SPOQ lp-Over-lq Regularization for Sparse Signal Recovery applied to Mass Spectrometry
Alef Cherni, Emilie Chouzenoux, Laurent Duval, Jean-Christophe Pesquet

To cite this version:
Alef Cherni, Emilie Chouzenoux, Laurent Duval, Jean-Christophe Pesquet. SPOQ lp-Over-lq Regularization for Sparse Signal Recovery applied to Mass Spectrometry. 2020. hal-02454518

HAL Id: hal-02454518
https://hal.archives-ouvertes.fr/hal-02454518
Submitted on 24 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
SPOQ $\ell_p$-Over-$\ell_q$ Regularization for Sparse Signal Recovery applied to Mass Spectrometry

Alef Cherni, Student member, IEEE, Emilie Chouzenoux, 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 require additional information for sound solutions with tractable optimization algorithms. Sparsity yields consequent heuristics to inquire additional information for sound solutions with tractable invest in computable proxies, such as the $\ell_1$ norm. However, the latter does not exhibit the desirable property of scale invariance for sparse data. Generalizing the SOOT Euclidean/Taxicab $\ell_1/\ell_2$ norm-ratio initially introduced for blind deconvolution, we propose SPOQ, a family of smoothed scale-invariant penalty functions. It consists of a Lipschitz-differentiable surrogate for $\ell_p$-over-$\ell_q$ 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 $\ell_0$, $\ell_1$, Cauchy, Welsch, and 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. On the role of sparsity measures in data science

The law of parsimony (or Occam’s razor) is an important heuristic principle and a guideline in history, social and empirical sciences. In modern terms, a preference to simpler models, when they possess — on observed phenomena — a power of explanation comparable to more complex ones. In statistical data processing, it can limit the degrees of freedom for parametric models, reduce a search space, define stopping criteria, bound filter support, simplify signals or images with meaningful structures. For processes that inherently generate sparse information (spiking neurons, chemical sensing), degraded by smoothing kernels and noise, sparsity may provide a quantitative target on restored data. On partial observations, it becomes a means to selecting one solution, among all potential solutions that are consistent with observations.

A. Cherni is with the Department of Aix-Marseille Univ, CNRS, Centrale Marseille, 12M, Marseille, France. E. Chouzenoux and J.-C. Pesquet are with Université Paris-Saclay, Centrale-Supélec, Inria, Centre de Vision Numérique, 91190 Gif-sur-Yvette, France. L. Duval is with ESIEE Paris, University Paris-Est, 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.

1Named after William of Ockham, stated as “Entities should not be multiplied without necessity” (”Non sunt multiplicanda entia sine necessitate”).

A natural playground for sparsity, in discrete time series analysis, is $c_{00}(\mathbb{R})$, the space of almost-zero real sequences, which is closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution, closed under finite addition and convolution. It is neither a norm nor a quasinorm. Its pseudonorm moniker depends on the definition of the subhomogeneity axiom.

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

2It is neither a norm nor a quasinorm. Its pseudonorm moniker depends on the definition of the subhomogeneity axiom.

3The lowest $K$, modulus of concavity of the quasinorm, saturates to 1 for norms. For $0 < p \leq 1$, $\ell_p(x + y) \leq 2^{1/p} (\ell_p(x) + \ell_p(y)).$

4Other denominations are subset selection, minimum weight solution, sparse null-space, or minimum set cover.
filtering and source separation in analytical chemistry [17]. A convex combination of ridge and lasso regularizations yields the elastic net regularization [18], [19]. Other convex p-norms (p ≥ 1) regularizations have been addressed, as in bridge regression which interpolates between lasso and ridge [20]. Expecting sparser solutions in practice, non-convex least-squares plus ℓ_p (p < 1) problems have been addressed [21]. Although a priori appealing for sparse data restoration, such problems retain NP-hard complexities [22]. Another caveat to using 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 (ℓ_p(λx) = |λ|ℓ_p(x), for λ ∈ ℝ). Either to copycat the 0-degree homogeneity of ℓ_0, or to cope with scaling ambiguity, scale-invariant contrast functions were suggested [23]. The work [24] (SOOT: Smoothed One-Over-Two norm ratio) also proposed an efficient optimization algorithm with theoretically guaranteed convergence. We now investigate a broader family of (quasi-)norm ratios ℓ_p(x)/ℓ_q(x) with couples (p, q) ∈ [0, 2] × [2, ∞[, based on both their counting properties and probabilistic interpretation.

First, we have equivalence relations in finite dimension:

\[
ℓ_q(x) \leq ℓ_p(x) \leq ℓ_0(x) \overset{p \leq q}{\leq} \frac{1}{p} ℓ_q(x) \leq N^{1 - \frac{1}{q}} ℓ_q(x)
\]

(1)

with p ≤ q, from the standard power-mean inequality [25] implying classical ℓ_p-space embeddings and generalized Rogers-Hölder’s inequalities. The LHS in (1) is attained when x realizes an instance of the most prototypical sparse signals of c_{00}(ℝ), with only one non-zero component. The RHS is reached by a maximally non-sparse x, where all the samples are set to a non-zero constant. Thus, ℓ_p/ℓ_q quasinorm-ratios provide interesting proxies for a sparseness measure of x, to quantify how much the “action” or “energy” of a discrete signal is concentrated into only a few of its components. 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.

