



**HAL**  
open science

# An Unrolled Half-Quadratic Approach for Sparse Signal Recovery in Spectroscopy

Mouna Gharbi, Emilie Chouzenoux, Jean-Christophe Pesquet

► **To cite this version:**

Mouna Gharbi, Emilie Chouzenoux, Jean-Christophe Pesquet. An Unrolled Half-Quadratic Approach for Sparse Signal Recovery in Spectroscopy. Inria Saclay. 2023. hal-04229774v1

**HAL Id: hal-04229774**

**<https://inria.hal.science/hal-04229774v1>**

Submitted on 5 Oct 2023 (v1), last revised 22 Dec 2023 (v2)

**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.



Distributed under a Creative Commons Attribution 4.0 International License

# An Unrolled Half-Quadratic Approach for Sparse Signal Recovery in Spectroscopy

Mouna Gharbi<sup>a</sup>, Emilie Chouzenoux<sup>a</sup>, Jean-Christophe Pesquet<sup>a</sup>

<sup>a</sup>*CVN, CentraleSupélec, Inria Saclay, University Paris Saclay, Gif-sur-Yvette, France*

---

## Abstract

This paper addresses the problem of sparse signal reconstruction frequently arising in analytical chemistry. Spectroscopic restoration is an important task that was tackled using two distinct approaches. First, model-based iterative methods were deployed. These methods rely on a strong theoretical background. However they face practical obstacles in terms of hyperparameter tuning. Conversely, approaches based on deep neural networks are fast and easy to implement. Nonetheless, they lack explainability, and their stability/robustness are ongoing research topics. In this paper, we propose a deep neural network based on unrolling a Majorization-Minimization (MM) Half-Quadratic (HQ) algorithm. This allows us to build interpretable layers mirroring iterations, making it possible to learn automatically data-driven hyperparameters such as regularization and stepsizes. Furthermore, we propose a dictionary of custom activation functions derived from potentials used in the original variational model. This interpretation of activations can be useful for analyzing the stability of neural networks. The efficiency of our method in comparison to iterative and learning-based methods is showcased through various experiments conducted on realistic mass spectrometry (MS) databases with various blur kernels and noise levels. All experiments show an improved reconstruction quality of chemical signals, while maintaining a low execution time at the test phase.

*Keywords:* Unrolling, Signal reconstruction, Half-Quadratic, Spectroscopy.

---

## 1. Introduction

Spectroscopy gathers data acquisition/processing pipelines aiming to decipher the chemical properties of a material or a compound. Let us cite,

for instance, mass spectrometry (MS) [1] and nuclear magnetic resonance (NMR) [2]. Despite constant improvements in physical acquisition devices, the acquisition process often remains noisy and prone to degradation. More formally, let  $\bar{x} \in \mathbb{R}^n$  be an original chemical signal we aim to estimate, and let  $y \in \mathbb{R}^m$  be an acquired measurement. In many spectroscopy applications (e.g., RMN DOSY [3], RMN relaxometry [4], MS [5]), both are linked by the following model

$$y = H\bar{x} + e, \quad (1)$$

where  $H \in \mathbb{R}^{m \times n}$  represents a measurement degradation process modeling for example a blur kernel [6] or a discrete Laplace transform [7], and  $e \in \mathbb{R}^m$  is an additive noise corrupting the data.

To draw relevant chemical interpretations, measurements  $y$  need to be processed to recover an estimate of the original data  $\bar{x}$ , which amounts to inverting Model (1).

In a real-life application context, this inversion process may encounter several difficulties. On the one hand, the ill-posedness of (1), linked to the poor conditioning of  $H$ , prevents the use of basic inversion or filtering approaches. On the other hand, the necessity to solve (1) multiple times, for each new acquisition, induces hard constraints on the processing time. To cope with the aforementioned challenges, two main strategies in literature have been considered. A first set of methods relies on optimization, where a criterion combining a data fidelity term and a penalization term is defined and minimized to estimate the signal of interest. This approach was employed in many spectroscopy restoration works such as [5, 8] for MS data processing, [4, 9, 10, 11, 12] in the context of NMR relaxometry, and in DOSY NMR [3].

Recently, a second set of approaches has emerged, which relies on supervised machine learning techniques applied on real or synthetic chemical signal databases to learn mappings (typically, deep neural networks) capable of producing an estimate of the sought signal from the degraded data. For example, dense convolutional neural networks (CNNs) were used to remove artifacts from high-dimensional NMR signals in [13] and, in [14], a residual convolutional neural network, was designed for compressive sensing spectroscopy.

Supervised learning-based methods can lead to high-quality results, as soon as a sufficiently large labeled database is available. Furthermore, they have proven to be fast since they benefit from the advances of GPU-based

deep learning programming frameworks. However, the interpretability and explainability of these methods remain open issues. They might suffer from instability issues [15][16], and guaranteeing their robustness remains a challenge [17]. In contrast, optimization-based methods offer the required theoretical guarantees, but tuning their hyperparameters might be a tedious task, and their execution time for CPU based implementations may be too high in large scale applicative contexts. For instance, the Matlab implementation<sup>1</sup> of the MS signal restoration method from [5] requires more than 1 minute to restore an MS spectra on a grid of  $n = 10^3$  mass values.

Lately, the two previous approaches have been combined, yielding an appealing class of techniques based on ‘unrolling’ (or ‘unfolding’). The core idea is to re-interpret existing iterative optimization algorithms as neural network architectures [18], where layers mirror iterations. A recent survey of unrolling techniques can be found in [19]. This new paradigm presents several advantages. First, new architectures can be proposed, which enjoy an interpretable network structure. Second, prior and/or physical knowledge about the problem at hand can be naturally embedded into the architecture. Third, a reduced execution time is obtained during the test phase, thanks to deep learning GPU-based frameworks. Lastly, unrolled architectures are more flexible than optimization algorithms, and in particular, they offer the possibility to learn some hyperparameters and/or operators of the original algorithm. For instance, stepsizes and regularization weights are learnt in [20], operators and penalty function parameters in [21], and shrinkage functions and biases are learnt in [22].

In various applications [19], unrolling techniques have been shown to outperform (sometimes, by far) existing methods, with a noticeable gain in explainability and robustness with respect to plain deep learning techniques, while leading to similar (fast) computing times. This paper investigates unrolled algorithms dedicated to the resolution of spectroscopy inverse problems. We explore specifically the unrolling of a popular class of optimization algorithms called Majorization-Minimization (MM) approaches [23]. Namely, we focus on the MM half-quadratic (HQ) approach [24] which minimizes sequentially quadratic majorizing surrogates of the objective function.

HQ was initially developed for image restoration applications [25], then

---

<sup>1</sup><https://fr.mathworks.com/matlabcentral/fileexchange/88897-spoq-smooth-sparse-p-over-q-ratio-regularization-toolbox>

deployed for chemical signal recovery in [26, 27]. Processing large-scale data with MM HQ has been explored in [25] by relying on subspace acceleration strategy [28, 29], or parallel CPU-based implementations [30, 31, 32, 33].

*Contributions:* The goal of this work is (i) the proposition of an unrolled MM HQ algorithm, for which an efficient GPU-based implementation is available, (ii) the construction of new activation functions within the proposed architecture in connection with choices of penalty functions, (iii) the design of architectures allowing to learn the involved regularization parameters, (iv) the validation of the proposed ideas on the restoration of signals arising from realistic databases of MS signals. *Link with existing works:* First, as we mentioned, deep learning strategies or optimization-based ones, have already been considered for tackling spectroscopy inverse problems. Up to our knowledge, this is however the first time unrolling is explored in such application field. Second, unrolling of MM methods remains scarcely explored in the literature. We can only mention [34] which proposes to unroll an HQ method for an application to image deblurring. Third, our analysis connecting penalty functions and activation layers generalizes the one presented in [35].

The paper is organized as follows, Section 2 introduces the optimization problem, and the MM HQ algorithm to solve it. Section 3 details our contributions, mainly the unrolling of HQ algorithm, the proposition of new activation functions, and the design of architectures for hyperparameter learning. Section 4 presents our experimental results and discussions. Lastly, Section 5 concludes the work.

## 2. Background on MM-based restoration methods

### 2.1. Problem formulation

A classical strategy for solving inverse problem (1) and producing an estimate  $\hat{x} \in \mathbb{R}^n$  of  $\bar{x}$  from observed data  $y$ , consists in solving a penalized least-squares minimization problem [36] of the form

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \left( F(x) = \frac{1}{2} \|Hx - y\|^2 + \lambda_1 \Psi_1(x) + \lambda_2 \Psi_2(x) \right). \quad (2)$$

Hereabove, the regularization functions  $\Psi_i: \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $i \in \{1, 2\}$ , aim at ensuring the stability of the solution  $\hat{x}$ , by enforcing *a priori* assumptions related to sparsity. A generic expression for such sparsity-promoting regularization

functions is given, for  $i \in \{1, 2\}$ , by

$$(\forall x \in \mathbb{R}^n) \quad \Psi_i(x) = \sum_{s=1}^n \psi_i([x]_s), \quad (3)$$

where  $\psi_i : \mathbb{R} \rightarrow [0, +\infty)$  is some positive-valued potentials, and  $[\cdot]_s$  with  $s \in \{1, \dots, n\}$  denotes the  $s$ -th component of a vector in  $\mathbb{R}^n$ .  $\lambda_1 > 0$  and  $\lambda_2 > 0$  are regularization hyperparameters balancing the weight between both regularization terms  $(\Psi_i)_{i \in \{1, 2\}}$ , and the least-squares fidelity in the objective function  $F$ .

Let us provide the rationale for adopting such a hybrid two-part penalty term. Sparsity-based priors are widely used in the signal processing literature so as to promote few nonzero entries in the sought signal (or in its representation in a transformed domain [37]). Two main families of such priors can be distinguished, namely (i) convex ones, (ii) non-convex ones. Convex sparse penalties include, as a notorious example, the  $\ell_1$  norm (related to the LASSO approach in machine learning [38]). An extension of this penalty is the elastic net regularizer [39]. Smooth approximations for it have also been very popular, such as the Huber penalty or the Fair potential [40]. Convex penalties are convenient since they lead to more tractable optimization problems associated with efficient solvers with sound convergence guarantees. However, they may lead to a bias in the estimated non-zero coefficients, which might be detrimental to the final quality of the results or to their quantitative interpretation. There is a plethora of non-convex priors, starting with the popular  $\ell_0$  quasi-norm up to its various approximations through norm ratios [41], piecewise defined penalties [42], and smoothed  $\ell_0$  potentials [43]. Such choices are closer, in essence, to the true goal of sparse restoration, as they directly aim at minimizing the number of significant components in the sought signal, without altering (too much) their amplitudes. Thus, they can lead to much better restoration quality, especially for very sparse signals, and avoid the biasing effect of the convex penalties. However, non-convexity has a significant cost in terms of difficulty of the optimization procedure, making the algorithms sometimes unstable and prone to getting stuck in spurious local minima. In this work, we propose to get the best of both worlds by mixing the two strategies, setting  $\psi_1$  as a convex sparsity-enhancing potential and  $\psi_2$  as a non-convex one. Furthermore, we opt for smooth (i.e., differentiable) regularizers, so as to preserve an efficient optimization procedure. It can be understood that, in such a context, the choice of proper regularization

parameters  $(\lambda_1, \lambda_2)$  will play a critical role. The unrolling approach will serve to decipher, in an automatic, supervised manner, the setting of them, as we will explain in Section 3.

## 2.2. Quadratic majorization-minimization algorithm

The estimate  $\hat{x}$  of  $\bar{x}$  is obtained by minimizing the penalized least squares formulation (2) by assuming that this optimization problem admits a solution. This is, in particular, guaranteed if the cost function in (2) is coercive, which is satisfied, for standard choices of function  $\Psi_1$ , provided that  $\lambda_1 > 0$ . Given our choices for  $\psi_1$  and  $\psi_2$ , a popular class of minimization schemes rely on the key concept of quadratic majorization-minimization we introduce hereafter.

Before presenting our algorithm, let us be more specific about the assumption on our potential functions  $\psi_1$  and  $\psi_2$ .

