

# Recipes on Hard Thresholding Methods

Anastasios Kyrillidis and Volkan Cevher

Laboratory for Information and Inference Systems  
Ecole Polytechnique Federale de Lausanne  
{anastasios.kyrillidis,volkan.cevher}@epfl.ch

**Abstract**—We present and analyze a new set of sparse recovery algorithms within the class of hard thresholding methods. We provide optimal strategies on how to set up these algorithms via basic “ingredients” for different configurations to achieve complexity vs. accuracy tradeoffs. Simulation results demonstrate notable performance improvements compared to state-of-the-art algorithms both in terms of data reconstruction and computational complexity.

## I. INTRODUCTION

We consider the following underdetermined linear inverse problem: assume that high-dimensional signal  $x^* \in \mathbb{R}^N$  is observed through a low-dimensional observation vector  $u \in \mathbb{R}^M$  ( $M < N$ ) via:

$$u = \Phi x^* + n. \quad (1)$$

In this setting,  $\Phi \in \mathbb{R}^{M \times N}$  represents the regression/sensing matrix and  $n \in \mathbb{R}^M$  is an additive noise term. Given  $u$  and  $\Phi$ , unconstrained least-squares method is the classic approach to the solution of linear systems by minimizing the data error function  $f(x) \triangleq \|u - \Phi x\|_2^2$ . Nevertheless, the reconstruction of  $x^*$  from  $u$  is an ill-posed problem since  $M < N$  and there is no hope in finding the *true vector* without ambiguity; additional prior information is needed. Therefore, we assume that  $x^*$  is a sparse vector with structure defined by a combinatorial sparsity model (CSM)  $\mathcal{C}_K$  where  $K \ll N$  represents the sparsity parameter (c.f., [1] for details on such sets).

In this paper, we concentrate on the following constrained minimization problem to recover  $x^*$ :

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{C}_K. \quad (2)$$

Unfortunately, prior knowledge on signal structure does not guarantee successful recovery of the true vector for *any sensing matrix*. Many conditions on  $\Phi$  have been proposed in the literature to establish solution uniqueness and reconstruction stability such as null space property, spark, unique representation property to name a few. Here, we focus on the so-called restricted isometry property (RIP). Given CSM  $\mathcal{C}_K$ ,  $\Phi$  satisfies the RIP with constant  $\delta_K$  if and only if

$$(1 - \delta_K)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)\|x\|_2^2, \quad \forall x \in \mathcal{C}_K. \quad (3)$$

While the majority of CS results assume (3) is satisfied with symmetry, we further consider the non-symmetric analog of the RIP:

$$\alpha_K \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq \beta_K \|x\|_2^2, \quad \forall x \in \mathcal{C}_K, \quad (4)$$

for positive constants  $\alpha_K, \beta_K$ .

In contrast to the conventional convex relaxation of this problem [2], we maintain the combinatorial nature of (2) and focus on the class of hard thresholding methods; c.f., [3] for a brief overview of existing variants. As a running example for our analysis, model-based

Iterative Hard Thresholding (IHT) algorithm [4] is used, characterized by the following two-step recursion:

$$\bar{x}_i = x_i - \frac{\mu}{2} \nabla f(x_i), \quad x_{i+1} = \mathcal{P}_{\mathcal{C}_K}(\bar{x}_i). \quad (5)$$

Here,  $i$  is the iteration number,  $\mu$  is the gradient descent step size,  $\nabla f(x) \triangleq -2\Phi^T(u - \Phi x)$  denotes the gradient of the objective function  $f(x)$ , and  $\mathcal{P}_{\mathcal{C}_K}(\cdot)$  is the combinatorial projection onto the subspace defined by CSM  $\mathcal{C}_K$  according to:

$$\mathcal{P}_{\mathcal{C}_K}(y) = \underset{x: x \in \mathcal{C}_K}{\text{argmin}} \|x - y\|_2. \quad (6)$$

In this paper, we concentrate on CSM cases where  $\mathcal{P}_{\mathcal{C}_K}(\cdot)$  projection is exactly computed in polynomial time (defined as PMAP<sub>0</sub> in [1]); examples include the simple sparsity model, and models with matroid or totally unimodular constraints.