B. Penalties with quasinorm and norm ratios

For every p ∈ [0, 2] and q ∈ [2, +∞[, we thus define:

\[
(ℓ_p/ℓ_q(x))^p = \sum_{n=1}^{N} \left( \frac{|x_n|^q}{\sum_{n'=1}^{N} |x_{n'}|^q} \right)^{p/q}.
\]

(2)

Expounding the term, peered in Jensen’s inequalities [25],

\[
p_n = \frac{|x_n|^q}{\sum_{n'=1}^{N} |x_{n'}|^q}
\]

(3)
as a discrete probability distribution, then ℓ_p/ℓ_q rewrites as an increasing function (u → u^{1/p}) of a sum of concave functions (u → u^{1/q} when p ≤ q) of probabilities. The minimization of such an additive information cost function [26], [27], a special case of Schur-concave functionals [28], [29], is used for instance in best basis selection [30]. Thus, special cases of ℓ_p/ℓ_q quasinorm ratios have served as sparsity-inducing penalties in the long history of blind signal deconvolution or image deblurring, as as stopping criteria (for instance in NMF, non-negative matrix factorization [31], measures of sparsity satisfying a number of sound parsimony-prone axioms [32], estimates for time-frequency spread or concentration [33]. The ℓ_p quasinorm weighted by ℓ_2 is considered as a “possible sparseness criterion” [34] when scaled with the normalizing factor N^{-\frac{1}{2}}. It bears close connection with the kurtosis [35] for centered distributions and central moments, widely used in sparse sensory coding [36] or adaptive filtering [37].

The most frequent one with (p, q) = (1, 2) is used [38], [24] as a surrogate to ℓ_0 [39], [40]. This ratio was used to enhance lasso recovery on graphs [41]. Its early history includes the “minimum entropy deconvolution” proposed in [42], where the “varimax norm”, akin to kurtosis (ℓ_1/ℓ_2)^p, is maximized to yield visually simpler (spikier) signals. It was inspired by simplicity measures proposed in factor analysis [43], and meant to improve one of the earliest mentioned ℓ_0 regularization [44] in seismic. The relationship with the concept of entropy was explained later [45]. It was generalized to the so-called “variable norm deconvolution” by maximizing (ℓ_q/ℓ_1)^{1/q} [46]. Note that techniques in [42], [46] 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 [47] for an overview. Recently, [48] uses ℓ_1/ℓ_0 for sparse recovery, and [49] ℓ_∞/ℓ_0 for cardinality-penalized clustering. The family of entropy-based sparsity measures (ℓ_q/ℓ_1)^{1/q} [50] (termed q-ratio sparsity level in [51]), extends a previous work on squared ℓ_1/ℓ_2 ratios [52] for compressed sensing. Finally, [53] proposes an extension of [50] to an ℓ_q/ℓ_2 ratio to discriminate between sharp and blurry images, and [54], [55] 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 set of smooth-enough surrogates to ℓ_0 with sufficient Lipschitz regularity. The resulting penalties, called SPOQ (Smoothed p-Over-q), extend the ℓ_1/ℓ_2 SOOT [24]. A novel trust-region algorithm generalizes and improves the variable metric forward-backward algorithm from [56]. Section II recalls the parsimony role and introduces sparsity measures. Section III describes the observation model and the proposed SPOQ quasinorm-norm ratio regularization. We derive our trust-region minimization algorithm and analyze its convergence in Section IV. Section V illustrates the good performance of SPOQ regularization in recovering “naturally sparse” mass spectrometry signals, over a range of existing sparsity penalties.

II. PROPOSED FORMULATION

A. Sparse signal reconstruction

Let us consider the observation model

\[
y = Dx + b
\]

(4)

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

$$\min_{x \in \mathcal{C}} \Phi(x) + \Theta(x),$$

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

$$\Theta(x) = \iota_{B}(Dx).$$

Hereabove, $\xi > 0$ is a parameter depending on noise characteristics, $D^T$ is the Euclidean ball centered at $y$ with radius $\xi$, and $\iota_B$ denotes the indicator function of a set $S$, equal to zero for $x \in S$, and $+\infty$ otherwise. Furthermore, function $\Psi : \mathbb{R}^N \rightarrow ]-\infty, +\infty]$ is a regularization function used to enforce desirable properties on the solution. The choice of $\Psi$ is essential for reaching satisfying results by promoting desirable properties in the sought signal. When sparsity is expected, the $\ell_1$ norm is probably the most used regularization function. It is a convex envelope proxy to $\ell_0$, a key feature for deriving efficient minimizations to solve [5].

However, because it is not scale invariant, the $\ell_1$ 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 $\ell_p/\ell_q$ 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 $\ell_p$ and $\ell_q$ parametrized by constants $(\alpha, \eta) \in [0, +\infty]^2$: for every $x = (x_n)_{1 \leq n \leq N} \in \mathbb{R}^N$,

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

and

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

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

$$\Psi(x) = \log \left( \frac{\ell_{p, \alpha}(x) + \beta \eta}{\ell_{q, \eta}(x)} \right).$$

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 [57], [58] and of the $\ell_p/\ell_q$ ratio. It generalizes the Euclidean/Taxicab Smoothed One-Over-Two-norm (SOOT) penalty [24], 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 $\ell_0$ and $\ell_1$. It is worth noticing that, on the second row, the logarithm sharpens the $\ell_1/\ell_2$ behavior toward $\ell_0$. 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 $\ell_0$ 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 $\Psi$ at $x \in \mathbb{R}^N$:

$$\nabla \ell_{p, \alpha}(x) = \eta \left( \sum_{n=1}^{N} (x_n(x_n^2 + \alpha^2)^{p/2} - \alpha^p) \right)^{1/p} \mathbb{1}_{\{x_n \neq 0\}},$$

and

$$\nabla \ell_{q, \eta}(x) = \eta \left( \sum_{n=1}^{N} |x_n|^q \right)^{1/q} \mathbb{1}_{\{|x_n| > 0\}}.$$

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

$$\Psi(x) = \log \left( \frac{\ell_{p, \alpha}(x) + \beta \eta}{\ell_{q, \eta}(x)} \right).$$