**Assumption 1.** For every  $i \in \{1, 2\}$ ,  $\psi_i : \mathbb{R} \rightarrow \mathbb{R}$  is

- (i) a differentiable, even function,
- (ii) increasing on  $[0, +\infty)$ ,
- (iii) such that  $\psi_i(\sqrt{\cdot})$  is concave on  $[0, +\infty)$ .

In addition,  $\psi_1$  is convex.

Assumption 1 is satisfied by a wide class of convex, and non-convex potentials, listed in the second column of Table 1 (see also [44]).

In the context of Problem (2) involving penalties satisfying Assumption 1, an efficient strategy consists of adopting a half-quadratic (HQ) optimizer [45, 24]. HQ belongs to the class of Majorization-Minimization (MM) algorithms, which is a prominent family of methods in the field of optimization [46, 47], with successful applications in the context of signal/image processing, telecommunications, and machine learning [48, 49]. MM techniques substitute the often complex problem of minimizing  $F$  by a sequence of more tractable problems, based upon the minimization of sequences of *majorant functions* (also called *surrogate functions* or *auxiliary functions*), satisfying the following conditions: at every iteration  $k \in \mathbb{N}$  of the algorithm,

$$(\forall x \in \mathbb{R}^n) \quad F(x) \leq Q(x, x_k), \quad F(x_k) = Q(x_k, x_k), \quad (4)$$

where  $x_k \in \mathbb{R}^n$  is the current iterate. Under Assumption 1, a quadratic majorant function for  $F$  at  $x_k \in \mathbb{R}^n$  is given by

$$(\forall x \in \mathbb{R}^n) \quad Q(x, x_k) = F(x_k) + \nabla F(x_k)^\top (x - x_k) + \frac{1}{2}(x - x_k)^\top A(x_k)(x - x_k), \quad (5)$$

where  $A(x_k)$  is symmetric definite positive matrix in  $\mathbb{R}^{n \times n}$ , defined as

$$(\forall x \in \mathbb{R}^n) \quad A(x) = H^\top H + \lambda_1 \text{Diag}\{(\omega_1([x]_s))_{1 \leq s \leq n}\} + \lambda_2 \text{Diag}\{(\omega_2([x]_s))_{1 \leq s \leq n}\},$$

where, for every  $i \in \{1, 2\}$ ,

$$(\forall u \in \mathbb{R} \setminus \{0\}) \quad \omega_i(u) = \frac{\varrho_i(u)}{u}, \quad \omega_i(0) = \lim_{u \rightarrow 0} \frac{\varrho_i(u)}{u}, \quad (6)$$

with  $\varrho_i: \mathbb{R} \rightarrow \mathbb{R}$  the derivative of the scalar potential function  $\psi_i$ . Note that the above limit is well-defined, thanks to Assumption 1 [45]. The expression of the derivatives for a bunch of potential functions satisfying Assumption 1 is provided in the third column of Table 1. Additional properties for the potential functions, also given in the table, will be discussed in Subsection 3.2.

Solving Problem (2) using the MM algorithm associated with quadratic majorant function (5) yields

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - \gamma_k A(x_k)^{-1} \nabla F(x_k), \quad (7)$$

where  $x_0 \in \mathbb{R}^n$  and, for every  $k \in \mathbb{N}$ ,  $A(x_k)^{-1}$  is the inverse of  $A(x_k)$ ,

$$\nabla F(x_k) = H^\top (Hx_k - y) + \lambda_1 \Omega_{1,k} x_k + \lambda_2 \Omega_{2,k} x_k,$$

with, for  $i \in \{1, 2\}$ ,

$$\Omega_{i,k} = \text{Diag}\{(\omega_i([x_k]_s))_{1 \leq s \leq n}\}. \quad (8)$$

$\{\gamma_k\}_{k \in \mathbb{N}} \subset (0, 2)$  is a sequence of stepsizes. Algorithm (7) can be viewed as a preconditioned gradient method, or a quasi-Newton method, where the scaling matrix multiplying the gradient descent term is built following the MM principle, in such a way that  $(F(x_k))_{k \in \mathbb{N}}$  decreases monotonically. It can also be interpreted as an alternating minimization method acting in an extended space, where the function to minimize is quadratic with respect to some auxiliary vector variable. This interpretation explains the name given



to HQ algorithm, and is actually a key element for proving its convergence [50]. The performance of HQ method has been illustrated in the context of signal restoration [45].

The matrix inversion required in (7) can be detrimental to a stable and fast practical implementation of this method. Moreover, setting the stepsize sequence, and the penalization hyperparameters  $(\lambda_1, \lambda_2)$  in the hybrid regularizer might be tedious. To this end, we propose in the next section an unrolling procedure that helps reducing computational time by creating an architecture with a given number of layers, benefiting from the advantages of implementations on GPUs, and with automatized hyperparameter tuning.

### 3. Proposed unrolled network

#### 3.1. Unrolled HQ architecture

The first important step for algorithm unrolling consists in translating the algorithm of interest into a multi-layer network architecture. Let us consider a given number  $K > 0$  of iterations for HQ algorithm. We will show in this section that each iteration  $k \in \{0, \dots, K-1\}$  of it can be mirrored by a layer  $\mathcal{L}_k$  of a feedforward network with a residual connection [51]. Let us rewrite the update rule of HQ algorithm (7) as

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = x_k - W_k \nabla F(x_k), \quad (9)$$

where  $x_0 \in \mathbb{R}^n$ ,

$$W_k = \gamma_k A(x_k)^{-1}, \quad (10)$$

and the gradient expression is given in (8). We now introduce linear and non linear operators that will play the role of linear layers and non-linear functions in the unrolled architecture. Let

$$V_0 = H^\top H, \quad V_1 = I_n, \quad V_2 = I_n, \quad V_3 = [I_n \quad \lambda_1 I_n \quad \lambda_2 I_n], \quad (11)$$

$$b_0 = -H^\top y, \quad b_1 = 0, \quad b_2 = 0, \quad b_3 = 0, \quad (12)$$

where  $I_n$  is the identity matrix of size  $n \times n$ . Moreover, let  $R_0$  to  $R_3$  be defined as, for every  $x \in \mathbb{R}^n$ ,

$$R_0(x) = R_3(x) = x, \quad R_1(x) = (\varrho_1([x]_s))_{1 \leq s \leq n}, \quad R_2(x) = (\varrho_2([x]_s))_{1 \leq s \leq n}. \quad (13)$$

Note these activations operate componentwise. Given the expressions (8) and (9), and the above definitions, we can rewrite the  $K$  first iterations of the HQ algorithm (7) as

$$(\forall k \in \{0, \dots, K-1\}) \quad x_{k+1} = x_k - W_k \left( R_3 \left( V_3 \begin{bmatrix} R_0(V_0 x_k + b_0) \\ R_1(V_1 x_k + b_1) \\ R_2(V_2 x_k + b_2) \end{bmatrix} + b_3 \right) \right), \quad (14)$$

with  $x_0 \in \mathbb{R}^n$ . Equation (14) has a remarkable structure, where we can identify fundamental bricks of a standard deep neural network architecture. For every  $k \in \{0, \dots, K-1\}$ , the input  $x_k$  of layer  $\mathcal{L}_k$  first goes simultaneously through the 3-channel linear layer with weights  $V_0, V_1, V_2$ , and biases  $b_0, b_1, b_2$ , then through the activation operators  $R_0, R_1, R_2$ . The three outputs of these activation blocks are concatenated and go through a linear layer with weight matrix  $V_3$ , aggregating them. This layer has bias  $b_3$ , and activation  $R_3$  (here reducing to identity). The output  $x_{k+1}$  is finally obtained by computing the difference between the input  $x_k$  and the output of  $R_3$  composed with the linear operator  $W_k$ . Such a last step is similar to a residual connection. Each layer of the network depends on the noisy observation. For every  $k \in \{0, \dots, K-1\}$  the  $k$ -th layer is denoted as  $\mathcal{L}_k^y$ , so that, for every  $k \in \{0, \dots, K-1\}$ ,

$$x_{k+1} = \mathcal{L}_k^y(x_k). \quad (15)$$

Figure 1 summarizes the structure of layer  $\mathcal{L}_k^y$  for some  $k \in \{0, \dots, K-1\}$ .

This equivalent formulation of the HQ algorithm presents several advantages. As we will demonstrate in Subsection 3.3, it is a first step towards the supervised learning of its hyperparameters such as the regularization parameters and the stepsizes. The optimization of these parameters is based on a dedicated loss function directly related to the application at hand (e.g., mean squared error on the restored signal). This deep unrolling strategy leads to the design of a deep architecture with interpretable structure, which contrasts with traditional deep networks. Namely, the involved weight and bias operators have specific forms derived from the original inverse problem formulation. The nonlinear activation functions are not chosen in an ad hoc manner, but they are fully interpretable in an optimization sense. This will be further discussed in Subsection 3.2. Another specificity of our architecture with respect to conventional DL ones is that the weight operator  $W_k$  depends on the  $k$ -th layer input. This weight operator thus exhibits some

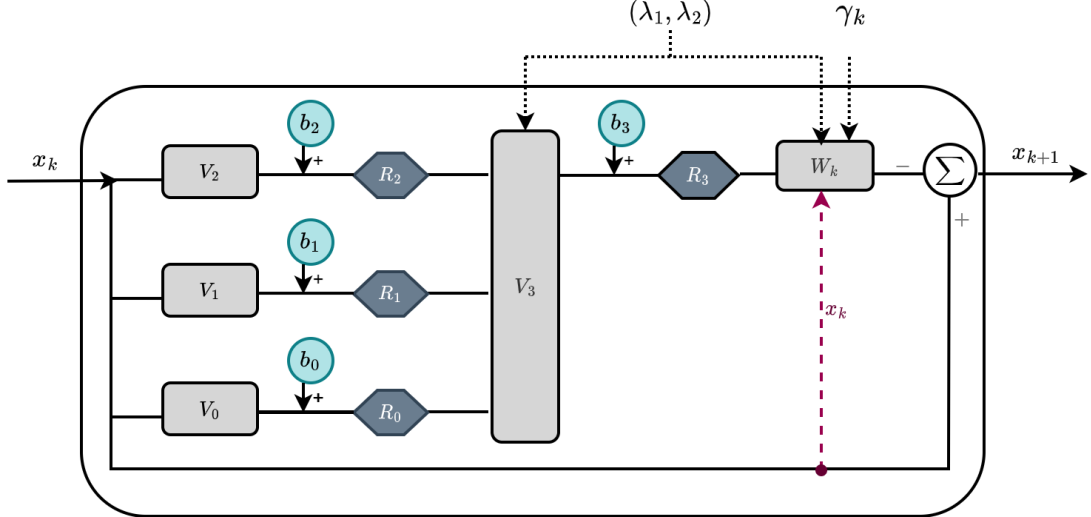


Figure 1: Unrolled HQ architecture: Layer  $\mathcal{L}_k^y$  mimicks the  $k$ -th iteration of the HQ algorithm (7), for  $k \in \{0, \dots, K - 1\}$ , and  $K > 0$ , as described in (14).  $V_0, V_1, V_2$  and  $V_3$  and  $W_k$  are weight operators while  $b_0, b_1, b_2, b_3$  are biases.  $R_0, R_1, R_2$  and  $R_3$  are activation functions.

similarity to a multivariate attention mechanism. Finally, from a practical viewpoint, the neural network-based interpretation of the HQ method allows the deployment of GPU-friendly tools to implement the different bricks of the optimizer. This is beneficial in terms of computational efficiency, during the training process, when supervised unrolling is performed. This is also useful for computational time reduction at the inference stage, with or without trained parameters, as we will demonstrate in our numerical experiments.

