful discussions, and Prof. Marc-André Delsuc (University of Strasbourg, France), for helping with MS problem modeling.

References

[1] J. Laird, "The law of parsimony," *Monist*, vol. 29, no. 3, pp. 321–344, Jul. 1919.

[2] T. D. Bontly, "Modified Occam's razor: Parsimony, pragmatics, and the acquisition of word meaning," *Mind Lang.*, vol. 20, no. 3, pp. 288–312, Jun. 2005.

[3] I. N. Bronshtein, K. A. Semendyayev, G. Musiol, and H. Mühlig, *Handbook of Mathematics*, 5th ed. Springer, 2007.

[4] F. Albiac and N. J. Kalton, *Topics in Banach Space Theory*. Springer, 2006.

[5] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[6] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[7] L. Fortnow, "The status of the P versus NP problem," *Commun. ACM*, vol. 52, no. 9, p. 78, Sep. 2009.

[8] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[9] C. Ramirez, V. Kreinovich, and M. Argaez, "Why ℓ_1 is a good approximation to ℓ_0: A geometric explanation," *J. Uncertain Syst.*, vol. 7, no. 3, pp. 203–207, 2013.

[10] R. Ben Mhenni, S. Bourguignon, and J. Ninin, "Global optimization for sparse solution of least squares problems," *PREPRINT*, 2019.

[11] A. N. Tikhonov, "On the solution of ill-posed problems and the method of regularization," *Dokl. Akad. Nauk SSSR*, vol. 151, pp. 501–504, 1963.

[12] L. Jacques, L. Duval, C. Chaux, and G. Peyré, "A panorama on multiscale geometric representations, intertwining spatial, directional and frequency selectivity," *Signal Process.*, vol. 91, no. 12, pp. 2699–2730, Dec. 2011.

[13] A. E. Hoerl and R. W. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, Feb. 1970.

[14] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[15] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)*, vol. 98, 1998, pp. 82–90.

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

[17] H. Fu, M. K. Ng, M. Nikolova, and J. L. Barlow, "Efficient minimization methods of mixed ℓ_2-ℓ_1 and ℓ_1-ℓ_1 norms for image restoration," *SIAM J. Sci. Comput.*, vol. 27, no. 6, pp. 1881–1902, Jan. 2006.

[18] X. Ning, I. W. Selesnick, and L. Duval, "Chromatogram baseline estimation and denoising using sparsity (BEADS)," *Chemometr. Intell. Lab. Syst.*, vol. 139, pp. 156–167, Dec. 2014.

[19] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.

[20] C. De Mol, E. De Vito, and L. Rosasco, "Elastic-net regularization in learning theory," *Complexity*, vol. 25, no. 2, pp. 201–230, Apr. 2009.

[21] W. J. Fu, "Penalized regressions: the bridge versus the lasso," *J. Comput. Graph. Stat.*, vol. 7, no. 3, pp. 397–416, Sep. 1998.

[22] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$," *Appl. Comput. Harmon. Analysis*, vol. 26, no. 3, pp. 395–407, May 2009.

[23] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization," *PREPRINT*, 2016.

[24] F. Wen, L. Chu, P. Liu, and R. C. Qiu, "A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning," *IEEE Access*, vol. 6, pp. 69 883–69 906, 2018.

[25] X. Chen, D. Ge, Z. Wang, and Y. Ye, "Complexity of unconstrained L_2-L_p minimization," *Math. Programm.*, vol. 143, no. 1-2, pp. 371–383, Nov. 2012.

[26] E. F. Beckenbach, "An inequality of Jensen," *Amer. Math. Monthly*, vol. 53, no. 9, pp. 501–505, Nov. 1946.

[27] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Trans. Inform. Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.

[28] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best-basis selection," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 713–718, Mar. 1992.

[29] H. Šikić and M. V. Wickerhauser, "Information cost functions," *Appl. Comput. Harmon. Analysis*, vol. 11, no. 2, pp. 147–166, Sep. 2001.

[30] J. Aczél and Z. Daróczy, *On Measures of Information and their Characterizations*. Academic Press, 1975.

[31] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*. Academic Press, 1979, vol. 143.

[32] D. Leporini, J.-C. Pesquet, and H. Krim, "Best basis representations based on prior statistical models," in *Bayesian Inference in wavelet based models*, ser. Lect. Notes Comput. Sci., P. Müller and B. Vidakovic, Eds. Springer, 1999, vol. 141, pp. 155–172.

[33] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[34] B. Ricaud and B. Torrésani, "A survey of uncertainty principles and some signal processing applications," *Adv. Comput. Math.*, vol. 40, no. 3, pp. 629–650, Oct. 2014.

