Gaussian Mixtures Based IRLS for Sparse Recovery With Quadratic Convergence

Gaussian mixtures based IRLS for sparse recovery with quadratic convergence

Chiara Ravazzi, *Member, IEEE*, and Enrico Magli, *Senior Member, IEEE*

***Abstract*—In this paper we propose a new class of iteratively re-weighted least squares (IRLS) for sparse recovery problems. The proposed methods are inspired by constrained maximum likelihood estimation under a Gaussian scale mixture (GSM) distribution assumption. In the noise-free setting, we provide sufficient conditions ensuring the convergence of the sequences generated by these algorithms to the set of fixed points of the maps that rule their dynamics and derive conditions verifiable a posteriori for the convergence to a sparse solution. We further prove that these algorithms are quadratically fast in a neighborhood of a sparse solution. We show through numerical experiments that the proposed methods outperform classical IRLS for ℓ_τ-minimization with $\tau \in (0, 1]$ in terms of speed and of sparsity-undersampling tradeoff and are robust even in presence of noise. The simplicity and the theoretical guarantees provided in this paper make this class of algorithms an attractive solution for sparse recovery problems.**

***Index Terms*—Compressed sensing, constrained maximum likelihood estimation, Gaussian scale mixtures, iterative support detection and estimation, iteratively re-weighted least squares methods.**

I. Introduction

Compressed sensing is an efficient technique for nonadaptive sampling and reconstruction of sparse signals, allowing us to recover the signal starting from fewer linear measurements than classical sampling theory demands [1].

The literature describes a large number of algorithms to recover a sparse signal from an under-determined linear system. The main approaches can be classified as optimization-based methods [2], pursuit strategies [3], [4], [5], [6], coding-theoretic tools [7], [8], and Bayesian methods (see [9] and reference therein). In particular, the optimization-based methods solve a convex or non-convex program whose minimizer is known to approximate the target signal. Examples include quadratic programming [10], such as interior-point methods [10], projected gradient procedures [11], iterative (hard and soft) thresholding algorithms [12], [13], and iteratively reweighted least squares (IRLS, [14]).

In this work, we focus on IRLS reconstruction schemes, which have been proposed as a valuable strategy for ℓ_τ-minimization problems in sparse recovery with $\tau \in (0, 1]$ [15]. These algorithms have been deeply studied in absence of noise [14]. In particular, under certain conditions, these methods have been proved to converge to the true signal globally linearly fast when $\tau = 1$ and locally superlinearly fast with rate $2 - \tau$ for $\tau \in (0, 1)$. IRLS for ℓ_τ-minimization can be interpreted as an instance of the Expectation Maximization (EM) algorithm for constrained maximum likelihood estimation under the assumption that the data are normal/independent random variables [16]. This connection is useful to extend theoretical results regarding convergence and speed of convergence to noisy scenarios [16], [17].

Although IRLS algorithms appear very robust and superlinearly fast with rate close to 2 for τ approching 0, such guarantees of rate of convergence are valid only in a neighborhood of a sparse solution. More precisely, the algorithm seems to converge properly when τ is not too small (say $\tau > 1/2$) and tends to fail to reach the region of guaranteed convergence when $\tau < 1/2$ [2]. Heuristic methods to avoid local minima are still an open issue.

In this paper our goal is to design IRLS procedures for sparse recovery with *quadratic rate* of convergence. In particular, our iterative schemes are based on the constrained maximum likelihood estimation techniques under the assumption that the data are two-state mixtures of normal with Bernoulli prior. More precisely, we model elements in the support of the signal and zero values as zero mean Gaussian distributions with high and low variances, respectively. Then we choose the probability mass function of the state variable to be a Bernoulli with probability depending on the sparsity level. In the last years, several authors devoted their attention to sparse signal recovery using prior information on the support [18], [19]. In our case the considered model is only used as a proxy for sparse signals and the maximization of the log-likelihood is shown to provide a valuable alternative to find the sparsest vector consistent to the data.

The constrained maximization of the log-likelihood function for a mixture model turns out to be a hard problem and there is no closed form solution for the model parameters. In order to overcome this issue, we propose three iterative techniques, which we refer to GSM-IRLS: ML-IRLS, EM-IRLS, and K-EM-IRLS. A common point to the three strategies is to consider the complete log-likelihood function based on the missing data. After choosing some initial values for the mixture parameters, the following updates are alternated: in the first step, we use the current values for the parameters to estimate the signal and to evaluate the posterior distribution of the signal coefficients; in the second step we use these probabilities to re-estimate the mixture parameters. In this sense, GSM-IRLS can be also seen as sparse reconstruction methods

The authors are with the Politecnico di Torino – Department of Electronics and Telecommunications, Italy. e-mail: {name.surname}@polito.it. This work has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 279848.

that combine iterative support detection (ISD) and estimation [3], [4], [5], [6], [20], [21]. We remark that our reconstruction schemes differ from that literature. For example, with respect to [3], [4], [5] the identification of the support is not nested or increasing over the time. While [6], [20] use greedy rules to identify the support, the proposed methods are based on Bayesian rules and a probabilistic model. Moreover, once the support has been identified, all elements in the support are used in the estimation task, unlike to [3], [4], [5], [6], [20].

Besides the design of the algorithms, our main contribution includes a rigorous proof of convergence of the algorithms. More precisely, under certain conditions, these reconstruction schemes converge to a fixed point of the map that rules their dynamics (see Theorem 3). Moreover, if there is a sparse solution to the inverse problem, then the limit of the proposed algorithm is that sparse solution (Theorem 3) and the algorithm is quadratically fast in a neighborhood of that solution (Theorem 4). Numerical simulations validate our predictions and show that GSM-IRLS are quadratically fast and avoid local minima, outperforming classical IRLS for sparse recovery in terms of sparsity-undersampling tradeoff. Moreover, they reach performance comparable to algorithms based on iteratively reweighed ℓ_1-minimization but with a lower computational effort. Finally, we show that GSM-IRLS converge even in presence of noise and are robust against noise, outperforming also in this case classical IRLS for sparse recovery in terms of speed of convergence and accuracy.

A. Outline of the paper

The paper is organized as follows. The general linear inverse problem and the classical IRLS algorithms for ℓ_τ-minimization are described in Section II. In Section III the basic ideas and the philosophy of the proposed algorithms are presented and the relations to the state of the art is discussed. Section IV introduces the GSM model for sparse signals and recast the sparse optimization into a constrained maximum-likelihood estimation problem. Then the GSM-IRLS are developed and described in Section V. Section VI summarizes the theoretical contribution. Numerical experiments are presented in Section VII and some concluding remarks (Section VIII) complete the paper. The theoretical results are rigorously proved in Appendix through intermediate steps. We conclude this section with some notations and preliminary definitions.

B. Notation and preliminary definitions

We denote column vectors with small letters, and matrices with capital letters. If $x \in \mathbb{R}^n$ we denote its j-th element as x_j and, given $S \in [n] := \{1, \dots, n\}$, we denote as $x|_S$ the subvector of x corresponding to the indexes in S. The support set of x is defined by $\mathrm{supp}(x) = \{i \in [n] : x^i \neq 0\}$. Later we use the ℓ_τ-norms

$$\|x\|_{\ell_\tau} = \big(\sum_{i\in[n]} |x_i|^\tau \big)^{1/\tau} \quad \tau \in (0, \infty),$$

the ℓ_0 pseudonorm $\|x\|_{\ell_0} = |\mathrm{supp}(x)|$. The Euclidean norm of a vector x is often denoted as $\|x\|$ instead of $\|x\|_{\ell_2}$. Let $w_1, \dots, w_n$ be a set of positive numbers. The corresponding weighted inner product and weighted norm on $\mathbb{R}^n$ are defined as

$$\langle x, y \rangle_w = \sum_{i\in[n]} w_i x_i y_i,$$
$$\|x\|_w = \sqrt{\langle x, x \rangle_w} = \sqrt{\sum_{i\in[n]} w_i x_i^2},$$

respectively. We denote as $r(x)$ the non increasing rearrangement of x $r(x) = (|x_{i_1}|, |x_{i_2}|, \dots, |x_{i_n}|)^\top$, where $|x_{i_\ell}| \geq |x_{i_{\ell+1}}|$, $\forall \ell = 1, \dots, n-1$. Given a matrix A, A^T denotes its transpose.

We give some definitions regarding the speed at which a convergent sequence approaches its limit. Formally, let $(\theta^{(t)})_{t\in\mathbb{N}}$ be a sequence which converges to a limit point θ_∞.

Definition 1 (Linear convergence)**.** *We say that $(\theta^{(t)})_{t\in\mathbb{N}}$ converges linearly, if there exists $\eta \in (0,1)$ such that*

$$\frac{\|\theta^{(t+1)} - \theta_\infty\|}{\|\theta^{(t)} - \theta_\infty\|} < \eta. \tag{1}$$

The following definition describes a stronger form of convergence than linear convergence.

Definition 2 (Super-linear convergence)**.** *We say that $(\theta^{(t)})_{t\in\mathbb{N}}$ converges super-linearly, if the condition* (1) *holds for all $\eta \in (0,1)$ or, equivalently,*

$$\limsup_{t\to\infty} \frac{\|\theta^{(t+1)} - \theta_\infty\|}{\|\theta^{(t)} - \theta_\infty\|} = 0 \tag{2}$$

The next definition is used to distinguish super-linear rates of convergence. We say that the sequence converges with order q to θ_∞ for $q > 1$ if

$$\limsup_{t\to\infty} \frac{\|\theta^{(t+1)} - \theta_\infty\|}{\|\theta^{(t)} - \theta_\infty\|^q} = \mu \tag{3}$$

with finite $\mu > 0$. In particular the convergence with order $q = 2$ is called *quadratic convergence*.

II. The sparse recovery problem

A. Sparse signal reconstruction

Compressed sensing [1] is a technique used to recover sparse signals starting from few noisy measurements. More precisely, a signal $x^\star \in \mathbb{R}^n$ has to be recovered starting from $m \leq n$ measurements of the form

$$y = Ax^\star + \eta, \tag{4}$$

where $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$ is the sensing matrix with $m < n$, and $\eta \in \mathbb{R}^m$ is some unknown perturbation bounded by $\|\eta\| \leq \delta$. A common approach in many applications is to estimate $x^\star \in \mathbb{R}^n$ under the assumption that $x^\star$ is sparse (*i.e.*, the number of nonzero components is much smaller than n). It is thus natural to consider the following optimization problem in order to estimate $x^\star$:

$$(P_{0,\delta}): \qquad \min \|x\|_{\ell_0} \quad \text{subject to } x \in \mathcal{F}(y, \delta), \tag{5}$$

with $\mathcal{F}(y, \delta) = \{x \in \mathbb{R}^n : \|Ax - y\| \leq \delta\}$.

In absence of noise (*i.e.* $\delta = 0$), the system in (4) is linear and underdetermined. Therefore, the problem (5) with $\delta = 0$

has infinitely many solutions. We denote with $\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ the null space of the matrix A and the set of solutions of (4) can be represented by the affine space $\mathcal{F}(y) := \mathcal{F}(y, 0) = x^\star + \mathcal{N}(A)$. It can be shown [22] that, if $x^\star$ is k-sparse and the following assumption is satisfied, then $x^\star$ is the unique solution to $(P_{0,0})$.

Assumption 1. *For every index set $\Gamma \subseteq \{1, \dots, n\}$ with $|\Gamma| = 2k$ the columns of A associated with Γ are linearly independent.*

Under certain assumptions on the sensing matrix and for a sufficient low level of the signal sparsity [23], *robust signal recovery* is possible in presence of noise. This means that the solution $\widehat{x}_{0,\delta}$ of $(P_{0,\delta})$ obeys $\|\widehat{x}_{0,\delta} - x^\star\| \leq \kappa\delta$ where κ is a positive constant.

B. Sparse recovery via IRLS for ℓ_τ-minimization

The problem $(P_{0,\delta})$ is known to be NP-hard and, despite its attractiveness, solving $(P_{0,\delta})$ is not a viable way to estimate $x^\star$. However, an attractive alternative is given by solving a slightly different problem that consists in selecting the element in $\mathcal{F}(y, \delta)$, which has minimal ℓ_τ-norm with $\tau \in (0, 1]$:

$$(P_{\tau,\delta}) \quad \min_{x \in \mathbb{R}^n} \|x\|_{\ell_\tau} \qquad \text{s.t. } x \in \mathcal{F}(y, \delta) \tag{6}$$

From now on, unless otherwise specified, we consider the noise-free case ($\delta = 0$). Problem (6) has strong performance guarantees, which are easily described by the following definition.

Definition 3. *The matrix A is said to satisfy the τ-null space property (τ-NSP) of order K if there exists $\gamma \in (0, 1)$ such that*

$$\|\eta_S\|_{\ell_\tau}^\tau \leq \gamma \|\eta_{S^c}\|_{\ell_\tau}^\tau$$

for all sets S with $|S| \leq K$ and all $\eta \in \mathcal{N}(A)$.

The null space property characterizes the exact recovery of sparse signals from noise-free measurements via ℓ_τ-minimization [24]: if A satisfies the τ-NSP of order k and there exists a k-sparse vector in $\mathcal{F}(y)$, then this k-sparse vector is the unique solution of $(P_{\tau,0})$.

The optimization in $(P_{\tau,0})$ can be carried out by an IRLS algorithm. In particular, given an initial guess $x^{(0)}$, this algorithm generates a sequence of estimates for the signal $x^\star$ as follows:

$$x^{(t+1)} = \arg\min_{x \in \mathcal{F}(y)} \|x\|_{w^{(t)}(\tau)}$$

with

$$w_i^{(t)} = ((\epsilon^{(t)})^2 + (x_i^{(t)})^2)^{\tau/2 - 1} \tag{7}$$

for $i \in [n]$ and a suitable non-increasing sequence $\epsilon^{(t)}$. At each iteration the IRLS algorithm corresponds to a constrained weighted least-squares problem and can be efficiently solved using standard convex optimization methods.