### 3.2. Activation functions

The architecture displayed in Fig. 1 makes use of several activation functions whose properties are further analysed in this section. Since  $R_0$  and  $R_3$  reduce to identity functions, we focus our discussion on  $R_1$  and  $R_2$ , which are directly related to our choice for the penalization functions  $\Psi_1$  and  $\Psi_2$  in (2). As can be seen in (13),  $R_1$  and  $R_2$  are applied componentwise. Their expressions involve the scalar potentials  $\varrho_1$  and  $\varrho_2$  that are the derivatives of  $\psi_1$  and  $\psi_2$ , respectively. We list in the third column of Table 1 the derivative  $\varrho$  for typical choices of function  $\psi$  satisfying Assumption 1. Interestingly, one can recognize that some correspond to commonly used activation functions

in neural networks, while others are less familiar. For example, the hyperbolic tangent is obtained for the Green potential, Eliott activation function corresponds to the Fair potential, and the Huber penalty leads to a shifted version of capped ReLU. As shown in [16], proximal properties of activation functions play an important role in the mathematical understanding of neural networks, in particular in their stability analysis. We analyse further the properties of  $\varrho$  in that respect. Let  $\Gamma_0(\mathbb{R})$  denote the class of lower-semicontinuous convex functions from  $\mathbb{R}$  to  $\mathbb{R} \cup \{+\infty\}$  which are proper (i.e., at least finite in one point). We recall that, if  $\varphi$  belongs to  $\Gamma_0(\mathbb{R})$ , its proximity operator  $\text{prox}_\varphi: \mathbb{R} \rightarrow \mathbb{R}$  is defined as

$$(\forall x \in \mathbb{R}) \quad \text{prox}_\varphi(x) = \underset{t \in \mathbb{R}}{\text{argmin}} \varphi(t) + \frac{1}{2}(t - x)^2. \quad (16)$$

We are now ready to state the following property.

**Proposition 1.** *Let  $\psi: \mathbb{R} \rightarrow \mathbb{R}$  be a differentiable even function and let  $\varrho$  be its derivative. Assume that  $\varrho$  is 1-Lipschitz on  $\mathbb{R}$ , i.e.*

$$(\forall (x, t) \in \mathbb{R}^2) \quad |\varrho(x) - \varrho(t)| \leq |x - t|.$$

1. *There exists  $\alpha \in [1/2, 1]$  and an even function  $\varphi \in \Gamma_0(\mathbb{R})$  such that*

$$(\forall x \in \mathbb{R}) \quad \varrho(x) = x + 2\alpha(\text{prox}_\varphi(x) - x). \quad (17)$$

2. *Let*

$$(\forall x \in \mathbb{R}) \quad \tilde{\varphi}(x) = \varphi(x) + \frac{x^2}{2} \quad (18)$$

*and let  $\tilde{\varphi}^* \in \Gamma_0(\mathbb{R})$  be the Fenchel-Young conjugate of  $\tilde{\varphi}$ .<sup>2</sup> Then*

$$(\forall x \in \mathbb{R}) \quad \psi(x) \stackrel{c}{=} (1 - 2\alpha)\frac{x^2}{2} + 2\alpha\tilde{\varphi}^*(x), \quad (19)$$

*where  $\stackrel{c}{=}$  designates equality up to an additive constant.*

3. *If  $\psi$  is convex, then  $\alpha = 1/2$ .*

*Proof.* See Section S1 of our supplementary material. □

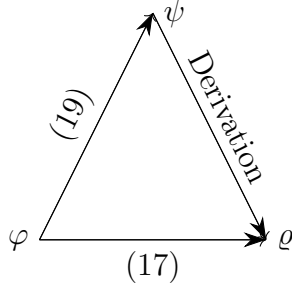


Figure 2: Interplay between penalty  $\psi$ , activation  $\rho$  and convex function  $\varphi$ .

The interplay between penalty function  $\psi$ , the activation  $\rho$ , and the underlying convex function  $\varphi$  is summarized in Fig. 2. The previous proposition ensures that, when  $\psi$  is a convex penalty function satisfying Assumption 1, there exists  $\varphi \in \Gamma_0(\mathbb{R})$  such that

$$\rho = \text{prox}_\varphi. \quad (20)$$

In this case,  $\rho$  enjoys the property of being firmly nonexpansive, i.e. increasing and 1-Lipschitz [52, Proposition 4.4]. As emphasized in [16], it turns out that most of the existing activation functions used in neural networks are proximity operators of convex functions (e.g. ReLU activation function). On the other hand, if  $\psi$  is non convex, then (17) shows that  $\rho$  is the over-relaxation of the proximity operator of a function in  $\Gamma_0(\mathbb{R})$ . According to (??),  $\rho$  is then  $\alpha$ -averaged, which means that it satisfies the inequality [52, Proposition 4.35] [53],

$$(\forall x, t \in \mathbb{R})^2 \quad |\rho(x) - \rho(t)|^2 \leq |x - t|^2 - \frac{1 - \alpha}{\alpha} |x - \rho(x) - t + \rho(t)|^2. \quad (21)$$

The smallest  $\alpha$ , the tightest the upper bound above. In other words, a small value of  $\alpha$  induces more stability. Equation (19) gives another interpretation of  $\alpha$ . It shows that  $\psi$  is a weakly-convex function [54, Proposition 4.3] with modulus  $2\alpha - 1$ . Thus,  $\alpha$  can be viewed as a measure of the degree of nonconvexity of potential  $\psi$ . To conclude, the activation functions in our unrolled architecture benefit from specific stability properties inherited from their optimization-based interpretation. This is in line with recent works [16, 35].

---

<sup>2</sup>The conjugate of  $\tilde{\varphi}$  is defined as  $(\forall u \in \mathbb{R}) \tilde{\varphi}^*(u) = \sup_{x \in \mathbb{R}} xu - \tilde{\varphi}(x)$ .

In [35], diffusion methods, wavelet approaches, and variational regularization are employed to perform a denoising task. The link with CNN architectures is made through residual neural blocks. Diffusivities, shrinkage functions, and regularizers are reinterpreted as activation functions. The authors emphasize the prospects of nonmonotone functions, rarely used in the deep learning community. Through our method, we obtain a dictionary of novel activation functions, associated to usual convex and non-convex sparsity promoting penalizations. We provide, the values for  $\alpha$  in the last column of Table 1. The convex function  $\varphi$  associated to a penalty  $\psi$  is implicitly defined by (18) and (19). As shown in Section S3 of our supplementary material, it is possible for most of the convex penalties in Table 1 to derive a closed form expression for  $\varphi$ .

Name	Penalization $\psi(t), t \in \mathbb{R}$	Activation $\varrho(t), t \in \mathbb{R}$	$\alpha$
<b>Convex penalties</b>			
Fair potential	$\delta( t  - \delta \log(\frac{ t }{\delta} + 1))$	$\frac{\delta t}{ t  + \delta}$	$\frac{1}{2}$
Huber	$\begin{cases} \frac{t^2}{2} & \text{if }  t  < \delta \\ \delta t  - \delta^2/2 & \text{otherwise} \end{cases}$	$\begin{cases} t & \text{if }  t  < \delta \\ \delta \text{sign}(t) & \text{otherwise} \end{cases}$	$\frac{1}{2}$
Green	$\log(\cosh(t))$	$\tanh(t)$	$\frac{1}{2}$
Hyperbolic	$\frac{\delta^2}{\kappa} \left( (1 + \frac{t^2}{\delta^2})^{\frac{\kappa}{2}} - 1 \right)$	$t(1 + \frac{t^2}{\delta^2})^{\frac{\kappa}{2}-1}$	$\frac{1}{2}$
<b>Nonconvex penalties</b>			
Welsch	$\delta^2 \left( 1 - \exp(-\frac{t^2}{2\delta^2}) \right)$	$t \exp(-\frac{t^2}{2\delta^2})$	$\exp(-\frac{3}{2}) + \frac{1}{2}$
Geman-McClure	$\frac{\delta^2 t^2}{2\delta^2 + t^2}$	$\frac{4\delta^4 t}{(2\delta^2 + t^2)^2}$	$\frac{5}{8}$
Tukey biweight	$\begin{cases} \delta^2(1 - (1 - \frac{t^2}{6\delta^2})^3) & \text{if }  t  \leq \sqrt{6}\delta \\ 1 & \text{otherwise} \end{cases}$	$\begin{cases} t(1 - \frac{t^2}{6\delta^2})^2 & \text{if }  t  \leq \sqrt{6}\delta \\ 0 & \text{otherwise} \end{cases}$	$\frac{9}{10}$
Hyperbolic tangent	$\delta^2 \tanh(\frac{t^2}{2\delta^2})$	$\frac{t}{(\cosh(\frac{t^2}{2\delta^2}))^2}$	0.9581
Cauchy	$\frac{\delta^2}{2} \log(1 + \frac{t^2}{\delta^2})$	$\frac{\delta^2 t}{t^2 + \delta^2}$	$\frac{9}{16}$

Table 1: Examples of penalties  $\psi$  satisfying Assumption 1. In the third column, we provide also the expression of their derivatives (i.e., nonlinear activation in our architecture)  $\varrho$  and, in the last column, the averaging constants  $\alpha$  (see Section 3.2). All expressions are valid for every  $t \in \mathbb{R}$ ,  $(\lambda, \delta) \in ]0, +\infty[^2$  and  $\kappa \in [1, 2]$ . The calculations for the averaging constants in non-convex cases are provided in Section S2 of our supplementary material.

### 3.3. Learning strategy

As aforementioned, the rewriting of HQ algorithm as a multi-layer neural network structure (see Fig. 1) paves the way for supervised learning of its

hyperparameters. In this subsection, we showcase our proposed strategy to tune automatically such parameters.

First, regularization parameters  $(\lambda_1, \lambda_2)$  are untied into two sequences  $(\lambda_{i,k})_{0 \leq k \leq K-1}$ ,  $i \in \{1, 2\}$ , varying with each layer of the proposed architecture. This is in contrast with classical optimization-based restoration procedures, which rely on setting those hyperparameters over all iterations, relying on empirical or statistical tuning techniques often difficult to implement [55]. Here, we propose instead to learn automatically data-driven parameters. For a given number of layers  $K > 0$ , each layer  $\mathcal{L}_k^y$  indexed by  $k \in \{0, \dots, K-1\}$  is attached to learnable units defined as

$$(\forall i \in \{1, 2\}) \quad \lambda_{i,k} = \text{ReLU}(\text{FC}_{i,k}([Hx_k - y]_{\ell})_{1 \leq \ell \leq m})). \quad (22)$$

Hereabove, ReLU is the rectified linear unit [56] applied componentwise, aiming at enforcing positivity on regularization parameters, and defined as follows:

$$(\forall z \in \mathbb{R}^m) \quad \text{ReLU}(z) = (\max(0, [z]_{\ell}))_{1 \leq \ell \leq m}. \quad (23)$$

Moreover,  $\text{FC}_{i,k}$ ,  $i \in \{1, 2\}$ ,  $k \in \{0, \dots, K-1\}$ , are fully connected layers acting on component-wise squares of the data fit residual at the entry of each layer  $\mathcal{L}_k^y$ . In such way, the regularization parameters automatically adapt to the noise level at hand, without any required prior knowledge. A similar strategy was employed in [20] in the context of image restoration.

Second, we suggest to also learn, in a supervised fashion, the stepsize sequence  $(\gamma_k)_{0 \leq k \leq K-1}$  involved in the linear weight layer sequence  $(W_k)_{0 \leq k \leq K-1}$ . The rationale is to tune optimally the contribution of each layer for delivering the best estimate  $\hat{x}$  at the output of the unrolled architecture. In order to reduce spurious local minima in the training phase, we adopt an overparametrization strategy [57], that consists in defining

$$(\forall k \in \{0, \dots, K-1\}) \quad \gamma_k = \text{ReLU}(c_{k,1} \times c_{k,2} \times \dots \times c_{k,N}), \quad (24)$$

with  $N \geq 1$  (typically,  $N = 10$ ), and  $(c_{k,j})_{0 \leq k \leq K-1, 1 \leq j \leq N}$  are scalar quantities that are learned. Vector  $\theta = (\theta_k)_{0 \leq k \leq K-1} \in \mathbb{R}^{K(2m+N)}$  gathers the parameters to be learnt, namely  $2mK$  weights for the fully-connected layers in (22), and  $NK$  values to learn the stepsize quantities as in (24).