To characterize the performance of the iterative process (5) both in terms of convergence rate and noise resilience, we use the following recursive expression:

$$\|x_{i+1} - x^*\|_2 \leq \rho \|x_i - x^*\|_2 + \gamma \|n\|_2. \quad (7)$$

In (7),  $\gamma$  denotes the approximation guarantee and provides insights into algorithm’s reconstruction capabilities when additive noise is present;  $\rho < 1$  expresses the convergence rate towards a region around  $x^*$ , whose radius is determined by  $\frac{\gamma}{1-\rho} \|n\|_2$ .

In each iteration, computational requirements of hard thresholding methods mainly depend on the total number of matrix-vector multiplication operations. Different problem configurations (e.g. comparable sparsity level with respect to the number of available measurements, etc.) lead to hard thresholding variants that guarantee stability and noise robustness but additional matrix  $\Phi$  applications (and its adjoint  $\Phi^T$ ) are required per iteration; hence, low iteration counts are desired to trade-off these operations. Furthermore, assuming CSM instead of the simple sparsity model introduces elaborate structure constraints, rendering the total number of combinatorial projections a non-negligible factor with respect to the overall complexity of the algorithm.

**Contributions:** We propose and analyze new recipes for hard thresholding methods. Three basic building blocks (“ingredients”) are studied: *i*) step size selection  $\mu_i$ , *ii*) memory exploitation, and *iii*) gradient or least-squares updates over restricted support sets. We highlight the impact of these blocks on the convergence rate and signal reconstruction performance and provide optimal and/or efficient strategies on how to set up these “ingredients” under different problem conditions. Finally, we provide empirical support for our claims for better data recovery performance and reduced complexity through experimental results on synthetic data.

**Notation:** We use  $[x]_j$  to denote the  $j$ -th element of  $x$ , and let  $x_i$  represent the  $i$ -th iterate of the hard thresholding method. The index set of  $N$  dimensions is denoted as  $\mathcal{N} = \{1, 2, \dots, N\}$ . Given  $\mathcal{S} \subseteq$

This work was supported in part by the European Commission under Grant MIRC-268398 and DARPA KeCoM program #11-DARPA-1055. VC also would like to acknowledge Rice University for his Faculty Fellowship.

$\mathcal{N}$ , we define the complement set  $\mathcal{S}^c = \mathcal{N} \setminus \mathcal{S}$ . Moreover, given a set  $\mathcal{S} \subseteq \mathcal{N}$  and a vector  $x \in \mathbb{R}^N$ ,  $x_{\mathcal{S}} \in \mathbb{R}^N$  denotes a vector with the following properties:  $[x_{\mathcal{S}}]_{\mathcal{S}} = [x]_{\mathcal{S}}$  and  $[x_{\mathcal{S}}]_{\mathcal{S}^c} = 0$ . The notation  $\nabla_{\mathcal{S}} f(x)$  is shorthand for  $[\nabla f(x)]_{\mathcal{S}}$ .  $\Phi_T$  represents the restriction of the matrix  $\Phi$  to a column submatrix whose columns are listed in the set  $T$ . The support set of  $x$  is defined as  $\text{supp}(x) = \{i : [x]_i \neq 0\}$ . We use  $|\mathcal{S}|$  to denote the cardinality of the set  $\mathcal{S}$ . The inner product between two vectors  $\alpha, \beta \in \mathbb{R}^N$  is denoted as  $\langle \alpha, \beta \rangle = \alpha^T \beta = \sum_{i=1}^N [\alpha]_i [\beta]_i$  where  $T$  is the transpose operation.  $\|\cdot\|_2$  denotes the  $l_2$ -norm where  $\|x\|_2 = \sqrt{\langle x, x \rangle}$ .  $\mathbb{I}$  represents an identity matrix with dimensions apparent from the context.

## II. STEP SIZE SELECTION

To emphasize how step size selection  $\mu_i$  affects both the convergence rate  $\rho$  and the approximation guarantee  $\gamma$ , we derive the convergence proof of model-based IHT where step size  $\mu_i$  is considered as a variable, using techniques described in [6]. Given non-symmetric RIP assumption, the following recursive formula holds true:

$$\|x_{i+1} - x^*\|_2 \leq 2\|\mathbb{I} - \mu_i \Phi_T^* \Phi_T\|_{2 \rightarrow 2} \|x_i - x^*\|_2 + 2\mu_i \sqrt{\beta_{2K}} \|n\|_2, \quad (8)$$

where  $T = \text{supp}(x^*) \cup \text{supp}(x_{i+1}) \cup \text{supp}(x_i)$  with  $|T| \leq 3K$  and

$$\|\mathbb{I} - \mu_i \Phi_T^* \Phi_T\|_{2 \rightarrow 2} \leq \max \left\{ \mu_i \lambda_{\max}(\Phi_T^* \Phi_T) - 1, 1 - \mu_i \lambda_{\min}(\Phi_T^* \Phi_T) \right\}. \quad (9)$$