The IRLS for ℓ_τ minimization can be also interpreted as the constrained maximum likelihood estimation under a GSM distribution [16], *i.e.*, assuming that $x^\star$ is a random variable with i.i.d. entries distributed according to

$$f_{x_i}(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\kappa(x_i^2)\right)$$

with

$$k(z) = \kappa(0) + \int_0^z \frac{1}{(\epsilon^2 + t)^{1-\tau/2}} \mathrm{d}t.$$

With this assumption the minimization of the rescaled negative log-likelihood takes the form

$$\min_{x \in \mathcal{F}(y)} \mathsf{L}_\tau^-(x) = \min_{x \in \mathcal{F}(y)} \sum_{i \in [n]} (x_i^2 + \epsilon^2)^{\tau/2} \tag{8}$$

which is the ϵ-smoothed version of ℓ_τ-norm of x with $\tau \in (0, 1]$. Using the formalism introduced in EM theory for normal/independent random variables [16], it can be shown that all IRLS algorithms with a weighting scheme with a completely monotone derivative [25] can be interpreted as an EM algorithm.

Although [16] shows a one-to-one correspondence between classical IRLS algorithms and a class of EM algorithms, this fact is not sufficient to guarantee the convergence of IRLS. In fact, as already observed in [16], the EM algorithm is not guaranteed to converge to a limit point but it is only guaranteed to converge to a set of points. If this set is discrete then the only connected components are singletons and the algorithm converges to some point. An ad-hoc proof for IRLS for ℓ_τ-minimization is provided in [14] in the noise-free setting, relying on the null space property of the constraint set. In [16] the proof is extended also for noisy-scenarios using the connection to the EM theory and the specific properties of the negative log-likelihood function in (8) (see [16], Appendix A).

The following theorem summarizes the convergence properties of IRLS for ℓ_τ-minimization in absence of noise. More precisely the following theorems have been proved in [14].

Theorem 1 (Theorem 7.7 in [14])**.** *Let K be chosen so that A satisfies the τ-NSP of order K with constant $\gamma < 1 - \frac{2}{K+2}$. For any $y \in \mathbb{R}^n$, let $\overline{\mathcal{Z}}_\tau(y)$ be the set of accumulation points of the sequence $(x^{(t)})_{t \in \mathbb{N}}$ produced by IRLS, and define $\overline{\epsilon} = \lim_{t \to \infty} \epsilon^{(t)}$. Then the following properties hold:*

- *(i) If $\overline{\epsilon} = 0$, then $\overline{\mathcal{Z}}_\tau(y)$ consists of a single point $\overline{x}$,* i.e., *$x^{(t)}$ converges to $\overline{x}$, and $\overline{x} \in \mathcal{F}(y)$ is an ℓ_τ-minimizer in $\mathcal{F}(y)$ that is also a K-sparse.*
- *(ii) If $\overline{\epsilon} > 0$, then for each $\overline{x} \in \overline{\mathcal{Z}}_\tau(y)$ we have $\langle \overline{x}, \eta \rangle_{w(\overline{x}, \epsilon, \tau)} = 0$ for all $\eta \in \mathcal{N}(y)$.*

The convergence is either globally linearly fast (if $\tau = 1$) or super-linearly fast (when $\tau \in (0, 1)$) in a neighborhood of the optimal solution.

Theorem 2 (Theorem 7.9 in [14])**.** *Assume that A satisfies the τ-NSP of order K with constant $\gamma \in (0, 1)$ and that $\mathcal{F}(y)$ contains a k-sparse vector $x^\star$ with $k \leq K$. Let $T = \operatorname{supp}(x^\star)$ and suppose that, for a given $\rho \in (0, 1)$, the local error $\|x^{(t_0)} - x^\star\| \leq R = \rho \min_{i \in T} |x_i^\star|^\tau$, then*

$$\|x^{(t+1)} - x^\star\| \leq \chi(\tau, K, \gamma, \tau, n) \|x^{(t)} - x^\star\|^{2-\tau} \tag{9}$$

for all $t \geq t_0$ if χ and R are such that $\chi R^{1-\tau} < 1$.

III. Support detection and signal reconstruction

In order to illustrate how the proposed algorithms work and evolve to solve (5), we will first describe their rationale and philosophy, which is very intuitive. Then, different procedures will be presented and their relations to the state of the art will be discussed.

A. Basic idea: an oracle estimator

Before presenting the proposed iterative algorithms to solve (5), we outline their basic principles. Let us start with some preliminary observations.

Proposition 1. *Suppose Assumption 1 holds and let $x^\star$ be the (unique) k-sparse solution of $(P_{0,0})$, $0 < \alpha \ll \beta = \frac{1}{k}\sum_{i=1}^n |x_i^\star|^2$. Then the (unique) solution $\widehat{x}^\alpha$ of the weighted least squares problem*

$$\min_{x\in\mathcal{F}(y)} \left[\sum_{j\in \mathrm{supp}(x^\star)} \frac{x_j^2}{\beta} + \sum_{j\notin \mathrm{supp}(x^\star)} \frac{x_j^2}{\alpha} \right] \tag{10}$$

is such that $\lim_{\alpha\to 0} \widehat{x}^\alpha = x^\star$.

Proof: Let us denote $\Lambda = \mathrm{supp}(x^\star)$ and $\eta = \widehat{x}^\alpha - x^\star \in \mathcal{N}(A)$. Since $\widehat{x}^\alpha$ is the optimal solution of (10) then for any $h \in \mathcal{N}(A)$, we have

$$\sum_{i\in\Lambda} \widehat{x}_i^\alpha h_i/\beta + \sum_{i\in\Lambda^c} \widehat{x}_i^\alpha h_i/\alpha = 0 \tag{11}$$

Using Cauchy-Schwarz inequality we obtain

$$\|\eta_\Lambda\|_1^2 \le \left(\sum_{i\in\Lambda} \frac{|\eta_i|^2}{\beta}\right) k\beta \le \left(\sum_{i\in\Lambda} \frac{|\eta_i|^2}{\beta} + \sum_{i\in\Lambda^c} \frac{|\eta_i|^2}{\alpha}\right) k\beta$$
$$= \left(\sum_{i\in\Lambda} \frac{(\widehat{x}_i^\alpha - x_i^\star)\eta_i}{\beta} + \sum_{i\in\Lambda^c} \frac{\widehat{x}_i^\alpha \eta_i}{\alpha}\right) k\beta \le \|x^\star\|_\infty \|\eta_\Lambda\|_1 k,$$

where the last equality follows from (11) with $h = \eta$. We evince that $\|\eta_\Lambda\|_1$ is bounded by $k\|x^\star\|_\infty$. Similarly, we have $\|\eta_{\Lambda^c}\|_1^2 \le \|x^\star\|_\infty \|\eta_\Lambda\|_1 (n-k)\alpha/\beta \le \|x^\star\|_\infty{}^2 k(n-k)\alpha/\beta$. We conclude that $\lim_{\alpha\to 0} \widehat{x}^\alpha = x^\star$ as, by Assumption 1, $x^\star$ is the unique k-sparse vector in $\mathcal{F}(y)$. ∎

This result guarantees that the optimization problem in (5) can be recast into a weighted ℓ_2-minimization. The advantage of this formulation consists in the constrained minimization of a smooth quadratic function which well approximates the nonsmooth ℓ_0 pseudonorm in (5) (see Fig. 1). In fact, by contuinuity of the weighted ℓ_2-minimization we get that the value of (10) is approximately equal to k when $\alpha \approx 0$. However, it is clear that we do not dispose neither of $x^\star$ nor $\mathrm{supp}(x^\star)$ and this approach is impractical.

B. Iterative support detection and estimation

Inspired by previous observation, we consider the solution of the following optimization problem

$$\min_{T\in[n],|T|\le n/2} \min_{x\in\mathcal{F}(y)} \left[\sum_{i\in T} \frac{x_i^2}{\beta} + \sum_{i\notin T} \frac{x_i^2}{\alpha} \right] \tag{12}$$

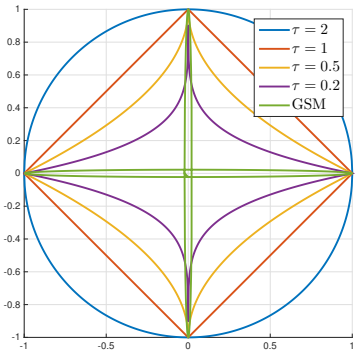


Fig. 1. The ℓ_τ-balls and the weighted ℓ_2-ball $x_1^2/\alpha + x_2^2/\beta = 1$ and $x_1^2/\beta + x_2^2/\alpha = 1$ with GSM mixture parameters $\alpha = 5\cdot 10^{-4}$ and $\beta = 1$.

with $0 \approx \alpha \ll \beta$. Fixed the set T, the idea is to penalize less the entries of x supported in T (see Fig. 1). Fixed the set T the solution is given by

$$x^{\alpha,\beta,T} = \Omega^{-1}A^\top (A\Omega^{-1}A^\top)^{-1} y,$$

where Ω is a diagonal matrix such that $\Omega_{\ell\ell} = 1/\beta$ if $\ell \in T$ and $\Omega_{\ell\ell} = 1/\alpha$ if $\ell \notin T$. Similar arguments used in Proposition 1 lead to conclude that, under Assumption 1, the solution $\widehat{x}^{\alpha,\beta}$ of (12) tends to the true signal $x^\star$ when $\alpha/\beta \to 0$.

Since the problem in (12) is still NP-hard, we are interested in designing low-complexity algorithms that incorporate both support detection and signal reconstruction. The methods we propose belong to a more general family of algorithms.

The first one performs *hard support detection* and signal reconstruction in an iterative fashion. The general lines of the procedure are described in Algorithm 1. The idea is that, starting from an initial guess T of the support of the signal and of the parameters $\alpha > 0$ and $\beta > 0$, the reconstruction is provided by the constrained weighted ℓ_2-least squares

$$\min_{x\in\mathcal{F}(y)} \left[\sum_{i\in T} \frac{x_i^2}{\beta} + \sum_{i\notin T} \frac{x_i^2}{\alpha} \right]$$

(see Step 3 in Algorithm 1). Then, given the estimation x of the signal, the support detection identifies a new atom set $T \subseteq [n]$ according to a certain rule. Step 4 of Algorithm 1 performs a hard assignment of indexes, as each component of the signal is associated uniquely to one cluster T or, alternatively, to $[n]\backslash T$. More precisely, the criterion upon which we decide whether $i \in T$ or in $i \in [n] \setminus T$ is described by a thresoshold test. Given a threshold $\delta = \delta(\alpha,\beta) > 0$ depending on parameters α and β, we assign the index i to T if $|x_i| > \delta$ and to $[n] \setminus T$ otherwise. Finally, parameters α and β are updated and the procedure is iterated until a specific stopping criterion is met. The choice of the threshold δ and the updates of α and β will be clear in Section V.A.

Another option is to alternatively perform a *soft support detection* and signal reconstruction. The second family of algorithms we consider is listed in Algorithm 2. It makes a soft assignment based on some prior information on the support,

Algorithm 1 Hard support detection and signal reconstruction

Input: Measurements $y \in \mathbb{R}^n$, data matrix $A \in \mathbb{R}^n$

1: Initialization: $\alpha^{(0)} = \alpha_0,\ \beta^{(0)} = \beta_0,\ T^{(0)} = T_0$

2: **for** $t = 0, 1, \ldots, StopIter$ **do**

3: Constrained weighted least square minimization:

$$x^{(t+1)} = \underset{x \in \mathcal{F}(y)}{\arg\min} \left[\sum_{i \in T} \frac{x_i^2}{\beta^{(t)}} + \sum_{i \notin T} \frac{x_i^2}{\alpha^{(t)}} \right]$$

4: Support detection: set threshold $\delta = \delta(\alpha^{(t)}, \beta^{(t)}) > 0$

$$T^{(t+1)} = \{ i \in [n] : |x_i^{(t+1)}| > \delta \}$$

5: Weights update:

$$\alpha^{(t+1)} = \alpha^{(t+1)}(x^{(t+1)}, T^{(t+1)})$$

$$\beta^{(t+1)} = \beta^{(t+1)}(x^{(t+1)}, T^{(t+1)})$$

6: **end for**

i.e. the assignment represents a belief that a certain component is in T.

Algorithm 2 Soft support detection and signal reconstruction

Input: Measurements $y \in \mathbb{R}^n$, data matrix $A \in \mathbb{R}^n$

1: Initialization: $\alpha^{(0)} = \alpha_0,\ \beta^{(0)} = \beta_0,\ \pi^{(0)} \in [0,1]^n$

2: **for** $t = 0, 1, \ldots, StopIter$ **do**

3: Constrained weighted least square minimization:

$$x^{(t+1)} = \underset{x \in \mathcal{F}(y)}{\arg\min} \left[\sum_{i \in T} \frac{(1 - \pi_i^{(t)}) x_i^2}{\beta^{(t)}} + \sum_{i \notin T} \frac{\pi_i^{(t)} x_i^2}{\alpha^{(t)}} \right]$$

4: Posterior beliefs of the signal coefficients:

$$\pi^{(t+1)} = \pi(x^{(t+1)}, \alpha^{(t)}, \beta^{(t)})$$

5: Weights update:

$$\alpha^{(t+1)} = \alpha^{(t+1)}(x^{(t+1)}, \pi^{(t+1)})$$

$$\beta^{(t+1)} = \beta^{(t+1)}(x^{(t+1)}, \pi^{(t+1)})$$

6: **end for**

Algorithm 1 and 2 do not specify a stopping criterion; standard criteria are to iterate until the estimate stops changing: $\|x^{(t+1)} - x^{(t)}\| / \|x^{(t)}\| < \mathsf{tol}$ for some $\mathsf{tol} > 0$.