The global multi-layer architecture, called **U-HQ** (**U**nrolled **H**alf-**Q**uadratic), is summarized in Fig. 3, with a detailed overview of a layer  $\mathcal{L}_k^{\theta_k}$  structure displayed in Fig. 4. Given a training set  $\{(\bar{x}^{(i)}, y^{(i)}), i \in \{1, \dots, S\}\}$  comprised

of  $S$  pairs of groundtruth and degraded data samples, the prediction feedforward model for one instance  $i \in \{1, \dots, S\}$  is given by

$$f_{\theta}(x_0^{(i)}; y^{(i)}) = \mathcal{L}_{K-1}^{y^{(i)}} \circ \dots \circ \mathcal{L}_k^{y^{(i)}} \circ \dots \circ \mathcal{L}_0^{y^{(i)}}(x_0^{(i)}), \quad (25)$$

$x_0^{(i)}$  denoting the input for the first layer of the network for the  $i$ -th sample. Then, the learnt hyperparameter vector  $\hat{\theta}$  is defined as the solution to the optimization problem

$$\underset{\theta \in \mathbb{R}^{K(2m+1)}}{\text{minimize}} \quad \frac{1}{S} \sum_{i=1}^S \mathcal{L}(f_{\theta}(x_0^{(i)}; y^{(i)}), \bar{x}^{(i)}), \quad (26)$$

with  $\mathcal{L}$  a predefined loss, for instance, the mean squared error. Problem (26) can be solved end-to-end through backpropagation with standard neural network training algorithms such as stochastic gradient or ADAM [58]. At the inference stage, for a given degraded signal  $y$ , and an initial prediction  $x_0$ , the inferred output is given by  $\hat{x} = f_{\hat{\theta}}(x_0; y)$ . The proposed approach benefits from automatic parameter tuning with layer-dependent and data driven structure, while having an interpretable structure.

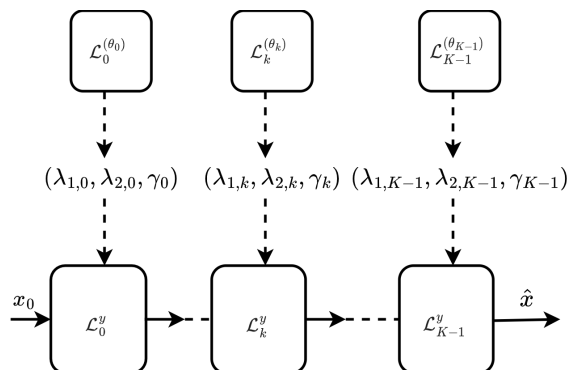


Figure 3: Supervised **U-HQ** network: For every  $k \in \{0, \dots, K-1\}$ , the detailed view of layer  $\mathcal{L}_k^y$  is given by Fig. 1. Details of the learning block  $\mathcal{L}_k^{\theta_k}$  are given in (22) and (24).

#### 4. Restoration of mass spectrometry signals

We now evaluate the performance of **U-HQ** on the restoration of MS sparse signals. We focus on the resolution of inverse problem (1) where



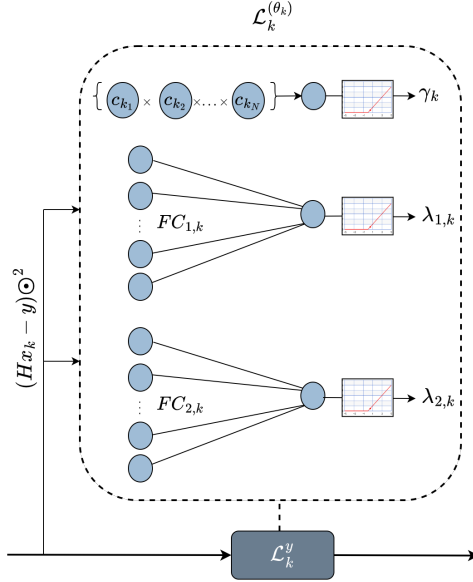


Figure 4: Overview of one layer of proposed architecture **U-HQ**.

$\bar{x} \in \mathbb{R}^n$  is a ground truth signal with length  $n = 2000$ ,  $H \in \mathbb{R}^{m \times n}$  is a Toeplitz matrix mimicking a circular padded convolution with a blur kernel, and  $e \in \mathbb{R}^m$  corresponds to an additive i.i.d. Gaussian noise with zero-mean and standard deviation  $\sigma > 0$ .

#### 4.1. Datasets

Let us first introduce the three datasets we employ to evaluate our method as well as its competitors. In all cases, the dataset is made of original signals  $\bar{x}$  paired with degraded signals  $y$  obtained through Model (1). Moreover, the noise standard variation is chosen randomly uniformly between 0 and 2, and assumed to be unknown. Finally, the blur support size is 50, so that  $m = 2050$ .

*Dataset 1.* Signals  $\bar{x}$  are real MS spectra extracted from the mass bank inventory of MS spectra [59], considering a monoisotopic mass range of [0,200] ppm and a charge state +1. The spectra intensities are rescaled to reach a maximum value of 100. The blur kernel is a centered Gaussian shape with standard deviation equal to one. Training, validation, and test sets contain respectively 4800, 100 and, 100 pairs  $(\bar{x}, y)$ .

*Dataset 2.* Signals  $\bar{x}$  are also real MS spectra extracted from the same mass spectrometry bank inventory [59], but we now consider the monoisotopic mass range [200,400] ppm and a charge state +1. Hereagain, the spectra intensities are rescaled to reach a maximum value of 100. The blur kernel is a “Ricker” shape with spreading factor 0.5 [60]. Training, validation, and test sets contain 1000, 100 and, 200 pairs  $(\bar{x}, y)$ , respectively.

*Dataset 3.* Groundtruth signals  $\bar{x}$  are identical to those from Dataset 1, but the blur kernel as well as the sets sizes are identical to the one used with Dataset 2. Training, validation, and test sets contain 1000, 100 and, 200 pairs  $(\bar{x}, y)$ , respectively.

#### 4.2. Experimental settings

*Evaluation of the results.* All compared supervised models are trained so as to solve (26), where  $\mathcal{L}$  is the averaged mean squared error (MSE) between the estimated signal  $\hat{x}$  and the ground truth  $\bar{x}$ ,  $\text{MSE}(\bar{x}, \hat{x}) = (1/n)\|\bar{x} - \hat{x}\|^2$ . MSE metric is also used to finetune regularization hyperparameters, as explained in a forthcoming section. In our comparative tables, we provide the signal-to-noise ratio (SNR, in dB) defined as  $\text{SNR}(\bar{x}, \hat{x}) = 20 \log_{10} (\|\bar{x}\|/\|\bar{x} - \hat{x}\|)$ , as well as the thresholded SNR metric (TSNR) in dB, defined as the SNR computed only on the support (i.e., indices of non-zero entries) of the sought signal  $\bar{x}$ .

*Penalties.* The regularization functions  $\Psi_1$  and  $\Psi_2$  in (2), are defined as in (3), with  $\psi_1$  the (convex) Fair potential and  $\psi_2$  the (non-convex) Cauchy penalty, whose mathematical expressions and properties are listed in Table 1. They depend on smoothing parameters  $\delta_1$  and  $\delta_2$ , respectively, which are set through a procedure precised hereafter. Within an ablation study, we will also compare to cases when only one of these penalty functions is employed (i.e., when we remove the network branch related to either  $\Psi_1$  or  $\Psi_2$ ).

*Initialization.* All the compared approaches are initialized by the null vector, in particular, all the compared neural networks take  $x_0 = 0$  as an input.

*Programming framework.* The methods are implemented using Pytorch (version 1.8.0) under Python (version 3.8.5). The GPU-based code is run on an Nvidia DGX server using a V100 SXM2 GPU card (32 GB of RAM). The CPU is an Intel® Xeon(R) W-2135 CPU @ 3.70 GHz with 12 active cores.

### 4.3. Comparison settings

The proposed **U-HQ** approach is compared to various methods, that can be classified into two categories, namely (i) optimization-based methods which rely on the HQ scheme to minimize (2), (ii) deep unrolling techniques relying on similar or distinct architecture than ours, with various training strategies. Let us specify hereafter the tuning for all the competing methods, including ours.

*Optimization-based methods.* We implement the HQ algorithm (7) described in Section 2.2 to minimize (2). We run HQ until convergence, that is either until it reaches a maximal number of iterations (here, 100), or until it shows a stabilization of its iterates through  $\|x_{k+1} - x_k\|^2 / \|x_k\|^2 < 10^{-5}$ . We denote **HQ-SC** this approach. We also consider a modified variant of HQ algorithm, where the number of iterations is fixed to a value  $K$  identical to the number of layers retained in the unrolled architecture. This method is inspired from early stopping (ES), therefore we name it **HQ-ES**.

For both methods, the stepsize is set to  $\gamma_k \equiv 1$  for simplicity. Moreover, the hyperparameters  $(\lambda_1, \lambda_2)$  are finetuned manually. Specifically, a gridsearch is performed on 200 signal randomly extracted from the training set to minimize the averaged MSE. The resulting parameters are retained later to assess the performance of these methods on the test set. Specifically, we perform the gridsearch using the convex and non-convex potentials separately to finetune both the regularization and smoothness hyperparameters  $(\lambda_1, \lambda_2, \delta_1, \delta_2)$ . Then, we run HQ algorithm using the hybrid penalty, with  $(\delta_1, \delta_2)$  fixed according to the previous disjoint gridsearch and we finetune the regularization hyperparameters  $(\lambda_1, \lambda_2)$ . The grid used for  $\lambda_1$  and  $\lambda_2$  is  $[0, 0.01, 0.1, 1, 2, 4, 6, 8, 10, 20]$ , while for  $\delta_1$  and  $\delta_2$  it is  $[0.01, 0.1, 0.5, 1, 2, 4, 6, 8]$ .

*Learning-based methods.* In order to assess the benefits of the proposed architecture in **U-HQ**, we provide comparisons with simplified versions of it, with an increasing number of free parameters, so as to perform an ablation study. We start with a low parametrized model inspired from deep equilibrium (DE) models [61]. Namely, the stepsize is set as  $\gamma_k \equiv 1$ , and the trained architecture shares the same two parameters  $(\lambda_1, \lambda_2)$  for all layers. So, for a fixed number of layers  $K$ , we have

$$\mathcal{L}_0^{(\theta_0)} = \mathcal{L}_1^{(\theta_1)} = \dots = \mathcal{L}_{K-1}^{(\theta_{K-1})} = \mathcal{L}^{(\theta)}. \quad (27)$$

This yields the so-called **U-HQ-DE** method. We then implement two architectures entitled **U-HQ-FixS** (**U-HQ Fix Stepsize**) and **U-HQ-FixN** (**U-HQ Fixed Noise**). In both cases, the noise estimation layer from Figure 4 is discarded, and the regularization hyperparameters  $(\lambda_{i,k})_{0 \leq k \leq K-1}$ , are simply defined as outputs of ReLU unit with  $2K$  learnable inputs  $(a_{i,k})_{0 \leq k \leq K-1}$ ,  $i \in \{1, 2\}$ . Moreover, in **U-HQ-FixS**, we furthermore impose  $\gamma_k \equiv 1$ , so that we learn only  $(\lambda_{i,k})_{0 \leq k \leq K-1}$ , while **U-HQ-FixN** learns stepsize hyperparameters  $(\gamma_k)_{0 \leq k \leq K-1}$  with enforced positivity using ReLU. Finally, we implement **U-HQ-FixN-OverP**. This last architecture uses **U-HQ-FixN** as a baseline, and includes the overparametrization strategy similar as the one described in (24), for the learning of both the regularization and stepsize hyperparameters per layer.

The unrolled schemes **U-HQ-FixS**, **U-HQ-FixN**, **U-HQ-FixN-OverP**, and the proposed **U-HQ**, are experimented using either  $\psi_1$  (convex) or  $\psi_2$  (non-convex), or  $(\psi_1, \psi_2)$  (hybrid), in their activation layers. Note that using only one of these two penalizations practically results in training half the number of weights. Regarding **U-HQ-DE** we present the results only for the hybrid penalty pair  $(\psi_1, \psi_2)$ , for the sake of conciseness. Furthermore, we make comparisons with two unrolled methods recently introduced for sparse signal recovery in [62], namely **U-ISTA** and **U-PD**, based on an unrolling of the ISTA algorithm [63] and the primal dual algorithm [64], respectively. Both methods rely on an  $\ell_1$  (thus convex) penalization on the signal. Regularization and stepsize parameters of both methods are learned, with enforced positivity using ReLU activations.