In the case of hard thresholding methods, recent works on the performance of IHT algorithm provide strong convergence rate guarantees in terms of RIP constants; c.f. [5]. However, as a prerequisite to achieve these strong isometry constant bounds, the step size is set  $\mu_i = 1, \forall i$ , given that  $\|\Phi\|_2^2 < 1$ . From a different perspective, [3] proposes a constant step size  $\mu_i = 1/(1 + \delta_{2K}), \forall i$ , based on a simple convergence analysis of the gradient descent method.

Unfortunately, most of the above problem assumptions are not naturally met; the authors in [7] provide an intuitive example where IHT algorithm behaves differently under various scalings of the sensing matrix  $\Phi$ . Violation of these configuration details usually lead to unpredictable signal recovery performance of hard thresholding methods. Therefore, more sophisticated step size selection procedures should be devised to tackle these computational issues during actual recovery. On the other hand, the computation of RIP constants has exponential time complexity for the strategy of [3] and exhaustive combinatorial search is necessary.

Existing approaches broadly fall into two categories: constant and adaptive step size selection. For both cases, we present efficient strategies to select the step size  $\mu_i$  that implies the fastest convergence rate but not necessarily the best approximation guarantee.

### A. Constant step size selection

As a first scenario, assume  $\Phi$  satisfies the non-symmetric RIP with known  $\alpha_{cK}, \beta_{cK}, (c = 2, 3)$  constants. In this case,

$$\lambda(\Phi_T^* \Phi_T) \in [\alpha_{3K}, \beta_{3K}]. \quad (10)$$

To optimize the convergence rate, we can pick  $\mu_i$  as the minimizer of the expression:

$$\min_{\mu_i} \|\mathbb{I} - \mu_i \Phi_T^* \Phi_T\|_{2 \rightarrow 2} \leq \min_{\mu_i} \max \left\{ \mu_i \beta_{3K} - 1, 1 - \mu_i \alpha_{3K} \right\}, \quad (11)$$

which leads to the following result, inspired by convex optimization constant step size strategies [9].

*Proposition 1 (Non-symmetric RIP constant step size strategy):* Assume  $\Phi$  satisfies the non-symmetric RIP with known upper/lower bounds  $\alpha_{cK}, \beta_{cK}, (c = 2, 3)$ . The step size  $\mu_i$  that implies the fastest convergence rate in (8) amounts to

$$\mu_i = \frac{2}{\alpha_{3K} + \beta_{3K}}, \forall i = \{1, 2, \dots\},$$

where  $\rho = \frac{2(\beta_{3K} - \alpha_{3K})}{\alpha_{3K} + \beta_{3K}} < 1$  for  $\beta_{3K} < 3\alpha_{3K}$  and  $\gamma = \frac{2\sqrt{\beta_{2K}}}{\alpha_{3K} + \beta_{3K}}$ . *Proof:* It is obvious that the step size  $\mu_i$  that minimizes (11) lies at the intersection of the linear functions  $\psi_1(\mu_i) \triangleq \mu_i \beta_{3K} - 1$ ,  $\psi_2(\mu_i) \triangleq 1 - \mu_i \alpha_{3K}$ . Hence, the minimum occurs when

$$\psi_1(\mu_i) = \psi_2(\mu_i) \Rightarrow \mu_i = \frac{2}{\alpha_{3K} + \beta_{3K}}. \quad (12)$$

In the special case where  $\Phi$  satisfies the RIP (3) for some constant  $\delta_{3K}$ , (8) becomes:

$$\|x_{i+1} - x^*\|_2 \leq 2\|\mathbb{I} - \mu_i \Phi_T^* \Phi_T\|_{2 \rightarrow 2} \|x_i - x^*\|_2 + 2\mu_i \sqrt{1 + \delta_{2K}} \|n\|_2 \quad (13)$$

Following the same proof technique, we conclude to the same convergence rate achieved in [6].