C. Relation to prior literature

As in greedy algorithms such as Matching Pursuit, (MP, [3]) Orthogonal Matching Pursuit (OMP, [4]), Regularized Orthogonal Matching Pursuits (ROMP, [5]), Stagewise OMP (StOMP, [26]) and CoSaMP (see [6]), Algorithm 1 identifies at each iteration a set of indexes T and updates the estimation of the signal based on the knowledge of T. However, regarding the set selection, Algorithm 1 differs from StOMP, ROMP and OMP (see [27] for comparisons of these methods), where the index set T is nested or increasing over the iterations. When the index set T has been computed, Algorithm 1 updates all the components of x, including both the detected and undetected ones, at the same time. This feature marks its difference with the mentioned algorithms which update only components in T.

In [20] an iterative support detection Threshold-ISD method is proposed that runs as fast as the best Basis Pursuit algorithm (BP, [1]). We remark that Algorithm 1 differs from Threshold-ISD in all steps 3, 4, and 5. Given a set T, Threshold-ISD solves a truncated BP instead of a constrained weighted Least Squares

$$\min_{x \in \mathcal{F}(y)} \sum_{i \notin T} |x_i|.$$

It should be noticed that, in this case, for the estimation of the signal a truncated ℓ_1-norm is used. While Threshold-ISD, like CoSaMP, uses a greedy rule to identify index set T, based essentially on thresholding techniques, Algorithm 1, as we will see in the next section, uses a Bayesian rule for the identification. Finally, it should be noted the update in Step 5 makes Algorithm 1 more flexible than Threshold-ISD.

As already said, both Algorithm 1 and 2 belong to the more general class of IRLS methods [14], [21], [28]. As will be discussed in next section, compared to the classical IRLS algorithms, the weights used in the constrained weighted Least Squares (see Step 3 in Algorithm 1 and Step 3 in Algorithm 2) depend on the energy of the signal supported in T or in the complementary set of T and this marks its difference with IRLS where the weights associated to component i of the signal (7), chosen with the aim of approximating the ℓ_τ-norm of x, turn out to depend exclusively on the value x_i.

IV. GSM models for sparse signals

A. Bayesian formulation of the signal

Let us now consider a two-state GSM model [7], [16] as a prior that describes our prior knowledge about the sparsity of the signal. Because our approximately sparse signal consists of a small number of nonzero coefficients and many zero coefficients, we associate each probability density function with a state variable that can take on two values. Large and small magnitudes correspond to zero mean Gaussian distributions with high and low variances. To ensure that we have few nonzero coefficients, we choose the probability mass function of the state variable to be Bernoulli with probability depending on the sparsity rate. The resulting model for signal coefficients is a two-state GSM distribution. More formally, let $x^\star$ be a random variable with components of the form

$$x_i = z_i \sqrt{\alpha} u_i + (1 - z_i) \sqrt{\beta} u_i \quad i \in [n]$$

where u_i are identically and independently distributed zero-mean Gaussians and z_i are identically and independently distributed positive scalar random variables with probability mass function

$$f_{z_i}(z_i) = \begin{cases} 1 & 1-p \\ 0 & p \end{cases}$$

with $p = k/n$, $\alpha \approx 0$, and $\beta \gg 0$. In our model we thus have (see Fig. 2)

$$f_{x_i}(x_i) = \frac{1-p}{\sqrt{2\pi\alpha}} \mathrm{e}^{-\frac{x^2}{2\alpha}} + \frac{p}{\sqrt{2\pi\beta}} \mathrm{e}^{-\frac{x^2}{2\beta}}. \tag{13}$$

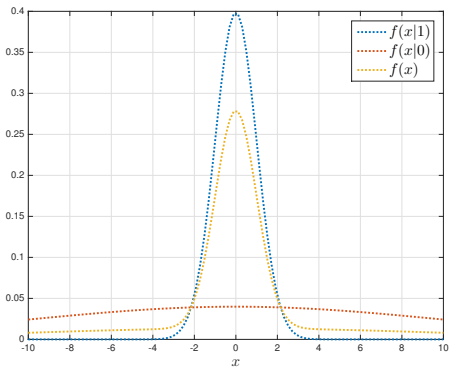


Fig. 2. Mixture Gaussian model for signal coefficients. The distribution of x conditioned on the two state variables, $z = 0$ and $z = 1$, is depicted with $1 = \alpha \ll \beta = 100$. the overall distribution is also shown for x with $p = k/n = 1/3$.

and

$$f(x,z) = \prod_{i=1}^{n} f(x_i|z_i) \prod_{i=1}^{n} f(z_i).$$

Since $\alpha \approx 0$ the family of GSM densities in (13) enhance sparsity in all coordinates (see Fig. 2).

In the last years, several authors devoted their attention to sparse signal recovery using prior information on the support [18], [19]. It is worth remarking that the GSM model is only used as a proxy for sparse signals.

This mixture model is completely characterized by three parameters: the sparsity ratio k/n, the variances α, β, and the configuration state z ($z_i = 1$ or $z_i = 0$).

With these options on the signal x, our goal is to estimate the parameters α, β and the specific configuration z_i of each components starting from linear measurements $y = Ax$. A natural approach to our problem would be to consider a joint maximum likelihood estimation (ML) of α, β and a maximum a-posteriori estimation of z_i (see [29]). Let $f(x, z|\alpha, \beta)$ be the joint distribution of x and z given the parameters α and β, and consider the negative rescaled log-likelihood function

$$\mathsf{L}^-(x, z, \alpha, \beta) := -\frac{1}{n} \log f(x, z|\alpha, \beta)$$

The hybrid ML/MAP solution, which for simplicity for now on we will refer to as the ML solution, prescribes to choose x, z, α, and β as

$$\begin{aligned}(x_{\mathrm{ML}}, z_{\mathrm{ML}}, \alpha_{\mathrm{ML}}, \beta_{\mathrm{ML}}) \\ = \underset{x\in\mathcal{F}(y), z\in\{0,1\}^n, \alpha\in\mathbb{R}^+, \beta\in\mathbb{R}^+}{\arg\min} \mathsf{L}^-(x, z, \alpha, \beta)\end{aligned} \tag{14}$$

Standard calculations lead to the following result

$$\begin{aligned}&\mathsf{L}^-(x, z, \alpha, \beta) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[z_i \left(\frac{x_i^2}{2\alpha} + \frac{1}{2} \log \alpha - \log(1-p) \right) \right. \\ &\left. + (1 - z_i) \left(\frac{x_i^2}{2\beta} + \frac{1}{2} \log \beta - \log p \right) \right] + c\end{aligned} \tag{15}$$

where c is a constant.

The minimization of (15) subject to $x \in \mathcal{F}(y)$, $z \in \{0,1\}^n$, $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+$ provides a valuable alternative to solve (5) even if the assumption on the distribution (13) of x is not satisfied. In particular, it should be noted that partial minimizations of $\mathsf{L}^-(x, z, \alpha, \beta)$ with respect to α and β have a simple form:

$$\widehat{\alpha}(x, z) = \arg\min_{\alpha} \mathsf{L}^-(x, z, \alpha, \beta) = \frac{1}{\sum_{i=1}^n z_i} \sum_{i=1}^{n} z_i x_i^2 \tag{16}$$

$$\begin{aligned}\widehat{\beta}(x, z) &= \arg\min_{\beta} \mathsf{L}^-(x, z, \alpha, \beta) \\ &= \frac{1}{\sum_{i=1}^n (1 - z_i)} \sum_{i=1}^{n} (1 - z_i) x_i^2.\end{aligned} \tag{17}$$

Putting (16) and (17) into (15) we obtain a non-convex penalty function which promotes the sparsity in the estimation:

$$\begin{aligned}&\mathsf{L}^-(x, z, \widehat{\alpha}(x, z), \widehat{\beta}(x, z)) \\ &= \frac{1}{n} \left[n + \sum_{i=1}^{n} z_i \left(\frac{1}{2} \log \frac{\sum_{i=1}^n z_i x_i^2}{\sum_{i=1}^n z_i} - \log(1-p) \right) \right. \\ &\left. + \sum_{i=1}^{n} (1 - z_i) \left(\frac{1}{2} \log \frac{\sum_{i=1}^n (1 - z_i) x_i^2}{\sum_{i=1}^n (1 - z_i)} - \log p \right) \right] + c\end{aligned}$$

The ML estimation can be performed by minimizing the following function

$$\begin{aligned}\mathsf{L}^-(T, x) = \left(1 - \frac{|T|}{n}\right) \log \frac{\|x_{T^c}\|^2}{(n - |T|)(1-p)^2} \\ + \frac{|T|}{n} \log \frac{\|x_T\|^2}{|T| p^2} + c\end{aligned} \tag{18}$$

subject to $T \in [n]$ and $x \in \mathcal{F}(y)$.

The obvious drawback of minimizing (15), and consequently (18), is that this problem is not well posed. In fact, if we consider the limit $\|x_T\|^2 \to 0$ or $\|x_{T^c}\|^2 \to 0$, then the cost function goes to $-\infty$. Such singularities will always be present and will occur whenever one of the Gaussian components collapses onto a specific data point. On the other hand, we can not expect that $\|x_T\|^2 \neq 0$ or $\|x_{T^c}\|^2 \neq 0$, since we desire a sparse solution. We can hope to avoid these pathological behaviors by minimizing the following modified cost function:

$$\begin{aligned}&\mathsf{J}(x, z, \alpha, \beta, \epsilon) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{z_i x_i^2 + \epsilon^2/n}{2\alpha} + \frac{z_i}{2} \log \alpha - z_i \log(1-p) \right. \\ &\left. + \frac{(1 - z_i) x_i^2 + \epsilon^2/n}{2\beta} + \frac{(1 - z_i)}{2} \log \beta - (1 - z_i) \log p \right] + c.\end{aligned} \tag{19}$$

V. GSM BASED IRLS

The considerations above suggest three methods to design IRLS algorithms for sparse recovery, which we refer to as ML-IRLS algorithm, EM-IRLS algorithm, and K-EM IRLS algorithm, respectively.

1) ML-IRLS is an alternating method for choosing minimizers and weights based on the functional $\mathsf{J}(x, z, \alpha, \beta, \epsilon)$ in

(19) and alternates two steps: hard support detection and estimation (see Algorithm 1);
2) EM-IRLS is a more refined iterative technique which is inspired by the so-called EM algorithm for parameters estimation and consists in an iterative soft support detection and signal estimation (see Algorithm 2);
3) K-EM IRLS is a modification of EM-IRLS and so belongs to the class of algorithms described in Algorithm 2. The idea behind such a modification is to modify the beliefs on the support in such a way that the $n-K$ entries of the signal that are expected to be small are penalized more.

A. ML-IRLS

We initialize the parameters $K = \widetilde{K}$, $p = K/n$, $\alpha^{(0)} = \alpha_0$, $z^{(0)} = \mathbb{1}$, $\epsilon^{(0)} = 1$. Then, for each time $t = 0, 1, \ldots, StopIter$, we compute

$$x^{(t+1)} = \underset{x \in \mathcal{F}(y)}{\arg\min}\, \mathsf{J}(x, z^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}) \quad (20)$$

$$z^{(t+1)} = \underset{z \in \{0,1\}}{\arg\min}\, \mathsf{J}(x^{(t+1)}, z, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}) \quad (21)$$

$$\epsilon^{(t+1)} = \min\left(\epsilon^{(t)}, \frac{r(x^{(t+1)})_{K+1}}{n}\right) \quad (22)$$

$$(\alpha^{(t+1)}, \beta^{(t+1)}) = \underset{\alpha,\beta}{\arg\min}\, \mathsf{J}(x^{(t+1)}, z^{(t+1)}, \alpha, \beta, \epsilon^{(t)}) \quad (23)$$

In particular, the computation of (20), which corresponds to Step 3 in Algorithm 1, requires the solution of a weighted least squares problem. If we define

$$\omega_i^{(t)} = z_i^{(t)}/\alpha^{(t)} + (1 - z_i^{(t)})/\beta^{(t)}$$

then

$$x^{(t+1)} = \Omega^{-1} A^\top (A\Omega^{-1} A^\top)^{-1} y$$

with Ω defined as the $n \times n$ diagonal matrix whose j-th diagonal entry is $\omega_j^{(t)}$. Once $x^{(t+1)}$ has been computed, (21) is given by

$$z_i^{(t+1)} = \begin{cases} 1 & \text{if } |x_i| < \delta(\alpha, \beta, p) \\ 0 & \text{otherwise} \end{cases}$$

and

$$\delta(\alpha, \beta, p) = \sqrt{2 \frac{\log\left(\frac{\beta}{\alpha}\frac{(1-p)^2}{p^2}\right)}{\frac{\beta-\alpha}{\alpha\beta}}}.$$

Finally estimation in (23) are performed according to the following rules

$$\alpha^{(t+1)} = \frac{\sum_{i=1}^n z_i |x_i^{(t+1)}|^2 + |\epsilon^{(t+1)}|^2}{\sum_{i=1}^n z_i}$$
$$\beta^{(t+1)} = \frac{\sum_{i=1}^n (1 - z_i) |x_i^{(t+1)}|^2 + |\epsilon^{(t+1)}|^2}{\sum_{i=1}^n (1 - z_i)}.$$

B. EM-IRLS

A more refined iterative solution is inspired by the so-called Expectation-Maximization (EM) algorithm [30], [31].