[35] A. M. Bronstein, M. M. Bronstein, M. Zibulevsky, and Y. Y. Zeevi, "Sparse ICA for blind separation of transmitted and reflected images," *Int. J. Imag. Syst. Tech.*, vol. 15, no. 1, pp. 84–91, 2005.

[36] R. A. Fisher, "Moments and product moments of sampling distributions," *Proc. Lond. Math. Soc.*, vol. s2-30, no. 1, pp. 199–238, Jan. 1930.

[37] D. J. Field, "What is the goal of sensory coding?" *Neural Comput.*, vol. 6, no. 4, pp. 559–601, Jul. 1994.

[38] O. Tanrikulu and A. G. Constantinides, "Least-mean kurtosis: a novel higher-order statistics based adaptive filtering algorithm," *Elec. Letters*, vol. 30, no. 3, pp. 189–190, Feb. 1994.

[39] D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, Jun. 21-25, 2011, pp. 233–240.

[40] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, and J.-C. Pesquet, "Euclid in a taxicab: Sparse blind deconvolution with smoothed ℓ_1/ℓ_2 regularization," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 539–543, May 2015.

[41] K. Nose-Filho, C. Jutten, and J. M. T. Romano, "Sparse blind deconvolution based on scale invariant smoothed ℓ_0-norm," in *Proc. Eur. Sig. Image Proc. Conf.*, Lisbon, Portugal, Sep. 1-5, 2014, pp. 461–465.

[42] Y. Li, K. Lee, and Y. Bresler, "Identifiability in blind deconvolution with subspace or sparsity constraints," *IEEE Trans. Inform. Theory*, vol. 62, no. 7, pp. 4266–4275, Jul. 2016.

[43] X. Bresson, T. Laurent, and J. von Brecht, "Enhanced Lasso recovery on graph," *Proc. Eur. Sig. Image Proc. Conf.*, Jun. 2015.

[44] R. A. Wiggins, "Minimum entropy deconvolution," *Geoexploration*, vol. 16, pp. 21–35, 1978.

[45] G. A. Ferguson, "The concept of parsimony in factor analysis," *Psychometrika*, vol. 19, no. 4, pp. 281–290, Dec. 1954.

[46] J. F. Claerbout and F. Muir, "Robust modeling with erratic data," *Geophysics*, vol. 38, no. 5, pp. 826–844, Oct. 1973.

[47] D. Donoho, "On minimum entropy deconvolution," in *Applied Time Series Analysis II*, 1981, pp. 565–608.

[48] W. C. Gray, "Variable norm deconvolution," Stanford Exploration Project, Tech. Rep. SEP-14, Apr. 1978.

[49] M. Castella, A. Chevreuil, and J.-C. Pesquet, "Convolutive mixtures," in *Handbook of Blind Source Separation. Independent Component Analysis and Applications*, P. Comon and C. Jutten, Eds. Oxford UK, Burlington USA: Academic Press, 2010.

[50] L. Demanet and P. Hand, "Scaling law for recovering the sparsest element in a subspace," *Inf. Inference*, vol. 3, no. 4, pp. 295–309, 2014.

[51] X. Chang, Y. Wang, R. Li, and Z. Xu, "Sparse k-means with ℓ_∞/ℓ_0 penalty for high-dimensional data clustering," *Stat. Sin.*, vol. 28, pp. 1265–1284, 2018.

[52] M. E. Lopes, "Unknown sparsity in compressed sensing: Denoising and inference," *IEEE Trans. Inform. Theory*, vol. 62, no. 9, pp. 5145–5166, Sep. 2016.

[53] Z. Zhou and J. Yu, "Sparse recovery based on q-ratio constrained minimal singular values," *Signal Process.*, vol. 155, pp. 247–258, Feb. 2019.

[54] M. E. Lopes, "Estimating unknown sparsity in compressed sensing," in *Proc. Int. Conf. Mach. Learn.*, Jun. 17-19, 2013.

[55] Y. Yu, N. Peng, and J. Gan, "Concave-convex norm ratio prior based double model and fast algorithm for blind deconvolution," *Neurocomputing*, vol. 171, no. 1, pp. 781–787, Jan. 2016.

[56] X. Jia, M. Zhao, Y. Di, P. Li, and J. Lee, "Sparse filtering with the generalized l_p/l_q norm and its applications to the condition monitoring of rotating machinery," *Mech. Syst. Signal Process.*, vol. 102, pp. 198–213, Mar. 2018.

[57] X. Jia, M. Zhao, M. Buzza, Y. Di, and J. Lee, "A geometrical investigation on the generalized l_p/l_q norm for blind deconvolution," *Signal Process.*, May 2017.

[58] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function," *J. Optim. Theory Appl.*, vol. 162, no. 1, pp. 107–132, Jul. 2014.