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

![Figure 1. Sparsity-promoting penalties, scaled to [0, 1]. From top to bottom and from left to right: $\ell_0$, $\ell_1$, smoothed $\ell_1/\ell_2$, SOOT, smoothed $\ell_p/\ell_q$ and SPOQ with $(\alpha, \beta, \eta) = (7 \times 10^{-7}, 3 \times 10^{-3}, 1 \times 10^{-1})$.](image-url)
Hence, \( \nabla \Psi = \nabla \Psi_1 - \nabla \Psi_2 \), and \( \nabla^2 \Psi = \nabla^2 \Psi_1 - \nabla^2 \Psi_2 \), with
\[
\nabla \Psi_1(x) = 1 (p_{\alpha}(x) + \beta p, \tag{15}
\]
\[
\nabla \Psi_2(x) = 1 (p_{\alpha}(x) + \beta p, \tag{16}
\]
\[
p \nabla^2 \Psi_1(x) = \frac{\nabla^2 p_{\alpha}(x)}{q} - \frac{\nabla^2 p_{\alpha}(x)}{q} (\nabla^2 p_{\alpha}(x))^T, \tag{17}
\]
\[
q \nabla^2 \Psi_2(x) = \frac{\nabla^2 q_{\alpha}(x)}{q} - \frac{\nabla^2 q_{\alpha}(x)}{q} (\nabla^2 q_{\alpha}(x))^T. \tag{18}
\]

From the above, we derive the following proposition, stating that for suitable parameter choices, function \( \Psi \) has \( 0_N \), 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, \( \nabla^2 \Psi(0_N) \) is a positive definite matrix and \( 0_N \) is a local minimizer of \( \Psi \). In addition, if
\[
\eta^2 \geq \beta^2 \max \left\{ \frac{2q-1}{4(q^2-1)}, \frac{2q-1}{8q^2} \right\}, \tag{19}
\]
then \( 0_N \) is a global minimizer of \( \Psi \).

**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 \( \Psi \).

**Proposition 2.** Let \( \Psi \) be defined by (7).

(i) \( \Psi \) is a \( p \)-Lipschitz differentiable function on \( \mathbb{R}^N \), i.e., for every \( (x, x') \in \mathbb{R}^N \),
\[
\|\nabla \Psi(x) - \nabla \Psi(x')\| \leq L \|x - x'\|, \tag{20}
\]
where
\[
L = p \frac{\alpha p-2}{\beta p} + \frac{p}{2 \alpha} \max \left\{ \frac{1}{2 \alpha}, \left( \frac{N p \alpha p-2}{\beta p} \right)^2 \right\}, \tag{21}
\]
In particular,
\[
\Psi(x') \leq \Psi(x) + (x' - x)^T \nabla \Psi(x) + \frac{L}{2} \|x' - x\|^2. \tag{22}
\]

(ii) For every \( \rho \in [0, +\infty[ \), define the \( \ell_{\rho} \)-ball complement:
\[
\mathcal{B}_{q, \rho} = \{ x = (x_n)_{1 \leq n \leq N} \in \mathbb{R}^N \mid \sum_{n=1}^{N} |x_n|^q \geq \rho^q \}. \tag{23}
\]
\( \Psi \) admits a quadratic tangent majorant at every \( x \in \mathcal{B}_{q, \rho} \), i.e.
\[
(\forall x' \in \mathcal{B}_{q, \rho}) \quad \Psi(x') \leq \Psi(x) + (x' - x)^T \nabla \Psi(x) + \frac{1}{2} (x' - x)^T A_{q, \rho}(x)(x' - x), \tag{24}
\]
where
\[
A_{q, \rho}(x) = \chi_{q, \rho} I_N + \frac{1}{2} p_{\alpha}(x) + \beta p, \tag{25}
\]
with
\[
\chi_{q, \rho} = \frac{q - 1}{q^2 + \rho^2}. \tag{26}
\]

Moreover, for every \( x \in \mathbb{R}^N \),
\[
\chi_{q, \rho} I_N \leq A_{q, \rho}(x) \leq (\chi_{q, \rho} + \beta^{-p} \alpha p^{-2}) I_N. \tag{27}
\]

**Proof:** See Appendix B

**Proposition 2** leads to a rather simple majorizing function for \( \Psi \), valid on the whole Euclidean space \( \mathbb{R}^N \). This extends our previous result established in [24] 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 \( \mathcal{B}_{q, \rho} \). By limiting the size of the region where majorization is imposed, one may expect more accurate approximations for \( \Psi \). This observation motivates the trust-region minimization algorithm we will 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:
\[
\min_{x \in \mathbb{R}^N} \Omega(x) \tag{28}
\]
where \( \Omega = \Psi + \Phi \) with \( \Psi \) defined in (9) and \( \Phi = \Theta + \iota C \). We will assume that function \( \Theta \) 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 \( \Theta \). Note that the assumptions made on set \( C \) implies that \( \Phi \) is coercive and belongs to \( \Gamma_0(\mathbb{R}^N) \). The particular structure of \( \Omega \), summing a Lipschitz differentiable function \( \Psi \) and the non-necessarily smooth convex term \( \Phi \) suits it well to the class of variable metric forward-backward (VMFB) optimization methods [59, 60, 56, 61]. In such methods, one alternates gradient steps on \( \Psi \) and proximity steps on \( \Phi \), 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) matrix \( A \in \mathbb{R}^{N \times N} \), the proximity operator of \( \Phi \) at \( x \in \mathbb{R}^N \) relative to the metric \( A \) is defined as
\[
\text{prox}_{A, \Phi}(x) = \text{argmin}_{z \in \mathbb{R}^N} \left( \frac{1}{2} \|z - x\|^2 + \Phi(z) \right) \tag{29}
\]
with notation \( \|u\|_A = \sqrt{u^T A u} \) for \( u \in \mathbb{R}^N \). Then, the VMFB method for solving Problem (28) reads, for every \( k \in \mathbb{N} \),
\[
x_{k+1} = \text{prox}_{\gamma_k^{-1} A_k, \Phi}(x_k - \gamma_k (A_k^{-1} \nabla \Psi(x_k))), \tag{30}
\]
where \( x_0 \in \mathbb{R}^N \), and \( (\gamma_k)_{k \in \mathbb{N}} \) and \( (A_k)_{k \in \mathbb{N}} \) are sequences of positive stepsizes and SPD metrics, respectively, chosen in such a way to guarantee the convergence of VMFB iterates to a solution to Problem (5) [60]. Two main challenges arise, when implementing the VMFB algorithm, namely (i) the choice for the preconditioning matrices \( (A_k)_{k \in \mathbb{N}} \), and (ii) the evaluation of the proximity operator involved in the update (30). In [60], a novel methodology was proposed based on the choice of pre-conditioning matrices satisfying a majorization condition for \( \Psi \). This methodology provides a practically efficient algorithm. Furthermore, it allows to establish convergence in the case of a non necessarily convex function \( \Psi \), as soon as it satisfies the so-called Kurdyka-Łojasiewicz inequality [62]. 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 the only act as smoothing constants for ensuring differentiability. As shown in Proposition 2(iii), 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 R^N. We extend [50] 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