We initialize the parameters $K = \widetilde{K}$, $p = K/n$, $\alpha^{(0)} = \alpha_0$, $\pi^{(0)} = \mathbb{1}$, $\epsilon^{(0)} = 1$. Then, for each time $t = 0, 1, \ldots, StopIter$, we compute

$$x^{(t+1)} = \underset{x \in \mathcal{F}(y)}{\arg\min}\, \mathsf{J}(x, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}) \quad (24)$$

$$\pi_i^{(t+1)} = f(z_i^{(t+1)} = 1 | x^{(t+1)}, \alpha^{(t)}, \beta^t, \epsilon^{(t)}) \quad (25)$$

$$\epsilon^{(t+1)} = \min\left(\epsilon^{(t)}, \frac{r(x^{(t+1)})_{K+1}}{n}\right) \quad (26)$$

$$(\alpha^{(t+1)}, \beta^{(t+1)}) = \underset{\alpha,\beta}{\arg\min}\, \mathsf{J}(x^{(t+1)}, \pi^{(t+1)}, \alpha, \beta, \epsilon^{(t)}) \quad (27)$$

In particular, (24) requires the solution of a weighted least squares problem. If we define

$$\omega_i^{(t)} = \pi_i^{(t)}/\alpha^{(t)} + (1 - \pi_i^{(t)})/\beta^{(t)} \quad i \in [n]$$

then

$$x^{(t+1)} = \Omega^{-1} A^\top (A\Omega^{-1} A^\top)^{-1} y \quad (28)$$

where Ω is the $n \times n$ diagonal matrix whose j-th diagonal entry is $\omega_j^{(t)}$. Once $x^{(t+1)}$ has been computed, the estimation (25) is given by

$$\pi_i^{(t+1)} = \frac{\mathrm{e}^{-\frac{|x_i^{(t+1)}|^2}{2\alpha} - \frac{\log(\alpha)}{2} + \log(1-p)}}{\mathrm{e}^{-\frac{|x_i^{(t+1)}|^2}{2\alpha} - \frac{1}{2}\log(\alpha) + \log(1-p)} + \mathrm{e}^{-\frac{|x_i^{(t+1)}|^2}{2\beta} - \frac{\log(\beta)}{2} + \log p}}$$

and the update (27) is performed according to the following rules

$$\alpha^{(t+1)} = \frac{\sum_{i=1}^n \pi_i |x_i^{(t+1)}|^2 + |\epsilon^{(t+1)}|^2}{\sum_{i=1}^n \pi_i},$$
$$\beta^{(t+1)} = \frac{\sum_{i=1}^n (1 - \pi_i) |x_i^{(t+1)}|^2 + |\epsilon^{(t+1)}|^2}{\sum_{i=1}^n (1 - \pi_i)}.$$

C. K-EM-IRLS

Finally, we propose a modification of EM-IRLS. K-EM-IRLS differs from the previous algorithm because in the E-step a thresholding operator σ_{n-K} is applied in order to promote the sparsity in the probability vector π. More precisely, it acts on π by taking the $n - K$ biggest elements

$$\sigma_{n-K}(v) = \underset{z \in \mathbb{R}^n : \|z\|_0 \le n-K}{\arg\min} \|z - v\|.$$

We initialize the parameters $K = \widetilde{K}$, $p = K/n$, $\alpha^{(0)} = \alpha_0$, $\pi^{(0)} = \mathbb{1}$, $\epsilon^{(0)} = 1$. Then, for each time $t = 0, 1, \ldots, StopIter$, we compute

$$\begin{aligned} x^{(t+1)} &= \underset{x \in \mathcal{F}(y)}{\arg\min}\, \mathsf{J}(x, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)}) \\ \pi_i^{(t+1)} &= f(z_i^{(t+1)} = 1 | x^{(t+1)}, \alpha^{(t)}, \beta^t, \epsilon^{(t)}) \\ \pi^{(t+1)} &= \sigma_{n-K}(\pi^{(t+1)}) \\ \epsilon^{(t+1)} &= \min\left(\epsilon^{(t)}, \frac{r(x^{(t+1)})_{K+1}}{n}\right) \\ (\alpha^{(t+1)}, \beta^{(t+1)}) &= \underset{\alpha,\beta}{\arg\min}\, \mathsf{J}(x^{(t+1)}, \pi^{(t+1)}, \alpha, \beta, \epsilon^{(t)}). \end{aligned} \quad (29)$$

VI. Theoretical results

A. Convergence of the proposed methods

In order to state our results in formal way, it is convenient to rewrite the dynamics of the proposed algorithms. In particular, we express the iterations as follows. The updates of GSM-IRLS are denoted as follows: given $\theta^{(0)}$, we consider the dynamical system

$$\theta^{(t+1)} = \Psi(\theta^{(t)}) \tag{30}$$

where $\Psi \in \{\Psi_{\mathrm{ML}}, \Psi_{\mathrm{EM}}, \Psi_{K\text{-EM}}\}$,

1) Ψ_{ML} is the map composition described by (20)-(23) and $\theta^{(t)} = (x^{(t)}, z^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)})$;
2) Ψ_{EM} is the map composition described by (24)-(27) and $\theta^{(t)} = (x^{(t)}, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)})$;
3) $\Psi_{K\text{-EM}}$ is the map composition described by equations in (29) and $\theta^{(t)} = (x^{(t)}, \pi^{(t)}, \alpha^{(t)}, \beta^{(t)}, \epsilon^{(t)})$.

In other terms, the recursive formula in (30) joins in one step the operations that in the GSM-IRLS algorithms are split into multiple steps, but the dynamics are actually the same.

The algorithms have been designed in such a way that there exists a function V which is non-increasing and convergent along the sequence of iterates:

$$V(\theta^{(t)}) \geq V(\Psi(\theta^{(t)})) = V(\theta^{(t+1)}).$$

More precisely, we have

1) $V_{\mathrm{ML}}(\theta) = V_{\mathrm{ML}}(x, z, \alpha, \beta, \epsilon) := \mathsf{J}(x, z, \alpha, \beta, \epsilon)$ is the function defined in (19);
2)

$$\begin{aligned} V_{\mathrm{EM}}(\theta) &= V_{\mathrm{EM}}(x, \pi, \alpha, \beta, \epsilon) \\ &:= -\frac{1}{n}\sum_{i=1}^{n} H(\pi_i) + \mathsf{J}(x, \pi, \alpha, \beta, \epsilon) \end{aligned} \tag{31}$$

where $H : [0,1] \to \mathbb{R}$ is the natural entropy function

$$H(\zeta) = -\zeta \log \zeta - (1-\zeta)\log(1-\zeta).$$

The next theorem ensures that, under certain conditions, also the sequence $(\theta^{(t)})$ is converging to the set of fixed points of Ψ.

Theorem 3 (Convergence)**.** *Let us assume that for every index set $\Gamma \subseteq [n]$ with $|\Gamma| = K$ the columns of A associated with Γ are linearly independent and define $\bar{\epsilon} = \lim_{t\to\infty} \epsilon^{(t)}$.*

For any $y \in \mathbb{R}^m$, the sequence

$$(\theta^{(t)})_{t\in\mathbb{N}} = (x^{(t)}, \pi^{(t)}, \epsilon^{(t)}, \alpha^{(t)}, \beta^{(t)})_{t\in\mathbb{N}}$$

produced by GSM-IRLS converges to a fixed point of Ψ. Moreover, if $\bar{\epsilon} = 0$ then $x^{(t)}$ converges to a point $\overline{x}$ that is also K-sparse.

The proof of Theorem 3 for EM-IRLS, postponed to Appendix A, is obtained using arguments of variational analysis. More precisely, the following facts are proved: (a) the sequence $(\theta^{(t)})_{t\in\mathbb{N}}$ is bounded; (b) two successive iterations of these algorithms become closer and closer: $\lim_{t\to\infty} \|\theta^{(t+1)} - \theta^{(t)}\| = 0$. The assertion is then concluded by continuity of Ψ. The convergence of K-EM-IRLS uses similar arguments, while the proof of ML-IRLS can be obtained by first proving (a), (b) and the stabilization of discrete variables of $z_i^{(t)}$ in finite time by modeling the system as a switching dynamical system (see [32]).

In other terms, as is typically the case with constrained ℓ_τ minimization problem (6) with $\tau \in (0,1)$, Theorem 3 does not ensure that Algorithm 1, 2, and 3 converge to the sparsest solution of (4). However, it does provide sufficient conditions that are verifiable a posteriori (*e.g.*, $\bar{\epsilon} = 0$) for such convergence. The reason for this weaker result lies in the non-convexity of the cost function in (19). Nevertheless, in the following theorem we show a local convergence result, which also highlights that the proposed techniques are quadratically fast in a neighborhood of a sparse solution.

Theorem 4 (Quadratic rate)**.** *Assume that A satisfies the NSP of order K with constant $\gamma \in (0,1)$ and that there exists a k-sparse vector $x^\star \in \mathcal{F}(y)$ with $k \leq K$. Let $\Lambda = \mathrm{supp}(x^\star)$ and*

$$\mathbb{1}_{\Lambda^c} = \begin{cases} 1 & i \in \Lambda^c \\ 0 & \textit{otherwise} \end{cases}.$$

There exist $R_1, R_2, R_3, R_4 > 0$, such that if

$$\begin{gathered} \|x^{(t_0)} - x^\star\| \leq R_1, \quad \|\pi^{(t_0)} - \mathbb{1}_{\Lambda^c}\| \leq R_2, \\ |\alpha^{(t_0)}| \leq R_3, \quad \left|\beta^{(t_0)} - \frac{\|x^\star\|^2}{k}\right| \leq R_4, \end{gathered}$$

then for all $t \geq t_0$

$$\begin{gathered} \|x^{(t+1)} - x^\star\| \leq \chi_1|\alpha^{(t)}|, \quad \|\pi^{(t+1)} - \mathbb{1}_{\Lambda^c}\| \leq \chi_2|\alpha^{(t)}|^{1/2} \\ |\alpha^{(t+1)}| \leq \chi_3|\alpha^{(t)}|^2, \quad \left|\beta^{(t+1)} - \frac{\|x^\star\|^2}{k}\right| \leq \chi_4|\alpha^{(t)}|^{1/2} \end{gathered} \tag{32}$$

with $\chi_i := \chi_i(R_1, R_2, R_4, \gamma, K, n)$ for $i \in \{1, \ldots, 4\}$ and such that

$$R_3 \leq \min\left\{\frac{R_1}{\chi_1}, \left(\frac{R_2}{\chi_2}\right)^2, \frac{1}{\chi_3}, \left(\frac{R_4}{\chi_4}\right)^2, 1\right\}. \tag{33}$$

This result is proved in Appendix B for EM-IRLS but similar arguments can be used for ML and K-EM-IRLS. Theorem 2, as Fact 2 for ℓ_τ-minimization, states that there exists a region where the convergence to a sparse solution is guaranteed and is quadratically fast. We conclude that the proposed methods outperform classical IRLS in terms of rate of convergence.

It should be emphasized that Theorem 2 provides only a sufficient condition for this convergence. In principle the region of quadratic convergence could be very small if at least one among χ_1, χ_2, χ_3, and χ_4 is quite large. In the following section we exhibit different simulations that show that this is not the case: while classical IRLS with $\tau < 1/2$ get trapped in local minima, the proposed method converge quadratically fast to the desired solution.

VII. Numerical experiments

In this section we discuss a series of experiments in order to assess the performance of the proposed GSM-IRLS methods in terms of convergence time and accuracy. We also show that these algorithms yield exact reconstruction in the noiseless

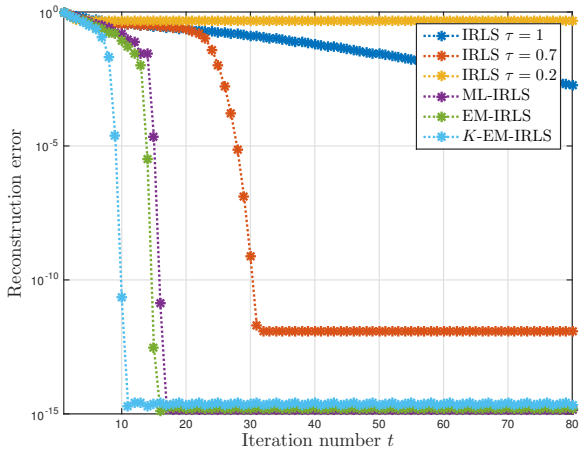


Fig. 3. A typical evolution of the approximation error $E(t) = \|x^{(t+1)} - x^\star\|/\|x^\star\|$ for classical IRLS algorithms (with $\tau = 1, 0.7, 0.2$) and IRLS based on ML and EM. The nonzero components of the signal $x^\star$ are drawn from a uniform distribution $\mathsf{U}([-10, 10])$.

scenario and are robust in presence of noise, in that small errors on the measurements produce small perturbation in the reconstruction.

A. Reconstruction from Noise-free Measurements

We begin with some experimental results describing the performance of GSM-IRLS methods in terms of convergence times and reconstruction capability in absence of noise. Our experiments include three different test sets: in the first and the second experiment we use synthetic signals and the last analysis considers a two-dimensional image, taken from the Sparco toolbox [33].

1) Rate of convergence (demo): A signal $x^\star$ to be recovered is generated choosing $k = 45$ nonzero components uniformly at random among the $n = 1500$ elements and drawing the amplitude of each nonzero component from a uniform distribution $\mathsf{U}([-10, 10])$ in order to introduce a mismatch in the signal model. The sensing matrix A with $m = 250$ rows is sampled from the Gaussian ensemble, n with zero mean and variance $1/m$. Such matrices are known to satisfy the NSP (see Definition 3) with high probability [34]. We have initialized the parameters $\alpha^{(0)} = 0.1$, $\pi^{(0)} = \mathbb{1}$ and $K = 55$.