*Training settings.* Let us now present the training specifications used for the learning-based methods described in the previous subsection. The number of layers is set, for each dataset and each method, through empirical search, to maintain a trade-off between a low validation loss, a comparable number of parameters between all tested methods, a reasonable computational training time and a stable learning. For architectures **U-HQ**, **U-HQ-FixS**, **U-HQ-FixN**, **U-HQ-FixN-OverP** and **U-HQ-DE**, we set  $\delta_1 = 0.1$  and  $\delta_2 = 1$  in the unrolled HQ methods (while these parameters were finetuned in optimization-based methods). We set initial values of learnable weights/parameters to 1, and the parameter  $N$  to 10 in all our overparametrized settings. Training is performed using Adam optimizer [65], with mini-batch size 5. Regarding the training strategy, we set a unique learning rate  $lr$  to update the weights in most architectures, at the excep-

tion of **U-HQ**. As this latter uses completely different architectures for the stepsize (i.e., simple ReLU layer) and the regularization weights (i.e., fully connected layers), we use two different learning rates, namely  $lr_\gamma$  and  $lr_\lambda$ , respectively. Training parameters, including learning rates and layer number, are summarized in Section S4 of our supplementary material.

#### 4.4. Results

Table 3 summarizes the results in terms of average SNR/TSNR on test sets, for the three datasets. We recall that the higher the SNR (resp. the TSNR), the better the estimation. At a first glance, we confirm the competitiveness of our proposed architecture **U-HQ** with respect to all other compared methods (last line of the table).

First, let us inspect the results of optimization-based methods. Using the proposed hybrid penalty term with iterative methods **HQ-SC** and **HQ-ES** reaches better metrics compared to using convex or non-convex potentials separately in most cases on all datasets. This demonstrates the advantage of a hybrid penalty in this applicative problem. Second, from a general glance, we can notice that optimization-based methods are outperformed by learning-based ones. In particular, our proposed approach **U-HQ** yields better restoration quality compared to the use of a standard iterative HQ algorithm **HQ-SC** (up to 4dB improvement in both SNR/TSNR). Results are slightly improved for the iterative method when early stopping is applied, so that **HQ-ES** slightly outperforms **U-SC**. Hybrid penalty is clearly beneficial, when compared to either convex or non-convex penalty only, as it brings flexibility in the regularization model. However, it is at the price of a tedious grid search. Both **HQ-ES** and **HQ-SC** yield inferior performance to all their learning-based competitors, notably because the latter perform a supervised data-driven tuning of its hyperparameters to optimize the quality metric (here, MSE, which is directly related to SNR).

The interpretation of HQ algorithm as layers of a feed-forward neural network has allowed us to implement it efficiently on the GPU-based Pytorch environment, which considerably reduces its execution time, when compared to a more standard CPU implementation. Since it requires, in average, a lower number of layers/iterations than its optimization-based counterpart, **U-HQ** benefits from a reduced execution time compared to **HQ-SC**, as illustrated in Table 2. For the sake of comparison, we also display the test time of the second best performing method, **U-HQ-FixN** (i.e., **U-HQ** without noise estimation layer). One can notice that this architecture has similar test

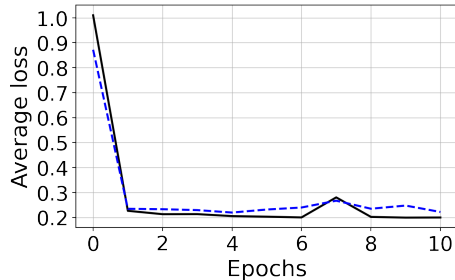


Figure 5: Evolution of training loss (solid black) and validation loss (dashed blue) along epochs for **U-HQ** for Dataset 1.

time than **U-HQ**, which shows that introducing data-driven fully connected layers (see Fig. 7) does not increase the network complexity.

In top of its good qualitative and computational performance, the training stability of **U-HQ**, and the absence of overfitting phenomenon can be assessed on the training loss curves on Dataset 1 given by Fig. 5. An example of restored signal in Dataset 1 is displayed in Fig. 6. We refer the reader to Section S5 of our supplementary file for similar plots on Datasets 2 and 3.

	Dataset 1	Dataset 2	Dataset 3
<b>HQ-SC</b>			
Iterations (averaged)	16.33	7.27	13.55
CPU time (s)	13.79	6.42	10.93
GPU time (s)	0.73	0.33	0.60
<b>U-HQ-FixN</b>			
Number of layers	8	16	12
GPU time (s)	0.17	0.32	0.24
<b>U-HQ-FixN-OverP</b>			
Number of layers	8	16	12
GPU time (s)	0.17	0.33	0.25
<b>U-HQ</b>			
Number of layers	8	8	12
GPU time (s)	0.17	0.17	0.25

Table 2: Averaged iteration/layer numbers and execution time on GPU per signal, for inference on test set on Dataset 1 to Dataset 3, using **HQ-SC**, **U-HQ-FixN**, **U-HQ-FixN-OverP** and **U-HQ**. For the former, we indicate also CPU time for the sake of comparison.

Third, the proposed architecture **U-HQ** appears superior to its simplified variants, namely **U-HQ-DE**, **U-HQ-FixS**, **U-HQ-FixN**, and **U-HQ-FixN-OverP**. This illustrates the importance of the learnable stepsize and hyper-parameter sequences, as well as the benefits of overparametrization.

A visual representation of learnt parameters is given in Fig. 7 for Dataset 1 (see Section S5 of our supplementary file for the other datasets). Interestingly, several regimes can be observed, namely boosting phases with higher regularization parameters and higher stepsizes, and stabilization phases where all parameters stabilize to constant values. Finally, it can be seen that, like for optimization-based methods, employing both penalty branches in the architecture of **U-HQ** has clear benefit on the performance when compared with the use of only the convex or non-convex term. The aforementioned observations confirm our choices for the final **U-HQ** architecture.

Lastly, the proposed **U-HQ** appears superior, in terms of both qualitative metrics, with respect to its competitors **U-PD** and **U-ISTA**, which again shows the benefit of considering a flexible and learnable regularization strategy in this applicative context.

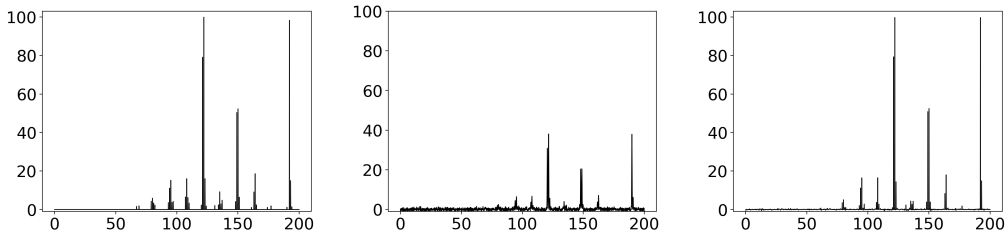


Figure 6: Groundtruth signals  $\bar{x}$  (left), degraded observations  $y$  (middle) and reconstruction  $\hat{x}$  using proposed method **U-HQ** (right), for an example from Dataset 1.  $x$ -axis indicates the mass-to-charge (m/z) ratio of chemical compounds and  $y$ -axis the abundance of corresponding compound. SNR/TSNR scores of restored signal (in dB) are 22.09/24.16, respectively.

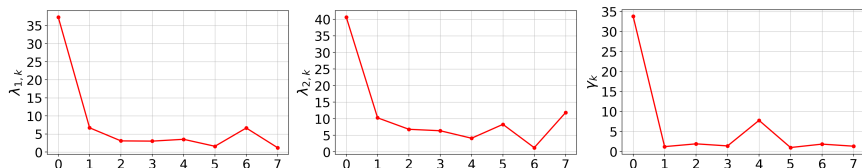


Figure 7: Learnt hyperparameter sequences, averaged on test set,  $(\lambda_{1,k})_{0 \leq k \leq K-1}$  (left),  $(\lambda_{2,k})_{0 \leq k \leq K-1}$  (middle) and  $(\gamma_k)_{0 \leq k \leq K-1}$  (right) with respect to layer index  $k$  ( $x$ -axis) for the proposed method **U-HQ**, for Dataset 1.

Penalty	Dataset 1	Dataset 2	Dataset 3
<b>HQ-SC</b>			
Convex	13.39/14.04	19.03/20.94	19.87/23.47
Non-convex	16.72/17.39	23.70/26.60	21.91/23.05
Hybrid	16.89/17.83	23.61/26.07	21.88/22.93
<b>HQ-ES</b>			
Convex	9.95/12.38	19.13/21.06	19.96/23.62
Non-convex	17.02/18.12	25.16/27.33	22.14/23.03
Hybrid	17.02/18.12	24.88/26.20	22.14/23.03
<b>U-ISTA</b>			
Convex	14.88/15.09	25.92/26.47	22.48/23.33
<b>U-PD</b>			
Convex	17.26/18.19	25.70/26.17	22.16/23.04
<b>U-HQ-DE</b>			
Hybrid	16.96/17.98	23.92/25.93	21.45/23.23
<b>U-HQ-FixS</b>			
Convex	12.09/13.77	22.97/24.58	20.71/22.43
Non-convex	17.12/18.05	25.09/26.31	21.85/23.08
Hybrid	17.13/17.88	25.08/26.36	21.86/23.11
<b>U-HQ-FixN</b>			
Convex	14.94/16.87	23.21/25.59	20.60/22.69
Non-convex	17.14/18.16	25.95/26.59	22.49/23.39
Hybrid	17.22/17.99	26.01/26.62	22.53/23.40
<b>U-HQ-FixN-OverP</b>			
Convex	17.15/18.34	25.05/27.06	22.20/23.51
Non-convex	17.35/18.58	25.87/26.24	22.22/22.81
Hybrid	17.52/18.48	26.00/26.50	22.54/23.53
<b>U-HQ</b>			
Convex	17.60/20.15	25.52/28.20	22.34/25.54
Non-convex	19.57/20.89	26.75/28.40	23.18/26.41
Hybrid	<b>19.73/21.16</b>	<b>26.82/28.90</b>	<b>24.00/25.83</b>

Table 3: Summary of results in terms of averaged SNR/TSNR in dB computed on test sets. Best scores are highlighted in bold.

## 5. Conclusion

In this work, we propose a novel approach to solve sparse recovery problems, motivated by a class of inverse problems arising in chemistry. The proposed method relies on merging a model-based half-quadratic iterative method and a data driven supervised approach, by adopting the unrolling paradigm. Unrolling the half-quadratic algorithm leads to an interpretable deep network architecture. Leveraging on automatic differentiation allows us to efficiently tune the algorithm hyperparameters. Furthermore, the resulting architecture turns out to be efficient in terms of reconstruction quality and computational budget, both of which are important features in any experimental setting. Our simulations illustrate the capability of our unrolling



technique on various realistic databases of mass spectrometry spectra degraded by different blur kernels and noise levels.

## Acknowledgments

This work has been supported by the ITN-ETN project TraDE-OPT funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 861137 and by the European Research Council Starting Grant MAJORIS ERC-2019-STG850925.