arg1, for solving Problem (28). At each iteration k ∈ N, we will make B ≥ 1 trials of values (ρk,i)k∈N,1≤i≤B for the trust-region radius. For each tested radius ρk,i ≥ 0, a VMFB update zk,i is computed within the majorizing metric Aq,ρk,i at xk, defined in Proposition 2(ii). Then, a test is performed for checking whether the update does belong to the region Ωq,ρk,i. If not, the region size is reduced with a factor θ ∈ [0, 1[, and a new VMFB step is performed. The trust-region loop stops as soon as zk,i ∈ Ωq,ρk,i. Note that, for the last trial, i.e. i = B, a radius equal to 0 is tested, which allows us to guarantee the well-definedness of our method. More precisely, this leads to the following sequence, for the radius values:

ρk,i = \begin{cases} \sum_{n=1}^{N} |x_{n,k}|^q & \text{if } i = 1 \\ θρk,i-1 & \text{if } 2 ≤ i ≤ B - 1 \\ 0 & \text{if } i = B. \end{cases} \label{eq:31}

Let us remark that xk ∈ Ωq,ρk,i and the following inclusion holds by construction:

Ωq,ρk,1 ⊂ Ωq,ρk,2 ⊂ ... ⊂ Ωq,ρk,i ⊂ R^N. \label{eq:32}

Algorithm 1 TR-VMFB algorithm

Initialize: x_0 ∈ domΦ, B ∈ N^*, θ ∈ [0, 1[, (γk)k∈N ∈ [0, +∞[ For k = 0, 1, ..., B :

For i = 1, ..., B :

Set trust-region radius ρk,i using (31)

Construct A_{k,i} = A_{q,ρk,i} (x_k) using (25)

z_{k,i} = prox_{γ_k^{-1}A_{k,i}Φ} (x_k − γ_k(A_{k,i})^{-1}∇Ψ(x_k))

If z_{k,i} ∈ Ωq,ρk,i : Stop loop

x_{k+1} = z_{k,i}

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 to a solution to Problem (28). We start with the following assumptions on the sequences (γk)k∈N and (A_{k,i})k∈N,1≤i≤B, that are necessary for our convergence analysis:

Assumption 1.

(i) There exists (γ, ν) ∈ [0, +∞[^2 such that for every k ∈ N, γ ≤ γk ≤ 2 − ν.

(ii) There exists (υ, ν) ∈ [0, +∞[^2, such that, for every k ∈ N and for every i ∈ {1, ..., B}, υ I_N ≤ A_{k,i} ≤ υ I_N.

Remark 1. By construction, iterates (x_k)k∈N produced by Algorithms 1 and 2 belong to the domain of Φ and therefore to the set C. This implies that sequence (x_k)k∈N is bounded, so that there exists p_{max} ≥ 0 such that, for every k ∈ N and i ∈ {1, ..., B}, we have ρk,i ≤ p_{max}. Assumption 1(ii) thus holds as a consequence of (27), by setting υ = χ_{q,ρ_{max}} and ν = χ_{q,0} + β^{-p_{max}}.

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

Lemma 1.

Under Assumption 1 for every i ∈ {1, ..., B}, there exist r_{k,i} ∈ ∂Ψ(z_{k,i}) such that conditions Alg. 2(a) and Alg. 2(b) are fulfilled, with

z_{k,i} = prox_{γ_k^{-1}A_{k,i}Φ} (x_k − γ_k(A_{k,i})^{-1}∇Ψ(x_k)) \label{eq:33}

and κ ≥ γ^{-1}√ν.

Algorithm 2 TR-VMFB algorithm — Inexact form

Initialize: x_0 ∈ domΦ, B ∈ N^*, θ ∈ [0, 1[, (γk)k∈N ∈ [0, +∞[ For k = 0, 1, ..., B :

For i = 1, ..., B :

Set trust-region radius ρk,i using (31)

Construct A_{k,i} = A_{q,ρk,i} (x_k) using (25)

Find z_{k,i} ∈ R^N such that:

(a) Φ(z_{k,i}) + (z_{k,i} - x_k) ^ τ Ψ(x_k) + γ_k^{-1}∥z_{k,i} - x_k∥^2_{A_{k,i}} ≤ Φ(x_k)

(b) ∥∇Ψ(x_k) + r_{k,i}∥ ≤ θ∥z_{k,i} - x_k∥_{A_{k,i}},

with r_{k,i} ∈ ∂Φ(z_{k,i}) and κ > 0

If z_{k,i} ∈ Ωq,ρk,i : Stop loop

x_{k+1} = z_{k,i}
Lemma 2, and Theorem 4.1, we deduce that the Kurdyka-Łojasiewicz inequality [62], [63]. Therefore, by using algebraic functions and logarithmic function, so that it satisfies the convexity of \( K \) such that

\[
\begin{aligned}
\langle r_{k,i}, \nabla \Phi(x_k) \rangle + \gamma_k^{-1} A_{k,i}(x_k - z_{k,i}) + \frac{\gamma_k^{-1}}{2} \| z_{k,i} - x_k \|_A^2 \leq \Phi(z_{k,i}) - \Phi(x_k).
\end{aligned}
\]

Thus, \( z_{k,i} \) satisfies:

\[
\Phi(z_{k,i}) + \langle z_{k,i} - x_k, \nabla \Phi(x_k) \rangle + \gamma_k^{-1} \| z_{k,i} - x_k \|_A^2 \leq \Phi(x_k).
\]

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

\[
\| r_{k,i} + \nabla \Phi(x_k) \| = \gamma_k^{-1} \| A_{k,i}(x_k - z_{k,i}) \|
\]

Hence the condition Alg. 2(b) holds for \( \kappa \geq \gamma^{-1}_{\sqrt{F}} \).