In Fig. 3, we compare the convergence rate of classical IRLS for different choices of $\tau = 1, 0.7, 0.2$ and GSM-based methods. In particular, the approximation error of the iterates

$$E(t) = \|x^{(t+1)} - x^\star\|/\|x^\star\|$$

to the unique sparsest solution of (5) is depicted as a function of the iteration step. The case with $\tau = 1$ shows linear convergence and for the smaller values of it (*i.e.*, $\tau = 0.7$), the error decay initially follows a linear, transient regime; however, once the iterates get sufficiently close to the sparse solution vector, the convergence is seen to speed up dramatically resulting in super-linear convergence. For smaller values of τ (*i.e.*, $\tau = 0.2$), we often do not observe convergence to the sparsest solution of (4). In fact, if $\tau < 0.5$, then the algorithm tends to fail to reach the region of guaranteed convergence. The GSM-IRLS are faster than classical IRLS methods: the transient linear regime lasts less and the local region of super-linear convergence is larger than classical IRLS methods based on ℓ_τ-minimization with $\tau < 0.5$.

2) Empirical probability of reconstruction: We compare the performance of GSM-IRLS with classical IRLS methods, Basis Pursuit (BP, [35]), Iterative support detection (Threshold-ISD, [20]) and Orthogonal Matching Pursuit (OMP, [4]), in terms of the empirical recovery success rate, averaged over 50 experiments, as a function of the sparsity level and number of measurements. The recovery is considered successfully when the reconstruction error is below 10^{-4}.

In the first experiment the matrix $A \in \mathbb{R}^{160\times 512}$ is generated as above; then, for each sparsity level $k \in [10, 100]$, the signal $x^\star$ is generated choosing k nonzero components uniformly at random among the $n = 512$ elements and drawing the amplitude of each nonzero component from a uniform distribution $\mathsf{U}([-10, 10])$. IRLS's smoothing parameter and the Threshold-ISD parameters have been optimized as suggested in [20]. For GSM-IRLS we have set $\alpha^{(0)} = 1$, $\pi^{(0)} = \mathbb{1}$, $K = k+30$ and fixed a total number of iterations equal to 200. In Figure 4, it should be noticed that GSM-IRLS outperform the classical ones with $\tau = 1$ and $\tau = 0.7$ in terms of sparsity-under-sampling tradeoff and the BP. Since classical IRLS with $\tau = 1$ has the global convergence property, we expect it to reach the accuracy obtained by BP. However, the convergence is extremely slow and before 200 iterations we get only an approximation with an error of order 10^{-2} in most of the cases. In the same picture are also reported the curves obtained using OMP with sparsity guess equal to k. In the considered setting OMP enjoys similar performance to BP only with the perfect knowledge of the sparsity of the signal. We leave it to the reader to verify that, as soon as the sparsity level is not perfectly known, the relative error is of order 10^{-1}. We would like to remark that IRLS methods are not affected by the choice of the sparsity guess if $K < n/2$. Moreover EM-IRLS reaches performance comparable to Threshold-ISD.

In the second experiment, we test the performance using sparse Bernoulli signals (see also Test set 4 in [20]): $x^\star$ is a column vector of length $n = 600$ with $k = 40$ nonzero components equal to ± 1 with probability $1/2$, respectively. The signal is measured by a Gaussian matrix, taking m random measurements. The GSM-IRLS parameters have been set as follows: $\alpha^{(0)} = 1$, $\pi^{(0)} = \mathbb{1}$, $K = 55$ and a total number of iterations has been fixed to 200. Figure 5 depicts the empirical recovery success rate (when the relative reconstruction error of the estimate provided is below 10^{-4}) averaged over 50 experiments, as a function of the number of measurements $m \in [80, 220]$. In particular, Threshold-ISD, EM-IRLS and K-EM IRLS achieve the same recoverability, which is significantly higher than that obtained using BP. The recoverability of the OMP method is the worst. It can be checked that the empirical probability of reconstruction obtained with other greedy methods like StOMP starts to become positive as soon as $m > 340$. It should be noticed that for this signal model also ML-IRLS does not achieve a good recovery and get trapped into local minima as classical IRLS with $\tau < 1/2$.

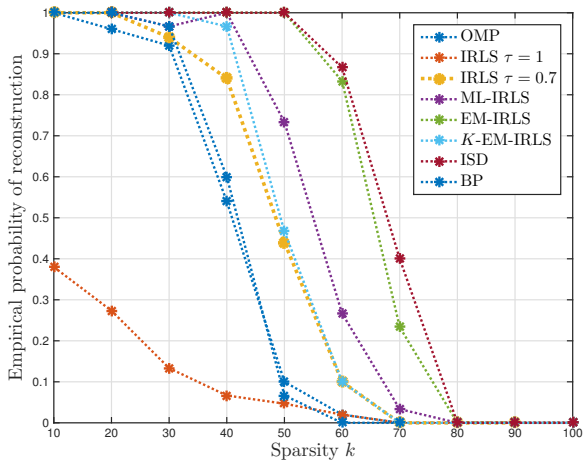


Fig. 4. Empirical probability of successful recovery as a function of the sparsity value k with $n = 512$ and $m = 160$.

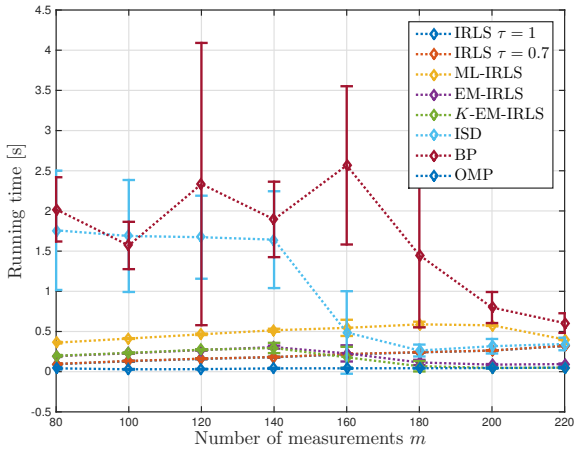


Fig. 6. Empirical probability of successful recovery as a function of the number of measurements $m \in [80, 220]$ with sparse Bernoulli signals $n = 600$ and $k = 40$.

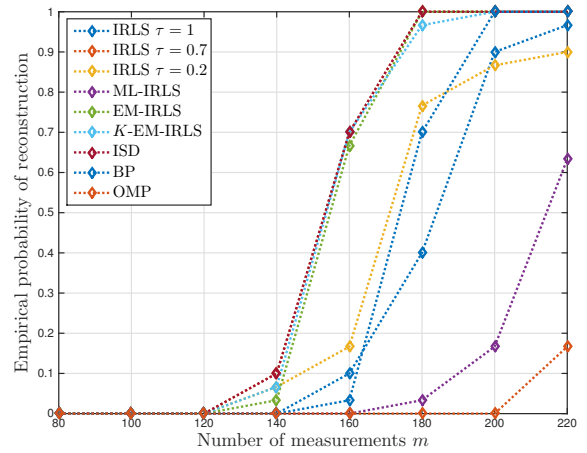


Fig. 5. Empirical probability of successful recovery as a function of the number of measurements $m \in [80, 220]$ with sparse Bernoulli signals $n = 600$ and $k = 40$.

Figure 6 shows the average running times (over 50 experiments) of the algorithms and the error bar represents the standard deviation of uncertainty. This picture testifies that IRLS methods are much faster than both Threshold-ISD and BP. We conclude that EM-IRLS and K-EM IRLS are not only the fastest algorithms but also required minimum number of measurements.

3) Sparco 501: In Sparco 501 and in the discussion that follows, a two-dimensional image is vectorized before the application of any linear operator. In particular, the vectorized form of an image $X^\star \in \mathbb{R}^{64\times 64}$ is trasformed into $x^\star$ defined as

$$x^\star = \text{vec}(X^\star) = (X_1^{\star\top}\ X_2^{\star\top}\ \dots\ X_n^{\star\top})^\top$$

We assume that the image has relatively sparse representations in Haar wavelets; *i.e.*, there exists $z^\star \in \mathbb{R}^{4096}$ such that

$$x^\star = Wz^\star$$

where $W \in \mathbb{R}^{4096\times 4096}$ is the Haar wavelet basis, and $z^\star$ is approximately sparse. We use a gaussian matrix $A \in \mathbb{R}^{2048\times 4096}$ to compute

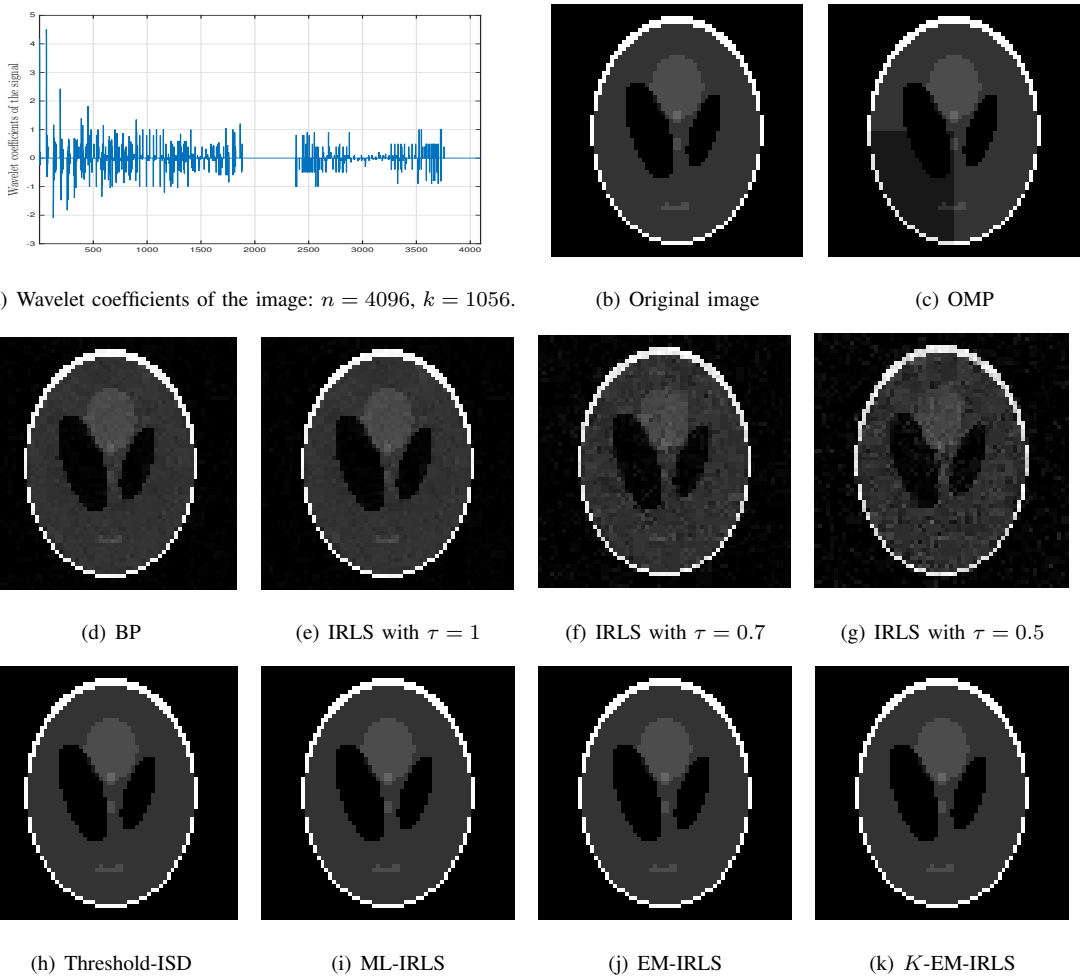
$$y = Ax^\star = AWz^\star$$

We try to obtain the wavelet coefficients $\widehat{z}$ by using IRLS methods. Finally, we complete the recovery by computing $\widehat{x} = W\widehat{z}$.

Fig. 7 depicts the relative error between the true Haar coefficients of the image and the recovered Haar coefficients as a function of iterations: after 30 iterations the GSM-IRLS provide an approximation with a relative error of order 10^{-14} and the classical methods are able to achieve just an error of order 10^{-2} for $\tau = 1$ and of order 10^{-1} for $\tau = 0.5, 0.7$. Fig. 8 and Table I compare the recovered images and the relative errors obtained by greedy algorithms such as OMP [4], Basis Pursuit (BP, [35]), classical IRLS methods for ℓ_τ-minimization with different values of τ, sparse reconstruction schemes via Threshold-ISD [20], and GSM-IRLS techniques. It can be shown that greedy algorithms are not able to obtain good recovery. Classical IRLS with $\tau = 1$ will reach the accuracy of BP but the decay of error is only linear and after 30 iterations the relative reconstruction error is of order $2.7\cdot 10^{-2}$. Classical IRLS with $\tau = 0.7$ and $\tau = 0.5$ are trapped into a local minima and the resulting reconstruction error is larger that the one obtained with $\tau = 1$. Only the algorithms based on the combination of iterative support detection and estimation (Threshold-ISD and GSM-IRLS) are able to reach very high accuracy in a reasonable number of iterations. In particular, Threshold-ISD has success after solving one ℓ_1-minimization and six Truncated-ℓ_1-minimization problems (about CPU Time = 132 [s]), while GSM-IRLS reach the accuracy after about 20 iterations (about CPU Time = 56 [s]).

B. Reconstruction from Noisy Measurements

1) Rate of convergence (demo): The IRLS algorithms can be extended for the minimization problem $(P_{\tau,\delta})$ in (6). At



(a) Wavelet coefficients of the image: $n = 4096$, $k = 1056$. (b) Original image (c) OMP

(d) BP (e) IRLS with $\tau = 1$ (f) IRLS with $\tau = 0.7$ (g) IRLS with $\tau = 0.5$

(h) Threshold-ISD (i) ML-IRLS (j) EM-IRLS (k) K-EM-IRLS

Fig. 8. Sparco 501: Original and recovered images for Shepp-Logan phantom. The image has been recovered using $m = 0.5n$ linear measurements. After 20 iterations the GSM-IRLS provide an approximation with a relative MSE of order 10^{-14} and the classical methods are able to achieve just an error of order 10^{-2} for $\tau = 1$ and of order 10^{-1} for $\tau = 0.5, 0.7$.