[59] R. Mazumder, J. H. Friedman, and T. Hastie, "SparseNet: Coordinate descent with nonconvex penalties," *J. Am. Stat. Assoc.*, vol. 106, no. 495, pp. 1125–1138, Sep. 2011.

[60] I. W. Selesnick and İ. Bayram, "Sparse signal estimation by maximally sparse convex optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1078–1092, Mar. 2014.

[61] S. Becker and M. J. Fadili, "A quasi-Newton proximal splitting method," in *Proc. Ann. Conf. Neur. Inform. Proc. Syst.*, vol. 2, Dec. 3-6, 2012, pp. 2618–2626.

[62] P. L. Combettes and B. C. Vũ, "Variable metric forward-backward splitting with applications to monotone inclusions in duality," *Optimization*, vol. 63, no. 9, pp. 1289–1318, dec 2014.

[63] A. Repetti and Y. Wiaux, "Variable metric forward-backward algorithm for composite minimization problems," *PREPRINT*, 2019.

[64] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss–Seidel methods," *Math. Programm.*, no. 137, pp. 91–129, 2013.

[65] J.-C. Pesquet and N. Pustelnik, "A parallel inertial proximal optimization method," *Pac. J. Optim.*, vol. 8, no. 2, pp. 273–305, Apr. 2012.

[66] F. Abboud, E. Chouzenoux, J.-C. Pesquet, J.-H. Chenot, and L. Laborelli, "Dual block-coordinate forward-backward algorithm with application to deconvolution and deinterlacing of video sequences," *J. Math. Imaging Vision*, vol. 59, no. 3, pp. 415–431, Nov. 2017.

[67] L. van den Dries, *Tame topology and o-minimal structures*. Cambridge University Press, 1998.

[68] K. E. Tan, B. C. Ellis, R. Lee, P. D. Stamper, S. X. Zhang, and K. C. Carroll, "Prospective evaluation of a matrix-assisted laser desorption ionization-time of flight mass spectrometry system in a hospital clinical microbiology laboratory for identification of bacteria and yeasts: a bench-by-bench study for assessing the impact on time to identification and cost-effectiveness," *J. Clin. Microbiol.*, vol. 50, no. 10, pp. 3301–3308, Oct. 2012.

[69] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[70] A. Scalbert, L. Brennan, O. Fiehn, T. Hankemeier, B. S. Kristal, B. van Ommen, E. Pujos-Guillot, E. Verheij, D. Wishart, and S. Wopereis, "Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research," *Metabolomics*, vol. 5, no. 4, pp. 435–458, Dec. 2009.

[71] N. Mano and J. Goto, "Biomedical and biological mass spectrometry," *Anal. Sci.*, vol. 19, no. 1, pp. 3–14, Jan. 2003.

[72] P. Schmitt-Kopplin and M. Frommberger, "Capillary electrophoresis-mass spectrometry: 15 years of developments and applications," *Electrophoresis*, vol. 24, no. 2223, pp. 3837–3867, 2003.

[73] S. A. Schwartz, R. J. Weil, M. D. Johnson, S. A. Toms, and R. M. Caprioli, "Protein profiling in brain tumors using mass spectrometry," *Clin. Cancer Res.*, vol. 10, no. 3, pp. 981–987, 2004.

[74] A. Panchaud, M. Affolter, and M. Kussmann, "Mass spectrometry for nutritional peptidomics: How to analyze food bioactives and their health effects," *Proteomics*, vol. 75, no. 12, pp. 3546–3559, 2012.

[75] M. W. Senko, J. P. Speir, and F. W. McLafferty, "Collisional activation of large multiply charged ions using Fourier transform mass spectrometry," *Anal. Chem.*, vol. 66, no. 18, pp. 2801–2808, 1994.

[76] M. W. Senko, S. C. Beu, and F. W. McLafferty, "Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions," *J. Am. Soc. Mass Spectrom.*, vol. 6, no. 4, pp. 229–233, Apr. 1995.

[77] A. Cherni, E. Chouzenoux, and M.-A. Delsuc, "Fast dictionary-based approach for mass spectrometry data analysis," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Apr. 15-20, 2018, pp. 816–820.

[78] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Am. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[79] W. J. J. Rey, *Introduction to Robust and Quasi-Robust Statistical Methods*. Springer, 1983.

[80] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 563–591, Mar. 2013.

[81] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem," *SIAM J. Imaging Sci.*, vol. 8, no. 3, pp. 1607–1639, 2015.

[82] E. Soubies, L. Blanc-Féraud, and G. Aubert, "New insights on the optimality conditions of the ℓ_2-ℓ_0 minimization problem," *J. Math. Imaging Vision*, 2019.

[83] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vision*, vol. 40, no. 1, pp. 120–145, 2011.

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

[85] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, Feb. 2004.