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 \( k \in \mathbb{N} \),

\[
\Omega(x_{k+1}) \leq \Omega(x_k) - \frac{\mu}{2} \| x_{k+1} - x_k \|^2
\]

with \( (x_k)_{k \in \mathbb{N}} \) defined in Algorithm 2.

Proof: We have

\[
(\forall k \in \mathbb{N}) \quad \Omega(x_{k+1}) = \Psi(x_{k+1}) + \Phi(x_{k+1})
\]

Under condition Alg. 2(a),

\[
\Psi(x_{k+1}) + \langle x_{k+1} - x_k, \nabla \Psi(x_k) \rangle + \gamma_k^{-1} \| x_{k+1} - x_k \|_A^2 \leq \Phi(x_k).
\]

By construction, \( x_{k+1} \in \mathcal{B}_{q,p,e_1} \) for some \( i \in \{1, \ldots, B\} \). Moreover, \( x_k \in \mathcal{B}_{q,p,e_1} \subset \mathcal{B}_{q,p,e_1} \). Therefore, by Proposition 2

\[
\Psi(x_{k+1}) \leq \Psi(x_k) + \langle x_{k+1} - x_k, \nabla \Psi(x_k) \rangle + \frac{1}{2} \| x_{k+1} - x_k \|_A^2.
\]

Thus,

\[
\Omega(x_{k+1}) \leq \Omega(x_k) + \frac{1}{2} \| x_{k+1} - x_k \|_A^2 - \gamma_k^{-1} \| x_{k+1} - x_k \|_A^2.
\]

Consequently, using Assumption 1 we deduce (37) by taking \( \mu = \frac{\gamma_k}{2(2^{-1}_{\sqrt{F}})} \).

Theorem 1.

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

Proof: Since \( C \) is compact, function \( \Omega \) is coercive. Moreover, it belongs to an o-minimal structure including semi-algebraic functions and logarithmic function, so that it satisfies Kurdyka-Łojasiewicz inequality [62], [63]. Therefore, by using Lemma 2 and [66] Theorem 4.1, we deduce that \( (x_k)_{k \in \mathbb{N}} \) converges to a critical point of \( \Omega \).

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 [64], anti-doping and proteomics [65], metabolomics [66], biomedical and biological analyses [67], [68], diagnosis process, cancer and tumors profiling [69], food contamination detection [70].

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 "averaging\(^{5}\)" model [71], [72].

Assuming that the charge state is known and mono-valued (see [73] for the multi-charged case), we propose to express the measured MS spectrum \( y \in \mathbb{R}^M \) as the sparse combination of individual isotopic patterns, i.e.

\[
y = \sum_{n=1}^{N} x_n d(m_{n,iso}^0, z) + b
\]

where \( d(m_{n,iso}^0, z) \in [0, +\infty[^M \) represents the mass distribution built with the "averaging\(^{5}\)" model at isotopic mass \( m_{n,iso}^0 \) 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 monoisotopic with mass \( m_{n,iso}^0 \). Moreover, \( b \in \mathbb{R}^M \) models the acquisition noise and some possible errors arising from the spectral analysis step. Let us form a dictionary matrix \( D \in \mathbb{R}^M \times N \) whose \( n \)-th column reads \( d(m_{n,iso}^0, z) \). Then, the above observation model (42) reads as (4), and the problem becomes the restoration of the sparse positive-valued signal \( x \), given \( y \) and \( D \). We proposed in [73] a restoration based on a penalized least squares problems with \( \ell_1 \) 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.

\(^{5}\)Determining an average amino acid from a statistical distribution.
B. Simulated datasets and settings

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

Figure 2 presents the sought isotopic distributions \( x \) and an example of associated MS spectra, for dataset A and B. In order to retrieve the original sparse signal, we will solve Problem (28) using \( \Theta \) defined in (5) and \( C = [0, x_{\text{max}}]^N \) with \( x_{\text{max}} = 10^5 \). Concerning the regularization function \( \Psi \), we will make comparisons between the \( \ell_1 \) norm, \( \ell_0 \), 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
\[
\Psi(x) = \sum_{n=1}^{N} \log(1 + x_n^2/\delta^2)
\]
with \( \delta > 0 \) [74], p. 111–112, the Welsch penalty
\[
\Psi(x) = \sum_{n=1}^{N} (1 - \exp(-x_n^2/\delta^2))
\]
with \( \delta > 0 \) [75], and the Continuous Exact \( \ell_0 \) penalty (CEL0) \( \Psi(x) = \sum_{n=1}^{N} (\delta - |x_n|/2 \sqrt{2/|x_n|})^2 1_{|x_n| < \sqrt{2/\delta}} \) where \( \delta > 0 \) and \( d_n \) is the \( n \)-th column of \( D \) [76, 77].

The resolution of (28) is performed by using the primal-dual splitting algorithm in [78, 79] in the case of \( \ell_1 \) norm, \( \ell_0 \) and CEL0 penalties. For Cauchy and Welsch penalties, we use the VMFB strategy, using the majorizing metrics described in [56]. 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 \( \Phi \) within the metric is computed by using the parallel proximal splitting algorithm from [80], with a maximum number of \( 5 \times 10^3 \) iterations. With the exception of \( \ell_1 \), 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 \( \ell_1 \) penalty. All algorithms were run until the stopping criterion defined as \( \|x_{k+1} - x_k\| \leq \epsilon \|x_k\| \) is satisfied (in our case, \( \epsilon = 10^{-4} \)), 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 \( \delta \) for Cauchy, Welsch and CEL0, and \((\alpha, \beta, \eta)\) for SPOQ. When not precised, hyperparameters were optimized with grid search to maximize the signal-to-noise ratio (SNR) defined as
\[
\text{SNR}(x, \hat{x}) = 20 \log_{10} \left( \frac{\| x \|_2}{\| x - \hat{x} \|_2} \right)
\]
where \( \hat{x} \) is the estimated signal and \( x \) the original one. Moreover, the bound \( \xi \) in (43) 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 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 \( x \) from the MS spectrum \( y \), especially when different isotopic peaks are present with different intensities in the same mass region.