Algorithm	Iter	MSE
OMP	55	$3.8 \cdot 10^{-1}$
BP	–	$2.7 \cdot 10^{-2}$
IRLS $\tau = 1$	20	$6.9 \cdot 10^{-2}$
IRLS $\tau = 0.7$	20	$1.4 \cdot 10^{-1}$
IRLS $\tau = 0.5$	20	$2.1 \cdot 10^{-1}$
Threshold-ISD	7	$4.1 \cdot 10^{-14}$
ML-IRLS	21	$3.1 \cdot 10^{-14}$
EM-IRLS	18	$1.1 \cdot 10^{-14}$
K-EM-IRLS	16	$1.8 \cdot 10^{-14}$

TABLE I
RELATIVE RECONSTRUCTION ERROR OF RECOVERED IMAGES FOR SHEPP-LOGAN PHANTOM (SEE ALSO FIGURE 8). ONLY THE ALGORITHMS BASED ON THE COMBINATION OF ITERATIVE SUPPORT DETECTION AND ESTIMATION (THRESHOLD-ISD AND GSM-IRLS) ARE ABLE TO REACH VERY HIGH ACCURACY IN A REASONABLE NUMBER OF ITERATIONS.

each iteration the IRLS algorithm corresponds to a constrained weighted least-squares problem ($x \in \mathcal{F}(y, \delta)$) and can be efficiently solved using standard convex optimization tools. As observed in [16], if the output SNR is greater than 1, then, at each iteration, 0 is not a feasible solution and the solutions lie on the boundary of $\mathcal{F}(y, \delta)$. Classical IRLS appears very robust and stable with a linear (if $\tau = 1$) or super-linear (when $\tau \in (0, 1)$) convergence in a neighborhood of the global minimizer of $(P_{\tau,\delta})$ (see [16] for a rigorous proof).

As a first experiment, we have considered a noisy scenario: the signal $x^\star$ to be recovered is generated choosing $k = 45$ nonzero components uniformly at random among the $n = 1500$ elements and drawing the amplitude of each nonzero component from a uniform distribution $\mathsf{U}([-10, 10])$. The sensing matrix A with $m = 250$ rows is sampled from the Gaussian ensemble, n with zero mean and variance $1/m$. The the vector η is an additive white Gaussian noise with standard deviation $\sigma = 0.01$. We assume the standard deviation of the noise is known in advance and, as $\mathbb{E}[\|\eta\|^2] = m\sigma^2$, we set $\delta = \sqrt{m}\sigma$. We have initialized the parameters $\alpha^0 = 0.1$, $\pi^{(0)} = \mathbb{1}$ and $K = 55$. Figure 9 shows that all the tested algorithms converge to a fixed point of the algorithm in few iterations. For smaller $\tau = 0.5$, we often do not observe convergence to the desired solution. In fact, since the ℓ_τ-norm is not a convex function, the algorithm easily gets trapped in local minima. It is worthwhile noting that the estimations obtained by the proposed GSM-IRLS are significantly more accurate in terms of the mean square error compared to those

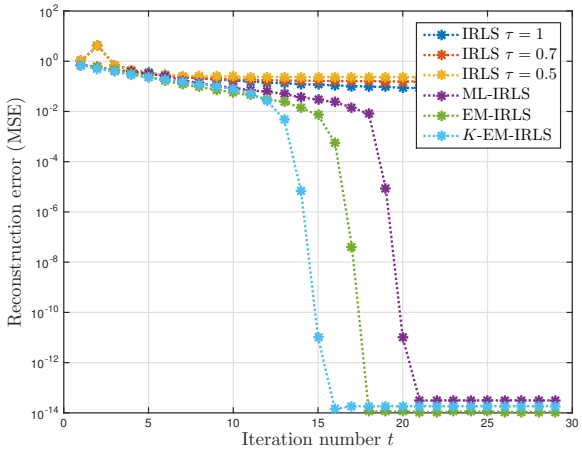


Fig. 7. Sparco 501: Evolution of the approximation error $E(t) = \|x^{(t+1)} - x^\star\|/\|x^\star\|$ for classical IRLS algorithms (with $\tau = 1, 0.7, 0.2$) and GSM-IRLS.

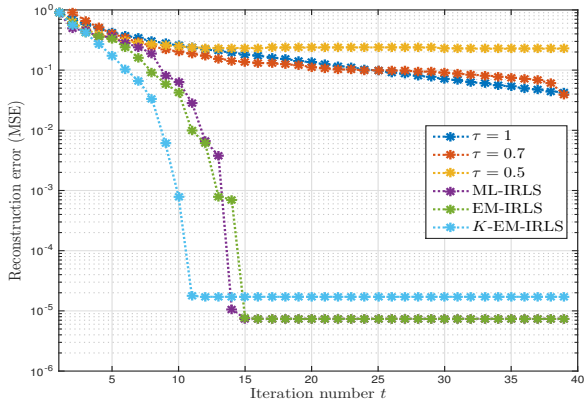


Fig. 9. Noisy scenario: A typical evolution of MSE as a function of the iterations for classical IRLS algorithms (with $\tau = 1, 0.7, 0.2$) and GSM-IRLS. The nonzero components of the signal $x^\star$ are drawn from a uniform distribution $\mathsf{U}([-10, 10])$ and the vector η is an additive white noise with standard deviation $\sigma = 0.01$.

obtained with the classical IRLS methods.

2) Robustness: In Figure 10 we use the same test set of the previous section and show that the proposed methods are robust against noise. More precisely, the mean square error, averaged over 50 runs and obtained after 50 iterations, is depicted as a function of Signal-to-Noise ratio (SNR). It should be noted that only few iterations are required to reach a satisfactory degree of accuracy. In all curves we can clearly identify the log-linear dependence of the MSE as a function of the SNR and, consequently, of the parameter δ. Moreover the MSE of the proposed algorithms are smaller than those obtained via classical IRLS algorithms with $\tau = 1, 0.7$ and $\tau = 0.5$. As already observed, the MSE of classical IRLS with $\tau = 0.5$ is very high compared to the other methods. Moreover, it does not decrease as the SNR increases and the algorithm turns out to be not robust against noise. This is due to the fact that the algorithm gets trapped into local minima and is unable to reach the global minima of $(P_{\tau,\delta})$.

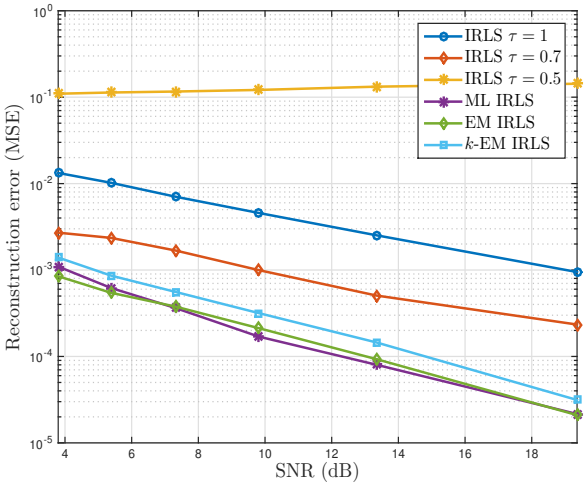


Fig. 10. Mean square error after 40 iterations as a function of the SNR for classical IRLS (with $\tau = 1$, $\tau = 0.7$ and $\tau = 0.5$) and the proposed GSM-IRLS.

VIII. Concluding remarks

In this paper, we designed three new techniques for sparse recovery problems. These iterative procedures, obtained by combining constrained likelihood maximization under GSMs assumptions and IRLS methods, are very powerful. The main theoretical contribution includes the proof of convergence of the algorithm to a fixed point and the conditions for local quadratic convergence in noise-free scenarios. Numerical results confirm our predictions and show that the proposed algorithms outperform classical IRLS schemes and related algorithms based on iteratively ℓ_1 minimization in terms of convergence rate and sparsity-undersampling tradeoff even in presence of noise.

Appendix

In this section we prove rigorously Theorem 3 and 4, which state the convergence of the algorithms to a fixed point of the map Ψ. We provide the proof for EM-based IRLS and similar arguments can be used also for ML and K-EM based IRLS.

A. EM-based IRLS: Convergence

In this section we prove the convergence of the EM-based IRLS (Theorem 3). We begin with the following preliminary results.

Lemma 1. *Let us assume that for every index set $\Gamma \subseteq [n]$ with $|\Gamma| = K$ the columns of A associated with Γ are linearly independent. Then, the set of feasible points $x \in \mathcal{F}(y)$ such that $\|x\|_0 \leq K$ is finite.*

Proof: From hypothesis, it should be deduced that the set of feasible points $x \in \mathcal{F}(y)$ such that $\|x\|_0 \leq K$ is finite. In fact, if we consider for any set $S \subseteq [n]$ so that $|S| \leq K$, submatrix A_S of A comprised of the columns corresponding to the set S, then by the assumption it follows that the matrix $A_S^\top A_S$ is nonsingular and $x = (A_S^\top A_S)^{-1} A_S^\top y$. To

summarize, for each set of indices S satisfying $|S| \leq K$, there is at most one candidate in $\mathcal{F}(y)$ with support S. Since the number of subsets of $[n]$ is finite, the result follows. ∎

Proposition 2. *The function V defined in* (31) *is nonincreasing along the iterates $(\theta^{(t)})_{t\in\mathbb{N}}$.*

Proof: Let $\widehat{\pi} = \widehat{\pi}(x,\alpha,\beta,\epsilon) = \arg\min_{\xi\in\mathbb{R}^n} V(x,\xi,\alpha,\beta,\epsilon)$ then, by differentiating V with respect to π_i we obtain

$$\widehat{\pi}_i = \frac{\mathrm{e}^{-\frac{x_i^2}{2\alpha}-\frac{\log(\alpha)}{2}+\log(1-p)}}{\mathrm{e}^{-\frac{x_i^2}{2\alpha}-\frac{1}{2}\log(\alpha)+\log(1-p)} + \mathrm{e}^{-\frac{x_i^2}{2\beta}-\frac{\log(\beta)}{2}+\log p}} \tag{34}$$

and $\partial^2 V/\partial\pi_i^2 \geq 0$.
It should be now noticed that, for each time $t \in \mathbb{N}$, we have

$$\begin{aligned} V(\theta^{(t+1)}) &= V(x^{(t+1)},\pi^{(t+1)},\alpha^{(t+1)},\beta^{(t+1)},\epsilon^{(t+1)}) \\ &\leq V(x^{(t+1)},\pi^{(t+1)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\leq V(x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\leq V(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\leq \dots \\ &\leq V(x^{(1)},\pi^{(1)},\alpha^{(1)},\beta^{(1)},\epsilon^{(1)}) = V(\theta^{(1)}). \end{aligned}$$

∎

Lemma 2. *Let us define the sequence of weights*

$$w_i^{(t)} = \frac{\pi_i^{(t)}}{2\alpha^{(t)}} + \frac{1-\pi_i^{(t)}}{2\beta^{(t)}} \tag{35}$$

and assume that for every index set $\Gamma \subseteq [n]$ with $|\Gamma| = K$ the columns of A associated with Γ are linearly independent. Then, the sequence $(x_i^{(t)})_{t\in\mathbb{N}}$ is bounded and $(w_i^{(t)})_{t\in\mathbb{N}}$ is lower bounded for all $i \in [n]$ and $t \in \mathbb{N}$.

Proof: From Proposition 2 we obtain

$$\begin{aligned} n &+ \sum_{i=1}^n \frac{\pi_i^{(t)}}{2}\log\alpha^{(t)} + \sum_{i=1}^n \frac{1-\pi_i^{(t)}}{2}\log\beta^{(t)} - \sum_{i=1}^n H(\pi_i) \\ &- \sum_{i=1}^n \pi_i^{(t)}\log(1-p) - \sum_{i=1}^n (1-\pi_i^{(t)})\log p \\ &= V(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\leq V(x^{(1)},\pi^{(1)},\alpha^{(1)},\beta^{(1)},\epsilon^{(1)}). \end{aligned}$$

Then, there exists a constant $C \in \mathbb{R}$ such that

$$\sum_{i=1}^n \frac{1-\pi_i^{(t)}}{2}\log\beta^{(t)} + \sum_{i=1}^n \frac{\pi_i^{(t)}}{2}\log\alpha^{(t)} \leq C. \tag{36}$$

We prove now that both $\alpha^{(t)}$ and $\beta^{(t)}$ must be upper bounded. Suppose ad absurdum that $\beta^{(t)} \to +\infty$, then by (36) $\alpha^{(t)} \to 0$ and

$$\lim_{t\to\infty} \epsilon^{(t)} \geq \lim_{t\to\infty} r(x^{(t)})_\ell = 0$$

for all $\ell \geq K+1$. This implies that there exists a subsequence $(x^{t_j})_{j\in\mathbb{N}}$ and a set S with $|S| \leq K$ such that

$$\lim_{j\to\infty} x_{S^c}^{(t_j)} = 0.$$

From Lemma 1 and the fact that $x^{(t_j)} \in \mathcal{F}(y)$ then there exists a K-sparse accumulation point $\widetilde{x}$ such that $\lim_{j\to\infty} x^{(t_j)} = \widetilde{x}$. This fact implies that the sequence $(x^{(t_j)})_{j\in\mathbb{N}}$ is bounded and, consequently $(\beta^{(t_j)})_{j\in\mathbb{N}}$ is bounded, as it is a convex combination of finite points. We conclude that

$$\lim_{j\to\infty} \pi_i^{(t_j)} = \begin{cases} 1 & \text{if } i \notin S \\ 0 & \text{if } i \in S \end{cases} \qquad \lim_{j\to\infty} \beta^{(t_j)} = \frac{\|\widetilde{x}\|^2}{K}.$$

From (28) and by continuity we get $\lim_{t\to\infty} x_i^{(t)} = \widetilde{x}$ and $\lim_{t\to\infty}\beta^{(t)} = \frac{\|\widetilde{x}\|^2}{K}$, hence the contradiction. The same arguments can be used if we suppose $\lim_{t\to\infty}\alpha^{(t)} = +\infty$. Since both sequences $(\alpha^{(t)})_{t\in\mathbb{N}}$ and $(\beta^{(t)})_{t\in\mathbb{N}}$ are upper bounded by a constant $C \in \mathbb{R}$, then also $(x^{(t)})_{t\in\mathbb{N}}$ is bounded and $(w_i^{(t)})_{t\in\mathbb{N}}$ is lower bounded for all $i \in [n]$:

$$\begin{aligned} \|x^{(t)}\|^2 &= \alpha^{(t)}\sum_{i=1}^n \pi_i^{(t)} + \beta^{(t)}\sum_{i=1}^n (1-\pi_i^{(t)}) \leq Cn \\ \|w_i^{(t)}\|^2 &= \frac{\pi_i^{(t)}}{\alpha^{(t)}} + \frac{1-\pi_i^{(t)}}{\beta^{(t)}} \geq \frac{1}{C}. \end{aligned}$$

∎

The following lemma ensures that two successive iterations of these algorithms become closer and closer. This property implies that these algorithms converge numerically when the number of iterations goes to infinity.