## References

- [1] L. Heaney, D. Jones, T. Suzuki, Mass spectrometry in medicine: a technology for the future?, *Future Science OA* 3 (Jun 2017).
- [2] R. R. Ernst, G. Bodenhausen, A. Wokaun, Principles of Nuclear Magnetic Resonance in One and Two Dimensions, 2nd ed., ser. International Series of Monographs on Chemistry. Oxford, NY: Oxford University Press, 19, 1987.
- [3] A. Cherni, E. Chouzenoux, M.-A. Delsuc, PALMA, an improved algorithm for DOSY signal processing, *Analyst* 142 (5) (2016) 772–779.
- [4] E. Chouzenoux, S. Moussaoui, J. Idier, F. Mariette, Optimization of a maximum entropy criterion for 2D Nuclear Magnetic Resonance reconstruction, in: Proceedings of the 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), 2010, pp. 4154–4157.
- [5] A. Cherni, E. Chouzenoux, L. Duval, J. C. Pesquet, SPOQ  $\ell_p$ -over- $\ell_q$  regularization for sparse signal recovery applied to mass spectrometry, *IEEE Transactions on Signal Processing* 68 (2020) 6070–6084.
- [6] H. Liu, T. Zhang, L. Yan, H. Fang, Y. Chang, A map-based algorithm for spectroscopic semi-blind deconvolution, *Analyst* 137 (2012) 3862–3873.
- [7] E. Lin, V.-V. Telkki, X. Lin, C. Huang, H. Zhan, Y. Yang, Y. Huang, Z. Chen, High-resolution reconstruction for multidimensional Laplace NMR, *The Journal of Physical Chemistry Letters* 12 (2021) 5085–5090.

- [8] A. Mohammad-Djafari, J.-F. Giovannelli, G. Demoment, J. Idier, Regularization, maximum entropy and probabilistic methods in mass spectrometry data processing problems, *International Journal of Mass Spectrometry* 215 (1) (2002) 175–193.
- [9] E. Chouzenoux, S. Moussaoui, J. Idier, F. Mariette, Efficient maximum entropy reconstruction of Nuclear Magnetic Resonance  $T_1 - T_2$  spectra, *IEEE Transactions on Signal Processing* 58 (12) (2010) 6040–6051.
- [10] Y.-Q. Song, L. Venkataramanan, M. D. Hürlimann, M. Flaum, P. Frulla, C. Straley,  $T_1 - T_2$  correlation spectra obtained using a fast two-dimensional Laplace inversion., *Journal of Magnetic Resonance* 154 2 (2002) 261–8.
- [11] C. Rondeau-Mouro, R. Kovrlija, E. Van Steenberge, S. Moussaoui, Two dimensional IR-FID-CPMG acquisition and adaptation of a maximum entropy reconstruction, *Journal of Magnetic Resonance* 265 (2016) 16–24.
- [12] L. Venkataramanan, Y.-Q. Song, M. Hurlimann, Solving Fredholm integrals of the first kind with tensor product structure in 2 and 2.5 dimensions, *IEEE Transactions on Signal Processing* 50 (5) (2002) 1017–1026.
- [13] X. Qu, Y. Huang, H. Lu, T. Qiu, D. Guo, T. Agback, V. Orekhov, Z. Chen, Accelerated Nuclear Magnetic Resonance spectroscopy with deep learning, *Angewandte Chemie (International ed. in English)* 59 (26) (2020) 10297–10300.
- [14] C. Kim, D. Park, H.-N. Lee, Compressive sensing spectroscopy using a residual convolutional neural network, *Sensors* 20 (2020) 594.
- [15] V. Antun, F. Renna, C. Poon, B. Adcock, A. Hansen, On instabilities of deep learning in image reconstruction and the potential costs of AI, *Proceedings of the National Academy of Sciences* 117 (2020) 30088–30095.
- [16] P. Combettes, J.-C. Pesquet, Lipschitz certificates for layered network structures driven by averaged activation operators, *SIAM Journal on Mathematics of Data Science* 2 (2020) 529–557.

- [17] M. Genzel, J. Macdonald, M. März, Solving inverse problems with deep neural networks – robustness included?, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (1) (2023) 1119–1134.
- [18] J. Hershey, J. Le Roux, F. Weninger, Deep unfolding: Model-based inspiration of novel deep architectures, Tech. rep., <https://arxiv.org/abs/1409.2574> (Sep. 2014).
- [19] V. Monga, Y. Li, Y. C. Eldar, Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing, *IEEE Signal Processing Magazine* 38 (2) (2021) 18–44.
- [20] C. Bertocchi, E. Chouzenoux, M. Corbineau, J.-C. Pesquet, M. Prato, Deep Unfolding of a Proximal Interior Point Method for Image Restoration, *Inverse Problems* 36 (3) (Feb. 2020.) 034005.
- [21] D. Ito, S. Takabe, T. Wadayama, Trainable ISTA for sparse signal recovery, *IEEE Transactions on Signal Processing* 67 (12) (2019) 3113–3125.
- [22] J. Sun, Z. Xu, Color image denoising via discriminatively learned iterative shrinkage, *IEEE Transactions on Image Processing* 24 (2015) 4148–59.
- [23] Y. Sun, P. Babu, D. Palomar, Majorization-Minimization algorithms in signal processing, communications, and machine learning, *IEEE Transactions on Signal Processing* 65 (2017) 794–816.
- [24] P. Charbonnier, L. Blanc-Feraud, G. Aubert, M. Barlaud, Deterministic edge-preserving regularization in computed imaging, *IEEE Transactions on Image Processing* 6 (2) (1997) 298–311.
- [25] D. Geman, C. Yang, Nonlinear image recovery with half-quadratic regularization, *IEEE Transactions on Image Processing* 4 (7) (1995) 932–946.
- [26] V. Mazet, D. Brie, J. Idier, Baseline spectrum estimation using half-quadratic minimization, in: *Proceedings of the 12th European Signal Processing Conference (EUSIPCO 2004)*, 2004, pp. 305–308.
- [27] J. Liu, J. Liu, J. Sun, X. Huang, G. Li, B. Liu, Goldindec: A novel algorithm for Raman spectrum baseline correction, *Applied Spectroscopy* 69 (7) (2015).

- [28] E. Chouzenoux, J. Idier, S. Moussaoui, A majorize–minimize strategy for subspace optimization applied to image restoration, *IEEE Transactions on Image Processing* 20 (6) (2011) 1517–1528.
- [29] E. Chouzenoux, A. Jeziarska, J.-C. Pesquet, H. Talbot, A Majorize–Minimize subspace approach for  $\ell_2$ - $\ell_0$  image regularization, *SIAM Journal on Imaging Sciences* 6 (1) (2013) 563–591.
- [30] S. Cadoni, E. Chouzenoux, J.-C. Pesquet, C. Chaux, A block parallel Majorize–Minimize memory gradient algorithm, in: *Proceedings of the 23rd IEEE International Conference on Image Processing (ICIP 2016)*, 2016, pp. 3194–3198.
- [31] M. Chalvidal, E. Chouzenoux, Block distributed 3MG algorithm and its application to 3D image restoration, in: *Proceedings of the 27th IEEE International Conference on Image Processing (ICIP 2020)*, 2020.
- [32] M. Hong, M. Razaviyayn, Z.-Q. Luo, J.-S. Pang, A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing, *IEEE Signal Processing Magazine* 33 (1) (2016) 57–77.
- [33] A. Breloy, Y. Sun, P. Babu, D. Palomar, Block Majorization–Minimization algorithms for low-rank clutter subspace estimation, in: *Proceedings of the 24th European Signal Processing Conference (EU-SIPCO 2016)*, 2016, pp. 2186–2190.
- [34] Y. Li, M.-R. Tofghi, V. Monga, Y. C. Eldar, An algorithm unrolling approach to deep image deblurring, *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)* (2019) 7675–7679.
- [35] T. Alt, J. Weickert, P. Peter, Translating diffusion, wavelets, and regularisation into residual networks, *Tech. rep.*, <https://arxiv.org/abs/2002.02753> (June 2020).
- [36] R. A. Willoughby, Solutions of ill-posed problems (A. N. tikhonov and V. Y. arsenin), *SIAM Reviews* 21 (2) (1979) 266–267.

- [37] M. Elad, P. Milanfar, R. Rubinstein, Analysis versus synthesis in signal priors, in: Proceedings of the 14th European Signal Processing Conference (EUSIPCO 2006), 2006, pp. 1–5.
- [38] T. Park, G. Casella, The Bayesian Lasso, *Journal of the American Statistical Association* 103 (482) (2008) 681–686.
- [39] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B* 67 (2005) 301–320.
- [40] A. Guitton, W. W. Symes, Robust inversion of seismic data using the Huber norm, *Geophysics* 68 (4) (2003) 1310–1319.
- [41] A. Repetti, M. Q. Pham, L. Duval, E. Chouzenoux, J. C. Pesquet, Euclid in a Taxicab: Sparse blind deconvolution with smoothed  $\ell_1/\ell_2$  regularization, *IEEE Signal Processing Letters* 22 (5) (2015) 539–543.
- [42] E. Soubies, L. Blanc-Féraud, G. Aubert, A continuous exact  $\ell_0$  penalty (CEL0) for least-squares regularized problem, *SIAM Journal on Imaging Science* 8 (3) (2015) 1574–1606.
- [43] H. Mohimani, M. Babaie-Zadeh, C. Jutten, A fast approach for overcomplete sparse decomposition based on smoothed  $\ell_0$  norm, *IEEE Transactions on Signal Processing* 57 (1) (2009) 289–301.
- [44] E. Chouzenoux, J. Pesquet, A stochastic Majorize-Minimize subspace algorithm for online penalized least squares estimation, *IEEE Transactions on Signal Processing* 65 (18) (2017) 4770–4783.
- [45] M. Nikolova, M. K. Ng, Analysis of half-quadratic minimization methods for signal and image recovery, *SIAM journal on Scientific Computing* 27 (2005) 937–966.
- [46] Z. Zhang, J. Kwok, D.-Y. Yeung, Surrogate maximization/minimization algorithms and extensions, *Machine Learning* 69 (2007) 1–33.
- [47] D. R. Hunter, K. Lange, A tutorial on MM algorithms, *The American Statistician* 58 (1) (2004) 30–37.
- [48] J. Mairal, Incremental majorization-minimization optimization with application to large-scale machine learning, *SIAM Journal on Optimization* 25 (02 2014).

- [49] Y. Sun, P. Babu, D. Palomar, Majorization-Minimization algorithms in signal processing, communications, and machine learning, *IEEE Transactions on Signal Processing* 65 (2017) 794–816.
- [50] M. Allain, J. Idier, Y. Goussard, On global and local convergence of half-quadratic algorithms, *Image Processing, IEEE Transactions on* 15 (2006) 1130 – 1142.
- [51] P. L. Combettes, J.-C. Pesquet, Deep neural network structures solving variational inequalities, *Set-Valued and Variational Analysis* 28 (2020) 491–518.
- [52] H. H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd Edition, Springer Publishing Company, Incorporated, 2017.
- [53] P. L. Combettes, J.-C. Pesquet, Fixed point strategies in data science, *IEEE Transactions on Signal Processing* 69 (2021) 3878–3905.
- [54] J.-P. Vial, Strong and weak convexity of sets and functions, *Mathematics of Operations Research* 8 (2) (1983) 231–259.
- [55] F. Bauer, M. A. Lukas, Comparing parameter choice methods for regularization of ill-posed problems, *Mathematics and Computers in Simulation* 81 (9) (2011) 1795–1841.
- [56] M. D. Zeiler, M. Ranzato, R. Monga, M. Z. Mao, K. Yang, Q. V. Le, P. Nguyen, A. W. Senior, V. Vanhoucke, J. Dean, G. E. Hinton, On rectified linear units for speech processing, *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)* (2013) 3517–3521.
- [57] S. Bell-Kligler, A. Shocher, M. Irani, Blind super-resolution kernel estimation using an internal-GAN, in: *Proceedings of the Conference on Neural Information Processing Systems 2019 (NEURIPS 2019, Vancouver, Canada, 2019)*.
- [58] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *Tech. rep.*, <https://arxiv.org/abs/1412.6980> (2015).
- [59] <https://massbank.eu/MassBank/>.

- [60] N. H. Ricker, The form and nature of seismic waves and the structure of seismograms, *Geophysics* 5 (4) (1940) 348–366.
- [61] D. Gilton, G. Ongie, R. Willett, Deep equilibrium architectures for inverse problems in imaging, *IEEE Transactions on Computational Imaging* 7 (2021) 1123–1133.
- [62] M. Gharbi, S. Villa, E. Chouzenoux, J.-C. Pesquet, Unrolled primal-dual deep network for sparse signal restoration, *Tech. rep.*, <https://hal.inria.fr/hal-03988686> (2022).
- [63] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring, in: *Proceedings of the 34th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, IEEE, 2009, pp. 693–696.
- [64] A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision* 40 (2011) 120–145.
- [65] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).