C. Numerical results

1) Comparison of sparse penalties: Tables [4] and [11] 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 always yields the best performance, for a suitable choice of \( p \) and \( q \). Moreover, its estimated sparsity degree is the closest to the reference. 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 \( q \in \{ 2, 3 \} \). The \( \ell_0 \), \( \ell_1 \) and CEL0 regularization functions ensure a good TSNR. However, SPOQ shows its clear superiority, in terms of SNR, as it is able to better estimate the sought support of the signal. Finally, Cauchy and Welsch perform slightly below the other regularization methods, possibly as a consequence of the smoothing induced by parameter \( \delta \). 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 \( A_k = A_{q,0}(x_k) \) and \( \gamma_k = 1.9 \) for every \( k \in \mathbb{N} \). FB algorithm is obtained by setting \( A_k = LI_N \) 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 behave here quite similarly. This illustrates the advantage of our local preconditioning scheme.

3) Setting SPOQ parameters: In all our tests, the smoothing parameter \( \delta \) for Cauchy, Welch and CEL0 penalties were chosen empirically, so as to maximize the final SNR. Aside, we provide a pairwise sensitivity analysis for the \( \alpha, \beta \) and \( \eta \) parameters 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

### 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.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( q )</th>
<th>( \ell_p / \ell_q )</th>
<th>( \ell_0 )</th>
<th>( \ell_1 )</th>
<th>Cauchy</th>
<th>Welch</th>
<th>CEL0</th>
<th>( \delta = 0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>1.25</td>
<td>0.05</td>
<td>35.54</td>
<td>20.41</td>
<td>52.20</td>
<td>52.20</td>
<td>52.20</td>
<td>52.20</td>
</tr>
<tr>
<td>0.25</td>
<td>2.5</td>
<td>0.15</td>
<td>42.81</td>
<td>47.85</td>
<td>41.20</td>
<td>30.16</td>
<td>37.68</td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>5</td>
<td>0.25</td>
<td>48.08</td>
<td>54.41</td>
<td>46.77</td>
<td>44.08</td>
<td>48.67</td>
<td></td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( q )</th>
<th>( \ell_p / \ell_q )</th>
<th>( \ell_0 )</th>
<th>( \ell_1 )</th>
<th>Cauchy</th>
<th>Welch</th>
<th>CEL0</th>
<th>( \delta = 0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>1.25</td>
<td>0.05</td>
<td>35.54</td>
<td>20.41</td>
<td>52.20</td>
<td>52.20</td>
<td>52.20</td>
<td>52.20</td>
</tr>
<tr>
<td>0.25</td>
<td>2.5</td>
<td>0.15</td>
<td>42.81</td>
<td>47.85</td>
<td>41.20</td>
<td>30.16</td>
<td>37.68</td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>5</td>
<td>0.25</td>
<td>48.08</td>
<td>54.41</td>
<td>46.77</td>
<td>44.08</td>
<td>48.67</td>
<td></td>
</tr>
</tbody>
</table>
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, $\alpha$ seems to have little impact, especially when $\beta$ and $\eta$ are optimized. Note that we did not display the variations for fixed $\alpha$ as we observed that the SNR exhibits non-noticeable value variations. Parameter $\alpha$ essentially controls the $L$-Lipschitz value $\|L\|_{\alpha}$ and the derivability of $\ell_{p,\alpha}$ at 0 (see (7)).

4) Noise level influence: Using different penalties ($\ell_0$, $\ell_1$, Cauchy, Welsch, CEL0 and two instances of SPOQ), we present 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 $\sigma$ 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.

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 5 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 4 case of $p = 0.25$ 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 of setting parameter $p$.

Figure 3. SNR evolution along time, for the proposed trust-region VMFB algorithm [1] VMFB algorithm [56] and FB algorithm, to process datasets A and B on a given noise realization (relative noise level: 0.1%).

![Figure 3](image3.png)

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

![Figure 4](image4.png)

Figure 5. Influence of noise level ($\sigma$ expressed as a percentage of the MS spectrum maximal amplitude) on quality reconstruction of datasets A (left) and dataset B (right) using various penalties: SPOQ $\ell_0/\ell_2$, SPOQ $\ell_1/\ell_2$ (or SOOT), $\ell_0$, $\ell_1$, Cauchy, Welsch and CEL0 (SNR values averaged over 10 noise realizations).

![Figure 5](image5.png)

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

![Figure 6](image6.png)

V. CONCLUSION

SPOQ offers scale-invariant penalties, based on ratios of smoothed quasinorms and norms. These surrogates to the $\ell_0$ 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.

**APPENDIX A**

**PROOF OF PROPOSITION 1**

First, we have \( \nabla^p \Psi(0_N) = \nabla^p \Psi(0_N) = 0_N \), so that \( 0_N \) is a critical point of \( \Psi \). By using (17),

\[
\nabla^2 \Psi_1(0_N) = \frac{\omega^{p-2}}{\beta^p} I_N, \tag{44}
\]

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

\[
\nabla^2 \Psi_2(0_N) = \frac{1}{q} \left( \frac{q(q-1)}{\eta^2} \text{Diag}((0^q-2)_{1 \leq n \leq N}) - \frac{q}{\eta^2 q} 0_N \times 0_N \right) = 0_N \times 0_N, \tag{45}
\]

otherwise, if \( q = 2 \),

\[
\nabla^2 \Psi_2(0_N) = \frac{1}{2} \left( \frac{2(2-1)}{\eta^2} \text{Diag}((0^2)_{1 \leq n \leq N}) - \frac{2}{\eta^2 q} 0_N \times 0_N \right) = \frac{1}{\eta^2} I_N \tag{46}
\]

since \( 0^0 = 1 \) by convention. Consequently

\[
\nabla^2 \Psi_2(0_N) = \begin{cases} \frac{1}{\eta^2} I_N & \text{if } q = 2, \\ 0_N \times 0_N & \text{elsewhere}. \end{cases} \tag{47}
\]