Lemma 3.

$$\lim_{t\to\infty} \|x^{(t+1)} - x^{(t)}\| = 0$$

Proof: Let us consider the weights $w_i^{(t)}$ defined in (35) with $i \in \mathbb{N}$. For each time $t \in \mathbb{N}$, we have

$$\begin{aligned} &V(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\quad - V(x^{(t+1)},\pi^{(t+1)},\alpha^{(t+1)},\beta^{(t+1)},\epsilon^{(t+1)}) \\ &\quad \geq V(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\quad - V(x^{(t+1)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\quad = \frac{1}{2}\langle x^{(t)},x^{(t)}\rangle_{w^{(t)}} - \frac{1}{2}\langle x^{(t+1)},x^{(t+1)}\rangle_{w^{(t)}} \\ &\quad = \frac{1}{2}\langle x^{(t)}+x^{(t+1)},x^{(t)}-x^{(t+1)}\rangle_{w^{(t)}} \end{aligned}$$

From (24) we have for any $\eta \in \mathcal{N}(A)$

$$\sum_i^n w_i^{(t)} x_i^{(t+1)} \eta_i = 0$$

and observing that $x^{(t)} - x^{(t+1)} \in \mathcal{N}(A)$ we obtain

$$\begin{aligned} &V(x^{(t)},\pi^{(t)},\alpha^{(t)},\beta^{(t)},\epsilon^{(t)}) \\ &\quad - V(x^{(t+1)},\pi^{(t+1)},\alpha^{(t+1)},\beta^{(t+1)},\epsilon^{(t+1)}) \\ &\quad \geq \frac{1}{2}\langle x^{(t)}-x^{(t+1)},x^{(t)}-x^{(t+1)}\rangle_{w^{(t)}} \\ &\quad = \frac{1}{2}w_j^{(t)}(x_j^{(t)}-x_j^{(t+1)})^2 \\ &\quad \geq \frac{w}{2}\sum_{i=1}^n (x_j^{(t)}-x_j^{(t+1)})^2 \end{aligned}$$

with $\underline{w} > 0$ (see Lemma 2). Summing both sides over times $t \in \mathbb{N}$ we get there exists a constant $C > 0$ such that

$$\sum_{t\in\mathbb{N}} \|x^{(t)} - x^{(t+1)}\|^2 \leq C$$

from which we evince the statement of the theorem. ∎

Proof of Theorem 3: (a) From Lemma 2 the sequence $(x^{(t)})$ is bounded and by the Bolzano Weierstrass Theorem, there exists a subsequence $(x^{p_j})_{j\in\mathbb{N}}$ such that

$$\lim_{t\to\infty} x^{(p_j)} = \widetilde{x}.$$

From (28) and by continuity we get $\lim_{t\to\infty} x_i^{(p_j+1)} = \lim_{t\to\infty} x_i^{(p_j)} = \widetilde{x}$. and using the continuity of Ψ we conclude $\Psi(\widetilde{x}) = \widetilde{x}$.

(b) Let us suppose that there exists a time t_0 such that $\epsilon^{(t_0)} = 0$, then the algorithm is stopped at $t = t_0$ and $\epsilon^{(t_0)} = r(x^{(t_0)})_{K+1} = 0$, *i.e.* $x^{(t_0)}$ is K-sparse. By assumption and by the fact that $x^{(t_0)} \in \mathcal{F}(y)$ we obtain that $x^{(t_0)} = \overline{x}$ which is the only K-sparse vector in $\mathcal{F}(y)$.

Let us assume now that $\epsilon^{(t)} > 0$ for all $t \in \mathbb{N}$ and $\overline{\epsilon} = 0$. Then there is an increasing sequence $(t_\ell)_{\ell\in\mathbb{N}}$ and a set $S \in [n]$ with $|S| \leq K$ such that $\lim_{t\to\infty} x_{S^c}^{(t_\ell)} = 0$. From Lemma 2 the sequence $(x^{(t)})$ is bounded and, consequently, there exists a subsequence $(p_j)_{j\in\mathbb{N}}$ of $(t_\ell)_{\ell\in\mathbb{N}}$ such that

$$\lim_{t\to\infty} x^{(p_j)} = \widetilde{x}$$

with $\widetilde{x}_{S^c} = 0$. By assumption we conclude again $\widetilde{x} = \overline{x}$ which is the only K-sparse solution in $\mathcal{F}(y)$. We conclude that

$$\lim_{j\to\infty} \pi_i^{(t_j)} = \begin{cases} 1 & \text{if } i \notin S \\ 0 & \text{if } i \in S \end{cases} \qquad \lim_{j\to\infty} \beta^{(t_j)} = \frac{\|\widetilde{x}\|^2}{K}.$$

From (28) and by continuity we get $\lim_{t\to\infty} x_i^{(t)} = \widetilde{x} = \overline{x}$.

B. EM-based IRLS: Quadratic rate of convergence

We prove now Theorem 4 through intermiediate steps. Let $\Lambda = \mathrm{supp}(x^\star)$ and $\eta^{(t)} = \|x^{(t)} - x^\star\|$.

Lemma 4. *The following equality holds*

$$\sum_{i\in[n]} \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} |\eta_i^{(t+1)}|^2 + \frac{1-\pi_i^{(t)}}{\beta^{(t)}} |\eta_i^{(t+1)}|^2 \right) = -\sum_{i\in\Lambda} \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} \eta_i^{(t+1)} x_i^\star + \frac{1-\pi_i^{(t)}}{\beta^{(t)}} \eta_i^{(t+1)} x_i^\star \right).$$

Proof: We start with the observation that

$$\begin{aligned}
&\sum_{i\in[n]} \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} |\eta_i^{(t+1)}|^2 + \frac{1-\pi_i^{(t)}}{\beta^{(t)}} |\eta_i^{(t+1)}|^2 \right) \\
&= \sum_{i\in[n]} \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} + \frac{(1-\pi_i^{(t)})}{\beta^{(t)}} \right) (x_i^{(t+1)} - x_i^\star) \eta_i^{(t+1)} \\
&= \sum_{i\in[n]} \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} + \frac{(1-\pi_i^{(t)})}{\beta^{(t)}} \right) x_i^{(t+1)} \eta_i^{(t+1)} \\
&- \sum_{i\in[n]} \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} + \frac{(1-\pi_i^{(t)})}{\beta^{(t)}} \right) x_i^\star \eta_i^{(t+1)} \\
&= -\sum_{i\in\Lambda} \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} + \frac{(1-\pi_i^{(t)})}{\beta^{(t)}} \right) x_i^\star \eta_i^{(t+1)}.
\end{aligned}$$

where the last step follows from the optimality condition in (24) and from $x^\star|_{\Lambda^c} = 0$. ∎

Proof of Theorem 4: We prove the statement by induction on $t \in \mathbb{N}$. Let $\Lambda = \mathrm{supp}(x^\star)$ and, for a given $\rho \in (0,1)$, define

$$R_1 = \rho \min_{i\in\Lambda} |x_i^\star|.$$

We assume (inductive assumption) that for some $R_2 \in (0,1)$, $R_3 \in (0,1)$, and $R_4 \in (0, \|x^\star\|^2/4)$

$$\|x^{(t)} - x^\star\| \leq R_1, \quad \|\pi^{(t)} - \mathbb{1}_{\Lambda^c}\| \leq R_2$$
$$|\alpha^{(t)}| \leq R_3, \quad \left|\beta^{(t)} - \frac{\|x^\star\|^2}{k}\right| \leq R_4$$

and we will derive (32).

If $\eta^{(t)} := x^{(t)} - x^\star$ is such that $\|\eta^{(t)}\| < R_1$, then we have for all $i \in \Lambda$

$$|\eta_i^{(t)}| \leq \|\eta^{(t)}\| \leq R_1 \leq \rho |x_i^\star|.$$

We evince that for all $i \in \Lambda$

$$\begin{aligned}
&|x_i^\star| \left(\frac{\pi_i^{(t)}}{\alpha^{(t)}} + \frac{(1-\pi_i^{(t)})}{\beta^{(t)}} \right) \\
&= |x_i^\star| \left(\frac{\pi_i^{(t)} \sum_{\ell\in[n]} \pi_\ell^{(t)}}{\sum_{j\in[n]} \pi_j^{(t)} |x_j^{(t)}|^2 + |\epsilon^{(t)}|^2} \right. \\
&\left. + \frac{(1-\pi_i^{(t)}) \sum_{\ell\in[n]} (1-\pi_\ell^{(t)})}{\sum_{j\in[n]} (1-\pi_j^{(t)}) |x_j^{(t)}|^2 + |\epsilon^{(t)}|^2} \right) \\
&\leq |x_i^\star| \left(\frac{\pi_i^{(t)} \sum_{\ell\in[n]} \pi_\ell^{(t)}}{\sum_{j\in[n]} \pi_j^{(t)} |x_j^{(t)}|^2} + \frac{(1-\pi_i^{(t)}) \sum_{\ell\in[n]} (1-\pi_\ell^{(t)})}{\sum_{j\in\Lambda} (1-\pi_j^{(t)}) |x_j^{(t)}|^2} \right) \\
&\leq \frac{|x_i^\star| \pi_i^{(t)} n}{\sum_{j\in\Lambda} \pi_j^{(t)} |x_j^\star + \eta_j^{(t)}|^2} + \frac{|x_i^\star| (1-\pi_i^{(t)}) n}{\sum_{j\in\Lambda} (1-\pi_j^{(t)}) |x_j^\star + \eta_j^{(t)}|^2} \\
&\leq \frac{|x_i^\star| \pi_i^{(t)} n}{\sum_{j\in\Lambda} \pi_j^{(t)} (|x_j^\star| - |\eta_j^{(t)}|)^2} + \frac{|x_i^\star| (1-\pi_i^{(t)}) n}{\sum_{j\in\Lambda} (1-\pi_j^{(t)}) (|x_j^\star| - |\eta_j^{(t)}|)^2} \\
&\leq \frac{2n}{(1-\rho)^2 |x_i^\star|} \leq \frac{2n}{(1-\rho)^2 \min_{i\in\Lambda} |x_i^\star|} := A.
\end{aligned}$$

The combination of the above inequality with Lemma 4 and the NSP (see Definition (3)) yields

$$\sum_{i\in[n]}\left(\frac{\pi_i^{(t)}}{\alpha^{(t)}}+\frac{1-\pi_i^{(t)}}{\beta^{(t)}}\right)|\eta_i^{(t+1)}|^2 \leq A\|\eta^{(t+1)}|_\Lambda\|_1 \leq A\gamma\|\eta^{(t+1)}|_{\Lambda^c}\|_1.$$

Using the Cauchy-Scharz inequality and defining

$$\omega_i^{(t)}=\frac{\pi_i^{(t)}}{\alpha^{(t)}}+\frac{1-\pi_i^{(t)}}{\beta^{(t)}}$$

we obtain

$$\|\eta^{(t+1)}|_{\Lambda^c}\|_1^2 \leq \left(\sum_{i\in\Lambda^c}|\eta_i^{(t+1)}|^2\omega_i\right)\left(\sum_{i\in\Lambda^c}|\omega_i^{(t)}|^{-1}\right) \leq A\gamma\|\eta^{(t+1)}|_{\Lambda^c}\|_1\left(\sum_{i\in\Lambda^c}|\omega_i^{(t)}|^{-1}\right)$$

If $\eta^{(t+1)}|_{\Lambda^c}=0$ then the algorithm is stopped and in finite time the algorithm reaches the sparse solution $x^\star$. Let us suppose now that $\eta^{(t+1)}|_{\Lambda^c}\neq 0$ then

$$\|\eta^{(t+1)}|_{\Lambda^c}\|_2 \leq \|\eta^{(t+1)}|_{\Lambda^c}\|_1 \leq A\gamma\sum_{i\in\Lambda^c}|\omega_i^{(t)}|^{-1}.$$

It should be noticed that

$$\sum_{i\in\Lambda^c}|\omega_i^{(t)}|^{-1}=\sum_{i\in\Lambda^c}\left(\frac{\pi_i^{(t)}}{\alpha^{(t)}}+\frac{1-\pi_i^{(t)}}{\beta^{(t)}}\right)^{-1}\leq\sum_{i\in\Lambda^c}\frac{\alpha^{(t)}}{\pi_i^{(t)}} \leq (n-k)\alpha^{(t)}\left(1-R_2\right)^{-1}$$

from which we conclude

$$\|\eta^{(t+1)}|_{\Lambda^c}\|_2 \leq \chi\alpha^{(t)} \tag{37}$$

with $\chi=A\gamma(n-k)\left(1-R_2\right)^{-1}$. We thus have

$$\begin{aligned}\|\eta^{(t+1)}\|^2 &= \|\eta^{(t+1)}|_\Lambda\|^2+\|\eta^{(t+1)}|_{\Lambda^c}\|^2\\ &\leq (1+\gamma^2(n-k))\|\eta^{(t+1)}|_{\Lambda^c}\|^2\\ &\leq (1+\gamma^2(n-k))\chi^2|\alpha^{(t)}|^2\end{aligned} \tag{38}$$

If we define

$$\chi_1:=\chi\sqrt{1+\gamma^2(n-k)}=\frac{2n\gamma(n-k)\sqrt{1+\gamma^2(n-k)}}{(1-\rho)^2\min_{i\in\Lambda}|x_i^\star|(1-R_2)}$$

we obtain the first inequality in (32).