# An Unrolled Half-Quadratic Approach for Sparse Signal Recovery in Spectroscopy (supplementary)

Mouna Gharbi<sup>a</sup>, Emilie Chouzenoux<sup>a</sup>, Jean-Christophe Pesquet<sup>a</sup>

<sup>a</sup>*CVN, CentraleSupélec, Inria Saclay, University Paris Saclay, Gif-sur-Yvette, France*

---

## Abstract

This supplementary file is organized as follows. In Section S1, we present our proof for Proposition 1. In Section S2, we provide the calculation details for  $\alpha$ -averaging constants of the activation functions  $\varrho$  of Table 1. Then, in Section S3, we derive the expression of the convex function  $\varphi$  which was introduced in Proposition 1. We finally provide our training settings in Section S4 and display in Section S5 additional figures, complementing the Section 4 of the paper.

---

## S1. Proof of Proposition 1

- If  $\varrho$  is 1-Lipschitz then there exists  $\alpha \in [\frac{1}{2}, 1]$  and a 1-Lipschitz function  $\vartheta$  such that

$$(\forall x \in \mathbb{R}) \quad \varrho(x) = (1 - \alpha)x + \alpha\vartheta(x). \quad (\text{E1})$$

(In the worst case, one can simply choose  $\alpha = 1$ .) A function  $\varrho$  satisfying (E1) is said to be  $\alpha$ -averaged. The existence of  $\varphi \in \Gamma_0(\mathbb{R})$  such that (17) holds then follows from [R2, Proposition 3.1(ii)].

Since  $\psi$  is even by assumption, its derivative  $\varrho$  is odd, and we deduce from (17) that  $\text{prox}_\varphi$  is odd too. For every  $p \in \mathbb{R}$ , let  $u \in \partial\varphi(p)$  and let  $x = u + p$ . Then

$$\begin{aligned} x - p &\in \partial\varphi(p) \\ \Leftrightarrow p &= \text{prox}_\varphi(x) \\ \Leftrightarrow -p &= \text{prox}_\varphi(-x) \\ \Leftrightarrow -u &= -x + p \in \partial\varphi(-p), \end{aligned} \quad (\text{E2})$$



where  $\partial\varphi$  denotes the subdifferential of  $\varphi$  (see [R3]). This shows that  $\partial\varphi$  is odd.

According to [R3, Proposition 16.17],

$$\emptyset \neq \text{int}(\text{dom } \varphi) \subset \text{dom } \partial\varphi, \quad (\text{E3})$$

where

$$\text{dom } \varphi = \{x \in \mathbb{R} \mid \varphi(x) < +\infty\} \quad (\text{E4})$$

$$\text{dom } \partial\varphi = \{x \in \mathbb{R} \mid \partial\varphi(x) \neq \emptyset\}, \quad (\text{E5})$$

and  $\text{int}(S)$  denotes the interior of a set  $S$ . There thus exists  $\bar{p} \in \mathbb{R}$  and  $\epsilon > 0$  such that

$$\begin{aligned} \bar{p} \in \text{int}(\text{dom } \varphi) &\subset \text{dom } \partial\varphi \\ &= \text{dom } \partial\check{\varphi} \end{aligned} \quad (\text{E6})$$

$$\subset \text{dom } \check{\varphi}, \quad (\text{E7})$$

where  $\check{\varphi}: x \mapsto \varphi(-x)$ . Equality in (E6) follows from the fact that, for every  $p \in \mathbb{R}$ ,  $\partial\check{\varphi}(p) = -\partial\varphi(-p) = \partial\varphi(p)$  since  $\partial\varphi$  is odd. Then, we deduce from (E7) that  $\bar{p} \in \text{int}(\text{dom } \varphi) \cap \text{dom } \check{\varphi}$ .

Let us consider the even part  $\varphi_e$  of  $\varphi$  defined as

$$\varphi_e = \frac{1}{2}(\varphi + \check{\varphi}). \quad (\text{E8})$$

Since we have shown that  $\text{int}(\text{dom } \varphi) \cap \text{dom } \check{\varphi}$  is nonempty, it follows from [R3, Theorem 16.47] that

$$\partial\varphi_e = \frac{1}{2}(\partial\varphi + \partial\check{\varphi}) = \partial\varphi. \quad (\text{E9})$$

As  $\varphi_e$  and  $\varphi$  have the same subdifferential,  $\text{prox}_{\varphi_e} = \text{prox}_{\varphi}$ . This shows that, in (17), we can assume that  $\varphi$  is even.

- It follows from [R3, Corollary 24.5] that  $\tilde{\varphi}^*$  is a differentiable function such that

$$\text{prox}_{\varphi} = (\tilde{\varphi}^*)' \quad (\text{E10})$$

where  $(\cdot)'$  denotes the derivative of the function in argument. We deduce from (17) that

$$(\forall x \in \mathbb{R}) \quad \psi'(x) = \varrho(x) = (1 - 2\alpha)x + 2\alpha(\tilde{\varphi}^*)'(x). \quad (\text{E11})$$

By integrating, this yields

$$(\forall x \in \mathbb{R}) \quad \psi(x) = (1 - 2\alpha)\frac{x^2}{2} + 2\alpha\tilde{\varphi}^*(x) + C, \quad (\text{E12})$$

where  $C \in \mathbb{R}$ .

- If  $\psi$  is convex, its derivative  $\varrho$  is increasing. Since  $\varrho$  is defined on  $\mathbb{R}$ , it follows from [R2, Proposition 3.1(iii)] that  $\alpha = 1/2$ .

## S2. $\alpha$ -averaging constants of activation functions

In this section, we present calculation details for  $\alpha$ -averaging constants of the activation functions  $\varrho$  of Table 1, as defined in Proposition 1. Firmly nonexpansive activations  $\varrho$  have an  $\alpha$  constant equal to  $\frac{1}{2}$ . This applies to the first five rows of Table 1. Let us now calculate  $\alpha$  for the non-monotonic activation functions  $\varrho$  listed in lines 5 – 10 of Table 1. It follows from (E1) that  $\alpha \in ]1/2, 1]$  is the smallest constant such that  $\vartheta = \text{Id} + \alpha^{-1}(\varrho - \text{Id})$  is 1-Lipschitz, i.e., it satisfies the condition

$$\sup_{t \in \mathbb{R}} |\vartheta'(t)| = 1 \quad (\text{E13})$$

### • Welsch:

It follows from the expression of Welsch activation function that (E13) is equivalent to

$$\sup_{t \in \mathbb{R}} \left| 1 + \alpha^{-1} \left( \exp\left(\frac{-t^2}{2\delta^2}\right) \left(1 - \frac{t^2}{\delta^2}\right) - 1 \right) \right| = 1. \quad (\text{E14})$$

The zeros of the derivative of

$\vartheta'$  are

$$t = 0, \quad t = \sqrt{3}\delta, \quad \text{and} \quad t = -\sqrt{3}\delta. \quad (\text{E15})$$

Condition (E14) is then equivalent to checking the latter two values, namely

$$\frac{2 \exp(-3/2) + 1}{\alpha} - 1 \leq 1 \quad (\text{E16})$$

The smallest averaging constant is thus  $\alpha = \exp(-3/2) + 1/2$ .

### • Geman-McClure:

From the expression of Geman-McClure activation function (E13) reads

$$\sup_{t \in \mathbb{R}} \left| 1 + \alpha^{-1} \left( \frac{4\delta^4(2\delta^2 - 3t^2)}{(2\delta^2 + t^2)^3} - 1 \right) \right| = 1. \quad (\text{E17})$$

The zeros of the derivative of  $\vartheta'$  are

$$t = 0, \quad t = \sqrt{2}\delta, \quad \text{and} \quad t = -\sqrt{2}\delta$$

Thus, the maximum in (E17) is reached for the latter two values, which yield the minimum value of the averaging constant  $\alpha = 5/8$ .

- **Tukey biweight:**

By using the piecewise form of the activation function, plugging in the expression of  $\varrho$  on  $[-\sqrt{6}\delta, \sqrt{6}\delta]$ , and deriving  $\vartheta$ , we obtain the equivalent formulation of (E13):

$$\sup_{t \in [-\sqrt{6}\delta, \sqrt{6}\delta]} \left| 1 + \alpha^{-1} \left( \left( 1 - \frac{t^2}{6\delta^2} \right) \left( 1 - \frac{5t^2}{6\delta^2} \right) - 1 \right) \right| = 1 \quad (\text{E18})$$

The maximum is reached for  $|t| = 3\sqrt{\frac{2}{5}}\delta$  and the averaging constant of Tukey biweight activation function is equal to  $\alpha = 9/10$ .

- **Cauchy:**

Condition (E13) reads

$$\sup_{t \in \mathbb{R}} \left| 1 + \alpha^{-1} \left( \frac{\delta^2(\delta^2 - t^2)}{(\delta^2 + t^2)^2} - 1 \right) \right| = 1,$$

the maximum being reached when  $|t| = \sqrt{3}\delta$ . This results in  $\alpha = \frac{9}{16}$ .

- **Hyperbolic Tangent:**

Condition (E13) is equivalent to

$$\sup_{t \in \mathbb{R}} \left| 1 + \alpha^{-1} \left( \frac{\delta^2 \cosh\left(\frac{t^2}{2\delta^2}\right) - 2t^2 \sinh\left(\frac{t^2}{2\delta^2}\right)}{\delta^2 \cosh^3\left(\frac{t^2}{2\delta^2}\right)} - 1 \right) \right| = 1. \quad (\text{E19})$$

The zeros of the derivative of  $v'$  satisfy

$$2t \left( 3t^2 \sinh^2\left(\frac{t^2}{2\delta^2}\right) - 3\delta^2 \cosh\left(\frac{t^2}{2\delta^2}\right) \sinh\left(\frac{t^2}{2\delta^2}\right) - t^2 \cosh^2\left(\frac{t^2}{2\delta^2}\right) \right) = 0.$$

The maximum in (E13) is obtained for the nonzero solutions to this equation. To determine the latter values, a numerical resolution is required. By using Newton-Raphson's method, we obtain numerically that the solutions are  $t/\delta \simeq \pm 1.5355$ . We deduce that  $\alpha \simeq 0.9581$ .

### S3. Associated convex functions

In this appendix, we derive the expression of the convex function  $\varphi$  which was introduced in Proposition 1.