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

Let us now show that, under suitable assumptions,

\[
(\forall x \in \mathbb{R}^N) \quad \Psi(x) \geq \Psi(0_N) \quad \iff \quad (p_{\omega, \alpha}(x) + \beta^p)^{1/p} \geq \frac{\beta}{\eta} \tag{48}
\]

that is

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

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,

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

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

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

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 \( I = \{1, \ldots, N\} \setminus I \). Since \( 2/p > 1 \),

\[
\left( 1 + \sum_{n=1}^{N} \frac{\alpha^p}{\beta^p} \left( \frac{z_n}{\alpha^2} + 1 \right)^{p/2} - 1 \right)^{2/p} \geq \left( 1 + \sum_{n \in I} \frac{\alpha^p}{\beta^p} \left( \frac{z_n}{\alpha^2} + 1 \right)^{p/2} - 1 \right)^{2/p}
\]

\[
+ \left( 1 + \sum_{n \in I} \frac{\alpha^p}{\beta^p} \left( \frac{z_n}{\alpha^2} + 1 \right)^{p/2} - 1 \right)^{2/p} \geq 1 + \sum_{n \in I} \frac{\alpha^p}{\beta^p} \left( 1 - \frac{2 - p}{4} \epsilon \right) z_n
\]

\[
+ \left( \sum_{n \in I} \frac{(\epsilon + 1)^{p/2} - 1}{\epsilon^{p/2} \beta^p} z_n^{p/2} \right)^{2/p}. \tag{53}
\]

If

\[
\frac{\alpha^p}{\beta^p} \left( 1 - \frac{2 - p}{4} \epsilon \right) \eta^2 \geq 1 \tag{54}
\]

\[
\left( \frac{(\epsilon + 1)^{p/2} - 1}{\epsilon^{p/2} \beta^p} \eta^2 \right)^{2/p} \eta^2 \geq 1, \tag{55}
\]

then

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

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

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

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

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

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

We deduce Inequality (49) by applying the triangle inequality for the \( L^{q/2} \) 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 \( \Psi \) is Lipschitz-differentiable by investigating the properties of \( \nabla^2 \Psi \). We start by studying
behavior of \(|\|\nabla^2 \Psi_1(x)\|\|\), where \(x \in \mathbb{R}^N\) and the spectral norm is denoted by \(|\|\cdot\|||\). Using (11) and (17), we obtain
\[
|||\nabla^2 \Psi_1(x)||| \leq \frac{|||\text{Diag}(Z)|||}{\ell_{p,\alpha}(x) + \beta p} + \frac{p|||Y|||^2}{(\ell_{p,\alpha}(x) + \beta p)^2}
\]
(60)
where we make use of the shorter notation:
\[
\begin{align*}
Y &= (x_n(x_n^2 + \alpha^2)^{q-1})_{1 \leq n \leq N} \\
Z &= \left(\left((p-1)x_n^2 + \alpha^2\right)(x_n^2 + \alpha^2)^{q-2}\right)_{1 \leq n \leq N}
\end{align*}
\]
(61)
First, we have
\[
\begin{align*}
|||\text{Diag}(Z)||| &= 1 \\
\leq \frac{1}{\ell_{p,\alpha}(x) + \beta p} \sup_{1 \leq n \leq N} (x_n^2 + \alpha^2)^{q-1}
\leq \frac{p\alpha p - 2}{\beta p}.
\end{align*}
\]
(62)
Since \(p < 2\) and, for every \(n \in \{1, \ldots, N\}\),
\[
|||\nabla^2 \Psi_1(x)||| \leq \frac{1}{\ell_{p,\alpha}(x) + \beta p} \sup_{1 \leq n \leq N} (x_n^2 + \alpha^2)^{q-1}
\leq \frac{p\alpha p - 2}{\beta p}.
\]
(63)
Besides, by setting \(\nu = \sum_{n=1}^{N} (x_n^2 + \alpha^2)^{p/2}\),
\[
\begin{align*}
\frac{|||Y|||^2}{(\ell_{p,\alpha}(x) + \beta p)^2} &= \frac{1}{(\ell_{p,\alpha}(x) + \beta p)^2} \sum_{n=1}^{N} x_n^2 (x_n^2 + \alpha^2)^{p/2 - 2}
\leq \frac{1}{(\ell_{p,\alpha}(x) + \beta p)^2} \sum_{n=1}^{N} x_n^2 (x_n^2 + \alpha^2)^{p/2 - 2}
\leq \frac{1}{(\ell_{p,\alpha}(x) + \beta p)^2} \sum_{n=1}^{N} (x_n^2 + \alpha^2)^{p/2 - 2}
\leq \frac{1}{(\ell_{p,\alpha}(x) + \beta p)^2} \sum_{n=1}^{N} (x_n^2 + \alpha^2)^{p/2 - 2}
\leq \frac{1}{2\alpha^2} \left(\frac{N\alpha p - \beta p}{\beta p}\right)^2
\leq \frac{1}{2\alpha^2} \left(1 + \frac{N\alpha p - \beta p}{\beta p}\right)^2
\leq \frac{1}{2\alpha^2} \max\left\{1, \left(\frac{N\alpha p - \beta p}{\beta p}\right)^2\right\}.
\end{align*}
\]
(64)
These results prove that \(\nabla^2 \Psi_1\) is bounded.
Let us now study the Hessian of \(\Psi_2\) at \(x \in \mathbb{R}^N\). Let \(\epsilon \in [0, +\infty]\), let
\[
\Lambda_\epsilon = \frac{\nabla^2 q_{q,\eta}(x) + q(q-1)\epsilon I_N}{\ell_{q,\eta}(x)},
\]
(67)
and let
\[
\nabla^2 \Psi_2(x) = \frac{1}{q} \left(\Lambda_\epsilon - \frac{\nabla q_{q,\eta}(x)}{\ell_{q,\eta}(x)}\right),
\]
(68)
By continuity,
\[
\lim_{\epsilon \to 0} |||\nabla^2 \Psi_2(x)||| = |||\nabla^2 \Psi_2(x)||| (69)
\]
On the other hand, since \(\Lambda_\epsilon\) is a positive definite matrix,
\[
\nabla^2 \Psi_2(x) = \frac{1}{q} \Lambda_\epsilon^{1/2}(I_N - v_e v_e^T)\Lambda_\epsilon^{1/2}
\]
(70)
where, by using (10),
\[
\begin{align*}
v_e &= \Lambda_\epsilon^{-1/2} \nabla q_{q,\eta}(x) \\
&= \Lambda_\epsilon^{-1/2} \frac{1}{\ell_{q,\eta}(x)} \left[\frac{\text{sign}(x_1)|x_1|^{q-1}}{\sqrt{|x_1|^{q-2} + \epsilon}}, \ldots, \frac{\text{sign}(x_N)|x_N|^{q-1}}{\sqrt{|x_N|^{q-2} + \epsilon}}\right]^T.
\end{align*}
\]
(71)
Therefore,
\[
|||\nabla^2 \Psi_2(x)||| = \frac{1}{q} |||\Lambda_\epsilon^{1/2}(I_N - v_e v_e^T)\Lambda_\epsilon^{1/2}||| (72)
\]
According to (67),
\[
\Lambda_\epsilon = \frac{q(q-1)}{\ell_{q,\eta}(x)} \text{Diag} \left((|x_n|^{q-2} + \epsilon)_{1 \leq n \leq N}\right). (73)
\]
Consequently,
\[
|||\Lambda_\epsilon||| = \frac{q(q-1)}{\ell_{q,\eta}(x)} \sup_{1 \leq n \leq N} (|x_n|^{q-2} + \epsilon). (74)
\]
We thus derive from (72) that
\[
|||\nabla^2 \Psi_2(x)||| = \frac{q(q-1)}{\ell_{q,\eta}(x)} \sup_{1 \leq n \leq N} (|x_n|^{q-2} + \epsilon)
\]
(75)
As \(\epsilon \to 0\), (69) yields
\[
|||\nabla^2 \Psi_2(x)||| \leq \frac{q - 1}{\ell_{q,\eta}(x)} \sup_{1 \leq n \leq N} |x_n|^{q-2} \max\{1, |||v|||^2 - 1\}
\]
(76)
where, according to (71),
\[
|||v|||^2 = \lim_{\epsilon \to 0} |||v_\epsilon|||^2 = \frac{q}{q - 1} \sum_{n=1}^{N} |x_n|^q (77)
\]
which is equivalent to
\[
|||v|||^2 - 1 = \frac{1}{q - 1} \left(1 - \frac{q}{\ell_{q,\eta}(x)}\right) (78)
\]
Since \((1 - \frac{q}{\ell_{q,\eta}(x)}) < 1\) and \(\frac{q}{q - 1} < 1\) for all \(q \in [2, +\infty]\), we deduce that \(|||v|||^2 - 1 < 1\). Resultingly,
\[
|||\nabla^2 \Psi_2(x)||| \leq \frac{q - 1}{\ell_{q,\eta}(x)} \sup_{1 \leq n \leq N} |x_n|^{q-2} (79)
\]
(79)
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 $\Psi$ is a Lipschitz-differentiable and
\[
|||\nabla^2 \Psi(x)||| = |||\nabla^2 \Psi_1(x) - \nabla^2 \Psi_2(x)|||
\leq |||\nabla^2 \Psi_1(x)||| + |||\nabla^2 \Psi_2(x)|||.
\] (81)