From the inductive hypothesis $|\beta^{(t)}-\frac{\|x^\star\|^2}{k}|<R_4$ and the fact that $(x^{(t)})_{t\in\mathbb{N}}$ is bounded (see Lemma 2) we have that there exists c so that $\forall i\in\Lambda$

$$\begin{aligned}\pi_i^{(t+1)} &\leq \sqrt{\frac{\beta^{(t)}}{\alpha^{(t)}}}\mathrm{e}^{-\frac{|x_i^{(t+1)}|^2}{2\alpha^{(t)}}+\frac{|x_i^{(t+1)}|^2}{2\beta^{(t)}}}\frac{1-p}{p}\\ &\leq \frac{c}{\sqrt{\alpha^{(t)}}}\mathrm{e}^{-\frac{(1-\rho)^2|x_i^\star|^2}{2\alpha^{(t)}}}\end{aligned} \tag{39}$$

and, consequently, $\pi_i^{(t+1)}\leq B|\alpha^{(t)}|^2$ for some positive constant B. Notice that, since $x_i^{(t+1)}$ are bounded and $\beta^{(t)}>\|x^\star\|^2/k-R_4>0$, then there exists a constant $C>0$ so that $\forall i\in\Lambda^c$

$$1-\pi_i^{(t+1)}\leq\sqrt{\frac{\alpha^{(t)}}{\beta^{(t)}}}\mathrm{e}^{\frac{|x_i^{(t+1)}|^2}{2}\left(\frac{1}{\alpha^{(t)}}-\frac{1}{\beta^{(t)}}\right)}\leq C\sqrt{\alpha^{(t)}}.$$

If we define $\chi_2=n\max(C,B)\sqrt{\alpha}$ we obtain the second inequality in (32).

From (26) and inequality (4.4) in [14] we have

$$|\epsilon^{(t+1)}|^2=\frac{(r(x^{(t+1)})_{K+1})^2}{n^2}\leq\frac{\|\eta^{(t+1)}\|_1^2}{(K+1-k)^2n^2}\leq D|\alpha^{(t)}|^2 \tag{40}$$

where $D:=\frac{\chi_1^2}{(K+1-k)^2n}$ and the last inequality follows from (38) and Cauchy-Schwarz inequality.

Let us consider now

$$\begin{aligned}\alpha^{(t+1)} &= \frac{\sum_{i\in[n]}\pi_i^{(t+1)}|x_i^{(t+1)}|^2+|\epsilon^{(t+1)}|^2}{\sum_{i\in[n]}\pi_i^{(t+1)}}\\ &= \frac{\sum_{i\in\Lambda}\pi_i^{(t+1)}|x_i^{(t+1)}|^2+\sum_{i\in\Lambda^c}\pi_i^{(t+1)}|\eta_i^{(t+1)}|^2+|\epsilon^{(t+1)}|^2}{\sum_{i\in[n]}\pi_i^{(t+1)}}\\ &\leq \frac{\sum_{i\in\Lambda}\pi_i^{(t+1)}|x_i^{(t+1)}|^2+\|\eta^{(t+1)}|_{\Lambda^c}\|^2+|\epsilon^{(t+1)}|^2}{\sum_{i\in[n]}\pi_i^{(t+1)}}.\end{aligned}$$

Combining the above inequality with (37), (39) and (40) and from hypothesis we obtain $\alpha^{(t+1)}\leq\chi_3|\alpha^{(t)}|^2$ with constant a positive constant χ_3.

Finally, we consider

$$\begin{aligned}&\left|\beta^{(t+1)}-\frac{\|x^\star\|^2}{k}\right|\leq\frac{|\epsilon^{(t+1)}|^2}{k(1-B|\alpha^{(t)}|^2)}\\ &+\frac{\left|k\sum_{i\in[n]}(1-\pi_i^{(t+1)})|x_i^{(t+1)}|^2-\|x^\star\|^2\sum_{i\in[n]}(1-\pi_i^{(t+1)})\right|}{k^2(1-B|\alpha^{(t)}|^2)}\\ &\leq\frac{D|\alpha^{(t)}|^2}{k(1-B|\alpha^{(t)}|^2)}+\frac{\left|k\sum_{i\in[n]}(|\eta_i^{(t+1)}|^2+2x_i^\star\eta_i^{(t+1)})\right|}{k^2(1-B|\alpha^{(t)}|^2)}\\ &+\frac{\left|k\sum_{i\in\Lambda}(1-\pi_i^{(t+1)})|x_i^\star|^2-\|x^\star\|^2\sum_{i\in[n]}(1-\pi_i^{(t+1)})\right|}{k^2(1-B|\alpha^{(t)}|^2)}\\ &\leq\frac{D|\alpha^{(t)}|^2}{k(1-B|\alpha^{(t)}|^2)}+\frac{\|\eta^{(t+1)}\|^2+2\max_{i\in\Lambda}|x_i^\star|\sqrt{n}\|\eta^{(t+1)}\|}{k(1-B|\alpha^{(t)}|^2)}\\ &+\frac{\left|k\sum_{i\in\Lambda}(1-\pi_i^{(t+1)})|x_i^\star|^2-\|x^\star\|^2\sum_{i\in\Lambda}(1-\pi_i^{(t+1)})\right|}{k^2(1-B|\alpha^{(t)}|^2)}\\ &+\frac{\|x^\star\|^2\sum_{i\in\Lambda^c}(1-\pi_i^{(t+1)})}{k^2(1-B|\alpha^{(t)}|^2)}\\ &\leq\frac{D|\alpha^{(t)}|^2}{k(1-B|\alpha^{(t)}|^2)}+\frac{\chi_1^2|\alpha^{(t)}|^2+2\max_{i\in\Lambda}|x_i^\star|\sqrt{n}\chi_1\alpha^{(t)}}{k(1-B|\alpha^{(t)}|^2)}\\ &+\frac{k\sum_{i\in\Lambda}\pi_i^{(t+1)}|x_i^\star|^2+\|x^\star\|^2\sum_{i\in\Lambda}\pi_i^{(t+1)}}{k^2(1-B|\alpha^{(t)}|^2)}\\ &+\frac{\|x^\star\|^2\sum_{i\in\Lambda^c}(1-\pi_i^{(t+1)})}{k^2(1-B|\alpha^{(t)}|^2)}\\ &\leq\frac{D|\alpha^{(t)}|^2}{k(1-B|\alpha^{(t)}|^2)}+\frac{\chi_1^2|\alpha^{(t)}|^2+2\max_{i\in\Lambda}|x_i^\star|\chi_1\sqrt{n}\alpha^{(t)}}{k(1-B|\alpha^{(t)}|^2)}\\ &+2\frac{\|x^\star\|^2B|\alpha^{(t)}|^2}{k(1-B|\alpha^{(t)}|^2)}+\frac{C\|x^\star\|^2(n-k)\sqrt{\alpha^{(t)}}}{k(1-B|\alpha^{(t)}|^2)}.\end{aligned}$$

We conclude that there exists a constant χ_4 such that $|\beta^{(t+1)}-\frac{\|x^\star\|^2}{k}|<\chi_4\sqrt{\alpha^{(t)}}$. If we consider the assumption in (33), we

now have

$$\|x^{(t+1)} - x^\star\| \leq R_1, \quad \|\pi^{(t+1)} - \mathbb{1}_{\Lambda^c}\| \leq R_2,$$
$$|\alpha^{(t+1)}| \leq R_3, \quad \left|\beta^{(t+1)} - \frac{\|x^\star\|^2}{k}\right| \leq R_4,$$

which completes the inductive step.

REFERENCES

[1] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207 – 1223, 2006.

[2] M. Fornasier, *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Radon Series on Computational and Applied Mathematics, 2010.

[3] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation: part i: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, Mar. 2006.

[4] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information theory*, vol. 53, pp. 4655–4666, 2007.

[5] D. Needell and R. Vershynin, "Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 310–316, 2010.

[6] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, Apr. 2008.

[7] D. Baron, S. Sarvotham, and R. Baraniuk, "Bayesian compressive sensing via belief propagation," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 269–280, Jan. 2010.

[8] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4299–4308, Sep. 2009.

[9] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, June 2008.

[10] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[11] E. Huebner and R. Tichatschke, "Relaxed proximal point algorithms for variational inequalities with multi-valued operators," *Optimization Methods Software*, vol. 23, no. 6, pp. 847–877, Dec. 2008.

[12] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[13] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[14] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Comm. Pure Appl. Math.*, vol. 63, no. 1, pp. 1–38, Jan. 2010.

[15] M. Wang, W. Xu, and A. Tang, "On the performance of sparse recovery via ℓ_p-minimization," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7255–7278, Nov 2011.

[16] D. E. Ba, B. Babadi, P. L. Purdon, and E. N. Brown, "Convergence and stability of iteratively re-weighted least squares algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 183–195, 2014.

[17] N. Bissantz, L. Dümbgen, A. Munk, and B. Stratmann, "Convergence analysis of generalized iteratively reweighted least squares algorithms on convex function spaces," *SIAM J. on Optimization*, vol. 19, no. 4, pp. 1828–1845, Feb. 2009.

[18] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4595–4607, Sep. 2010.

[19] M. P. Friedlander, H. Mansour, R. Saab, and O. Yilmaz, "Recovering compressively sampled signals using partial support information," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1122–1134, Feb. 2012.

[20] Y. Wang and W. Yin, "Sparse signal reconstruction via iterative support detection," *SIAM J. Img. Sci.*, vol. 3, no. 3, pp. 462–491, Aug. 2010.

[21] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *J. Mach. Learn. Res.*, vol. 11, pp. 1081–1107, Mar. 2010.

[22] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *J. Amer. Math. Soc*, pp. 211–231, 2009.

[23] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[24] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[25] K. Lange and J. S. Sinsheimer, "Normal/independent distributions and their applications in robust regression," *J. Compput. Graph Statist.*, vol. 2, no. 2, pp. 175–198, June 1993.

[26] D. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 58, no. 2, pp. 1094–1121, Feb. 2012.

[27] M. Tan, I. Tsang, and L. Wang, "Matching pursuit lasso part i: Sparse recovery over big dictionary," *IEEE Transactions on Signal Processing*, vol. 63, no. 3, pp. 727–741, Feb 2015.

[28] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing Sparsity by Reweighted ℓ_1 Minimization," Tech. Rep.

[29] A. Yeredor, "The joint MAP-ML criterion and its relation to ML and to extended least-squares," *IEEE Transactions on Signal Processing*, vol. 48, no. 12, pp. 3484 – 3492, 2000.

[30] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley, 1985.

[31] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, pp. 195 – 239, 1984.

[32] F. Fagnani, S. M. Fosson, and C. Ravazzi, "A distributed classification/estimation algorithm for sensor networks," *SIAM J. on Control and Optimization*, vol. 52, no. 1, pp. 189–218, Jan. 2014.

[33] E. van den Berg, M. P. Friedlander, G. Hennenfent, F. J. Herrmann, R. Saab, and Özgür Yilmaz", "Algorithm 890: Sparco: A testing framework for sparse reconstruction," *ACM Transactions on Mathematical Software*, vol. 35, no. 4, pp. 1–16, Feb. 2009.

[34] R. Baraniuk, M. Davenport, D. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[35] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences. Paris, France, ser. I*, vol. 346, pp. 589 – 592, 2008.

Chiara Ravazzi (M'13) received the B.Sc. and M.Sc., and Ph.D. degrees in applied mathematics from Politecnico di Torino, Italy, in 2005 and 2007, and 2011 respectively. She is currently a Postdoctoral Associate at the Department of Electronics and Telecommunications (DET), Politecnico di Torino, Italy. During 2010, she was a visiting student at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge. Specific themes of current interest include the mathematics of control and information theory, coding theory, and signal processing.

Enrico Magli (S'97–M'01–SM'07) received the M.Sc. and Ph.D. degrees from Politecnico di Torino, Torino, Italy, in 1997 and 2001, respectively. He is currently an Associate Professor with Politecnico di Torino, Torino, Italy. His research interests are in the field of compressive sensing, image and video coding, and vision. He is an associate editor of the IEEE Transactions on circuits and systems for video technology, the IEEE Transactions on multimedia, and the EURASIP Journal on image and video processing, an IEEE Distinguished Lecturer for 2015-2016, and a corecipient of the IEEE Geoscience and Remote Sensing Society 2011 Transactions Prize Paper Award.