- **Fair potential:**

By using (20) and (E10), we have

$$(\forall t \in \mathbb{R}) \quad \varrho(t) = (\tilde{\varphi}^*)'(t) = \frac{\delta t}{|t| + \delta}. \quad (\text{E20})$$

This function can be seen as a scaled version of the Elliot activation function investigated in [R1, Example 2.15]. We have thus

$$(\forall t \in \mathbb{R}) \quad \tilde{\varphi}^*(t) = \delta(|t| - \delta \log(|t| + \delta)) + \eta \quad (\text{E21})$$

with  $\eta \in \mathbb{R}$ . Since  $\tilde{\varphi} \in \Gamma_0(\mathbb{R})$ ,

$$(\forall t \in \mathbb{R}) \quad \tilde{\varphi}(t) = \tilde{\varphi}^{**}(t) = \sup_{u \in \mathbb{R}} tu - \tilde{\varphi}^*(u). \quad (\text{E22})$$

For every  $t \in \mathbb{R}$ , let us define

$$\begin{aligned} (\forall u \in \mathbb{R}) \quad g(u) &= tu - \tilde{\varphi}^*(u) \\ &= tu - \delta(|u| - \delta \log(|u| + \delta)) - \eta. \end{aligned} \quad (\text{E23})$$

If  $t \geq \delta$ , then  $g(u) \xrightarrow{u \rightarrow +\infty} +\infty$ , and if  $t \leq \delta$  then  $g(u) \xrightarrow{u \rightarrow -\infty} +\infty$ . For  $t \in ]-\delta, \delta[$ ,  $g$  reaches its maximum at  $u$  such that

$$t = \frac{\delta u}{|u| + \delta} \quad \Leftrightarrow \quad u = \frac{\delta t}{\delta - |t|}. \quad (\text{E24})$$

For this value of  $u$ , by setting  $\eta = 2\delta^2 \log \delta$ ,

$$g(u) = -\delta|t| - \delta^2 \log(\delta - |t|). \quad (\text{E25})$$

Therefore, we deduce from (18) that, for every  $t \in \mathbb{R}$ ,

$$\varphi(t) = \begin{cases} -\delta|t| - \delta^2 \log(\delta - |t|) - \frac{t^2}{2} & \text{if } |t| < \delta \\ +\infty & \text{otherwise} \end{cases} \quad (\text{E26})$$

- **Huber:**

$$\begin{aligned} (\forall t \in \mathbb{R}) \quad \varrho(t) &= \begin{cases} t & \text{if } |t| < \delta \\ \delta \operatorname{sign}(t) & \text{otherwise} \end{cases} \\ &= \operatorname{proj}_{[-\delta, \delta]}(t) \end{aligned} \quad (\text{E27})$$

where  $\operatorname{proj}_S$  denotes the projection onto a nonempty closed convex set  $S$ . Thus

$$(\forall t \in \mathbb{R}) \quad \varphi(t) = \iota_{[-\delta, \delta]}(t) = \begin{cases} 0 & \text{if } t \in [-\delta, \delta] \\ +\infty & \text{otherwise.} \end{cases} \quad (\text{E28})$$

- **Green:**

$$(\forall t \in \mathbb{R}) \quad \varrho(t) = \tanh(t). \quad (\text{E29})$$

This activation function was already studied in [1, Example 2.12] where it was proved that, for every  $t \in \mathbb{R}$ ,

$$\varphi(t) = \begin{cases} \frac{(1+t) \log(1+t) + (1-t) \log(1-t) - t^2}{2} & \text{if } |t| < 1 \\ \log 2 - 1/2 & \text{if } |t| = 1 \\ +\infty & \text{otherwise.} \end{cases} \quad (\text{E30})$$

- **Hyperbolic:**

We will just consider the cases when  $\kappa \in \{1, 2\}$  for which closed form expressions can be obtained.

- If  $\kappa = 1$ , then

$$(\forall t \in \mathbb{R}) \quad \varrho(t) = \frac{t}{\sqrt{1 + \frac{t^2}{\delta^2}}}. \quad (\text{E31})$$

We can thus set

$$(\forall t \in \mathbb{R}) \quad \tilde{\varphi}^*(t) = \delta^2 \sqrt{1 + \frac{t^2}{\delta^2}}. \quad (\text{E32})$$

By defining  $g$  as in (E23), if  $t \geq \delta$ , then  $g(u) \xrightarrow{u \rightarrow +\infty} +\infty$  and, if  $t \leq \delta$ , then  $g(u) \xrightarrow{u \rightarrow -\infty} +\infty$ . For  $t \in ]-\delta, \delta[$ , the maximum of  $g$  is reached at  $u$  such that

$$t = \frac{u}{\sqrt{1 + \frac{u^2}{\delta^2}}} \Leftrightarrow u = \frac{t}{\sqrt{1 - \frac{t^2}{\delta^2}}}. \quad (\text{E33})$$

We deduce that, for every  $t \in \mathbb{R}$ ,

$$\varphi(t) = \begin{cases} -\delta^2 \sqrt{1 - \frac{t^2}{\delta^2}} - \frac{t^2}{2} & \text{if } |t| < \delta \\ +\infty & \text{otherwise.} \end{cases} \quad (\text{E34})$$

- If  $\kappa = 2$ , then  $\varrho$  reduces to the identity function and  $\varphi = 0$ .

#### S4. Training settings for Section 4

We summarize in the Table T1 the training settings used for the unrolled methods experimented in the paper.

#### S5. Additional figures for Section 4

We display in Figure F1 the training loss curves on Dataset 2 and 3, for the proposed architecture **U-HQ**. Examples of restored signals in Datasets 2 and 3, are displayed in Figure F2. We provide in Figure F3 a visual representation of learnt parameters by the proposed architecture **U-HQ**, for Datasets 2 and 3.

Penalty	Dataset 1			Dataset 2			Dataset 3		
	$K$	$p$	$lr$	$K$	$p$	$lr$	$K$	$p$	$lr$
<b>U-ISTA</b>									
Convex	12	$2K$	$10^{-3}$	16	$2K$	$10^{-3}$	16	$2K$	$10^{-3}$
<b>U-PD</b>									
Convex	16	$3K$	$10^{-3}$	12	$3K$	$10^{-4}$	16	$3K$	$10^{-4}$
<b>U-HQ-DE</b>									
Hybrid	8	2	$10^{-3}$	16	2	$10^{-2}$	12	2	$10^{-2}$
<b>U-HQ-FixS</b>									
Convex	8	$K$	$10^{-3}$	16	$K$	$10^{-2}$	12	$K$	$10^{-2}$
Non-convex	8	$K$	$10^{-3}$	16	$K$	$10^{-2}$	12	$K$	$10^{-2}$
Hybrid	8	$2K$	$10^{-3}$	16	$2K$	$10^{-2}$	12	$2K$	$10^{-2}$
<b>U-HQ-FixN</b>									
Convex	8	$2K$	$10^{-3}$	16	$2K$	$10^{-2}$	12	$2K$	$10^{-2}$
Non-convex	8	$2K$	$10^{-3}$	16	$2K$	$10^{-2}$	12	$2K$	$10^{-2}$
Hybrid	8	$3K$	$10^{-2}$	16	$3K$	$10^{-2}$	12	$3K$	$10^{-2}$
<b>U-HQ-FixN-OverP</b>									
Convex	8	$2KN$	$10^{-2}$	16	$2KN$	$10^{-2}$	12	$2KN$	$10^{-2}$
Non-convex	8	$2KN$	$10^{-2}$	16	$2KN$	$10^{-2}$	12	$2KN$	$10^{-2}$
Hybrid	8	$3KN$	$10^{-2}$	16	$3KN$	$10^{-2}$	12	$3KN$	$10^{-2}$
<b>U-HQ</b>									
Convex	8	$K(m+N)$	$10^{-6}$	8	$K(m+N)$	$10^{-5}$	12	$K(m+N)$	$10^{-6}$
			$10^{-2}$						$10^{-2}$
Non-convex	8	$K(m+N)$	$10^{-6}$	8	$K(m+N)$	$10^{-6}$	12	$K(m+N)$	$10^{-6}$
			$10^{-2}$						$10^{-2}$
Hybrid	8	$2K(m+N)$	$10^{-6}$	8	$2K(m+N)$	$10^{-6}$	12	$2K(m+N)$	$10^{-6}$
			$10^{-2}$						$10^{-2}$

Table T1: Summary of training settings for experimented deep unrolled architectures: number of layers  $K$ , number of learnt parameters  $p$ , optimizer and learning rate (lr). For **U-HQ**, we indicate  $lr_\lambda$  (top) and  $lr_\gamma$  (bottom)

## References

- [R1]. P.L. Combettes and J.-C. Pesquet. Deep neural network structures solving variational inequalities. Set-Valued and Variational Analysis, 28:491-518, 2020.

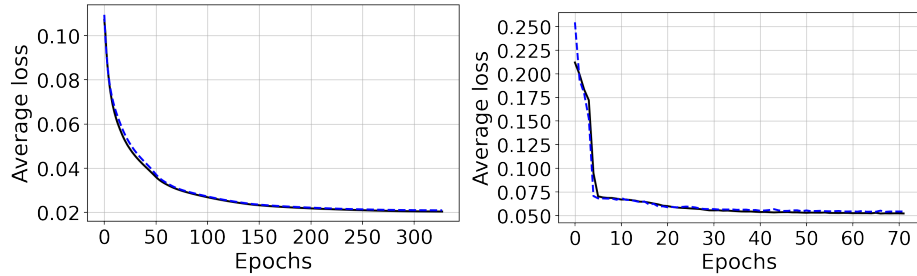


Figure F1: Evolution of training loss (solid black) and validation loss (dashed blue) along epochs for **U-HQ** for Dataset 2 (left) and Dataset 3 (right).

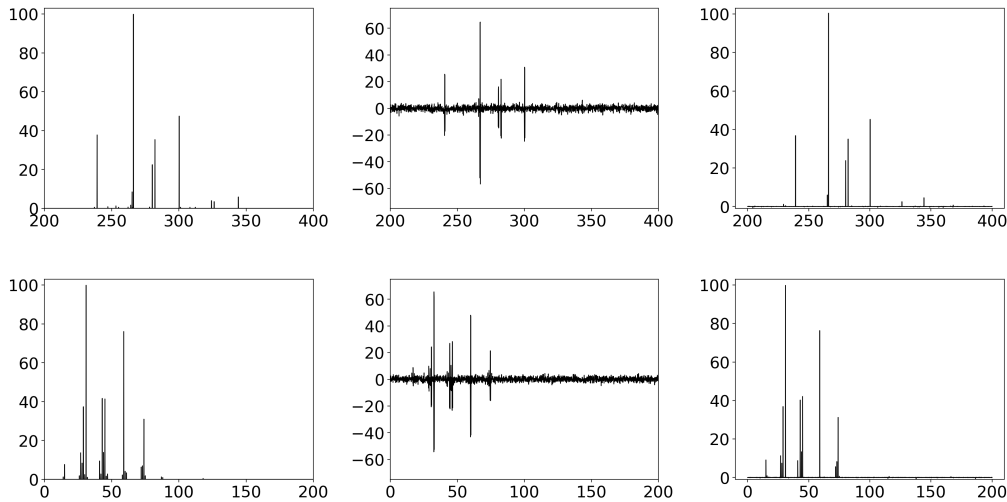


Figure F2: Groundtruth signals  $\bar{x}$  (top), degraded observations  $y$  (middle) and reconstruction  $\hat{x}$  using proposed method **U-HQ**, for examples from Datasets 2 (top) and 3 (bottom).  $x$ -axis indicates the mass-to-charge (m/z) ratio of chemical compounds and  $y$ -axis the abundance of corresponding compound. SNR/TSNR scores of restored signals (in dB) are (left) 22.09/24.16, (middle) 25.23/25.81, and (right) 24.09/24.52, respectively.

- [R2]. P.L. Combettes and J.-C. Pesquet. Lipschitz Certificates for Layered Network Structures Driven by Averaged Activation Operators. *SIAM Journal on Mathematics of Data Science*, 2:529-557, Jun. 2020.
- [R3]. H. H. Bauschke, P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd Edition, Springer Publishing Company, Incorporated, 2017.



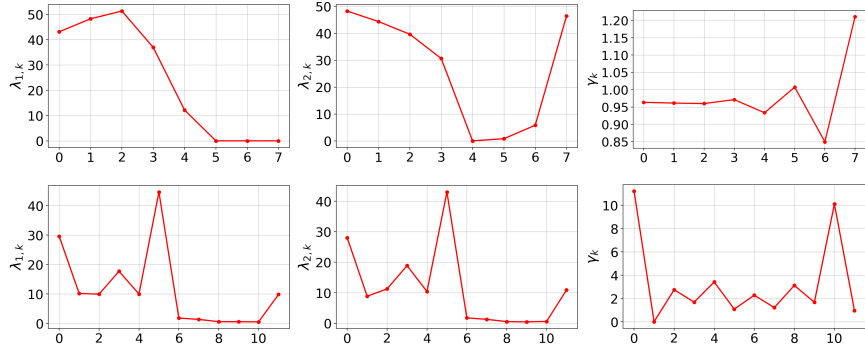


Figure F3: Learnt hyperparameter sequences  $(\lambda_{1,k})_{0 \leq k \leq K-1}$ ,  $(\lambda_{2,k})_{0 \leq k \leq K-1}$  and  $(\gamma_k)_{0 \leq k \leq K-1}$  with respect to layer index  $k$  ( $x$ -axis) for the proposed method **U-HQ**, for Dataset 2 (top) and Dataset 3 (bottom). Learnt parameters are averaged on test set.