Using (65), (66), and (80), we conclude that
\[
|||\nabla^2 \Psi(x)||| \leq \frac{\alpha p^2}{\beta p} + \frac{p}{2\alpha^2} \max \left\{1; \left(\frac{N\alpha p}{\beta p}\right)^2 + \frac{q - 1}{p}\right\},
\]

hence $\Psi$ is Lipschitz differentiable with constant $L$ as in (21).

(ii) Let us now prove that $\Psi$ 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 (83) that for every $(x, \epsilon, x') \in (\mathbb{R}^N)^2$, and $n \in \{1, \ldots, N\}$:
\[
(x_n^2 + \alpha^2)^{p/2} \leq (x_n^2 + \alpha^2)^{p/2} + p\epsilon_n (x_n^2 + \alpha^2)^{p/2-1} (x_n' - x_n) + \frac{p}{2} (x_n^2 + \alpha^2)^{p/2-1} (x_n' - x_n)^2.
\] (82)

As a consequence,
\[
\ell_{p,\alpha}^p (x') \leq \ell_{p,\alpha}^p (x) + \langle x' - x, \nabla \ell_{p,\alpha}^p (x) \rangle + \frac{p}{2} (x_n^2 + \alpha^2)^{p/2-1} A_1 (x)(x' - x)
\]
where $A_1 (x) = \text{Diag}(x_n^2 + \alpha^2)^{p/2-1}_{1 \leq n \leq N}$. By using the Napier inequality expressed as
\[
\forall (u, v) \in \mathbb{R}^2 \quad \log u \leq \log v + \frac{u - v}{v},
\]
we get
\[
\Psi_1 (x') \leq \Psi_1 (x) + \langle x' - x, \nabla \Psi_1 (x) \rangle + \frac{1}{2}\|
\ell_{p,\alpha}^p (x) + \beta \rho \|^\top A_1 (x)(x' - x).
\] (84)

By applying the descent lemma to function $-\Psi_2$, and using (79) we obtain
\[
\forall (x, x') \in \bar{B}_{\Psi_2}^2
\Psi_2(x') \leq -\Psi_2 (x) - \langle x' - x, \nabla \Psi_2 (x) \rangle + \chi_{\sigma, \rho} \frac{1}{2} \|x' - x\|^2.
\] (85)

The majorization property is then derived from (84) and (85). The inequality (27) can be deduced in a straightforward manner, by noticing that both $\ell_{p,\alpha}^p (x)$ and $(x_n^2 + \alpha^2)^{p/2-1}$ are minimal for $x = 0_N$.

ACKNOWLEDGMENT

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

REFERENCES