*Corollary 1 (RIP constant step size strategy):* Given  $\Phi$  satisfies the RIP for some  $\delta_{3K}$ , the step size  $\mu_i$  that implies the fastest convergence rate in (13) amounts to

$$\mu_i = 1, \forall i = \{1, 2, \dots\}, \quad (14)$$

with  $\rho = 2\delta_{3K}$  and  $\gamma = 2\sqrt{1 + \delta_{2K}}$ . Moreover, the iterations are contractive iff  $\delta_{3K} < 1/2 \Rightarrow \rho < 1$ .

*Remark 1:* If (4) holds and  $\alpha_{3K}, \beta_{3K}$  are known, we observe that  $\widehat{\Phi} = \sqrt{\frac{2}{\alpha_{3K} + \beta_{3K}}} \Phi$  satisfies (3)  $\forall x \in \mathcal{C}_{3K}$  with  $\delta_{3K} = \frac{\beta_{3K} - \alpha_{3K}}{\alpha_{3K} + \beta_{3K}}$ . In this case,  $\mu_i = 1$  implies the fastest convergence in (8).

### B. Adaptive step size selection

Since the computation of the exact RIP bounds is combinatorially hard, the assumptions made for constant step size selection strategies are unverifiable even for moderate-sized random matrices. To improve stability, an adaptive scheme is mandatory.

There is limited work on the adaptive step size selection for hard thresholding methods. To the best of our knowledge, [7]-[8] are the only studies that attempt this via line searching.

According to (5), let  $x_i \in \mathcal{C}_K$  be the  $K$ -sparse signal estimate with known support  $\mathcal{X}_i \triangleq \text{supp}(x_i)$  at the  $i$ -th iteration. It then holds that the non-zero elements  $[x_{i+1}]_j, \forall j \in \mathcal{X}_{i+1} \triangleq \text{supp}(x_{i+1})$  of the new estimate satisfy:

$$[x_{i+1}]_j = \begin{cases} -\frac{\mu_i}{2} [\nabla f(x_i)]_j & \text{if } [x_i]_j = 0, \\ [x_i]_j - \frac{\mu_i}{2} [\nabla f(x_i)]_j & \text{otherwise.} \end{cases}$$

for any step size  $\mu_i$ . Since  $|\mathcal{X}_{i+1}| \leq K$ , we easily deduce the following key observation:

*Remark 2:* Let  $\mathcal{S}_i$  be a  $2K$ -sparse support set defined as:

$$\mathcal{S}_i = \mathcal{X}_i \cup \text{supp}(\mathcal{P}_{\mathcal{C}_K}(\nabla_{\mathcal{X}_i^c} f(x_i))). \quad (15)$$

Given  $\mathcal{X}_{i+1}$  is unknown at the  $i$ -th iteration,  $\mathcal{S}_i$  is the smallest index set that contains it such that the following equality

$$\mathcal{P}_{\mathcal{C}_K} \left( x_i - \frac{\mu_i}{2} \nabla f(x_i) \right) = \mathcal{P}_{\mathcal{C}_K} \left( x_i - \frac{\mu_i}{2} \nabla_{\mathcal{S}_i} f(x_i) \right) \quad (16)$$

necessarily holds.

Using Remark 2, model-based IHT [4] can be equivalently written as

$$\bar{x}_i = x_i - \frac{\mu_i}{2} \nabla_{\mathcal{S}_i} f(x_i), \quad x_{i+1} = \mathcal{P}_{\mathcal{C}_K}(\bar{x}_i), \quad (17)$$

where  $\bar{x}_i \in \mathcal{C}_{2K}$  with  $\text{supp}(\bar{x}_i) \subseteq \mathcal{S}_i$ . To compute step-size  $\mu_i$ , we propose:

$$\mu_i = \underset{\mu}{\text{argmin}} \left\| u - \Phi \left( x_i - \frac{\mu}{2} \nabla_{\mathcal{S}_i} f(x_i) \right) \right\|_2^2 = \frac{\| \nabla_{\mathcal{S}_i} f(x_i) \|_2^2}{\| \Phi \nabla_{\mathcal{S}_i} f(x_i) \|_2^2}, \quad (18)$$

i.e.,  $\mu_i$  is the minimizer of the objective function. Note that  $1 - \delta_{2K} \leq 1/\mu_i \leq 1 + \delta_{2K}$  and  $\alpha_{2K} \leq 1/\mu_i \leq \beta_{2K}$  due to RIP and non-symmetric RIP, respectively. The proposed adaptive step size selection strategy leads to the following theorem, whose proof is omitted due to lack of space:

**Theorem 1 (Iteration Invariant):** Assume  $\Phi \in \mathbb{R}^{M \times N}$  satisfies (4) with  $\alpha_{cK}, \beta_{cK}, (c = 2, 3)$  unknown. In the worst case scenario, model-based IHT [4] with adaptive step size selection (18) satisfies the following recursive formula:

$$\|x_{i+1} - x^*\|_2 \leq \rho \|x_i - x^*\|_2 + \gamma \|n\|_2, \quad (19)$$

where  $\rho = 2 \max\{\frac{\beta_{3K}}{\alpha_{2K}} - 1, 1 - \frac{\alpha_{3K}}{\beta_{2K}}\}$  and  $\gamma = \frac{2\sqrt{\beta_{2K}}}{\alpha_{2K}}$ .

**Corollary 2:** Assuming RIP (3) with constants  $\delta_{cK}, (c = 2, 3)$ , (19) is rewritten as:

$$\|x_{i+1} - x^*\|_2 \leq 2 \frac{\delta_{3K} + \delta_{2K}}{1 - \delta_{2K}} \|x_i - x^*\|_2 + \frac{2\sqrt{1 + \delta_{2K}}}{1 - \delta_{2K}} \|n\|_2,$$

where  $\delta_{3K} < 1/5 \Rightarrow \rho = 2 \frac{\delta_{3K} + \delta_{2K}}{1 - \delta_{2K}} < 1$ .

We observe that adaptive  $\mu_i$  scheme results in more restrictive “worst-case” isometry constants compared to [6], [11], but faster convergence and better stability are empirically observed, as shown in Section V.

### III. MEMORY

Iterative algorithms can use memory to gain momentum in convergence. The success of the memory-based approaches depends on the iteration dependent momentum term by leveraging previous estimates. Based on Nesterov’s optimal gradient methods [9], [12] proposes the following hard thresholding variant:

$$x_i = \mathcal{P}_{\mathcal{C}_K} \left( y_i - \frac{\mu_i}{2} \nabla_{\mathcal{S}_i} f(y_i) \right), \quad y_{i+1} = x_i + \tau_i (x_i - x_{i-1}), \quad (20)$$

where  $\mathcal{Y}_i = \text{supp}(y_i)$ ,  $\mathcal{S}_i = \mathcal{Y}_i \cup \text{supp}(\mathcal{P}_{\mathcal{C}_K}(\nabla_{\mathcal{Y}_i^c} f(y_i)))$  with  $|\mathcal{S}_i| \leq 3K$  and  $\tau_i$  represents the momentum step size.

Similarly to  $\mu_i$  strategies,  $\tau_i$  can be preset as constant or adaptively computed at each iteration. Constant momentum step size selection has no additional computational cost but convergence rate acceleration is not guaranteed for a wide range of problem formulations. On the other hand, empirical evidence has shown that adaptive  $\tau_i$  selection strategies result to faster convergence with (almost) *equivalent complexity* to zero-memory methods.

For the case of strongly convex objective functions, Nesterov [9] proposed the following constant momentum step size selection scheme for (20)<sup>1</sup>:  $\tau_i = \frac{\alpha_i(1-\alpha_i)}{\alpha_i^2 + \alpha_{i+1}}$ , where  $\alpha_0 \in (0, 1)$  and  $\alpha_{i+1}$  is computed as the root  $\in (0, 1)$  of

$$\alpha_{i+1}^2 = (1 - \alpha_{i+1})\alpha_i^2 + q\alpha_{i+1}, \quad \text{for } q \triangleq \frac{1}{\kappa^2(\Phi)} = \frac{\sigma_{\min}^2(\Phi)}{\sigma_{\max}^2(\Phi)},$$

where  $\kappa(\Phi)$  denotes the condition number of  $\Phi$  and  $\sigma_{\min}(\Phi)$ ,  $\sigma_{\max}(\Phi)$  denote the minimum and maximum singular values of  $\Phi$ . In this scheme, exact calculation of  $q$  parameter is computationally expensive for large-scale data problems and approximation schemes are leveraged to compensate this complexity bottleneck.

<sup>1</sup>Authors thank Francis Bach for pointing out this scheme.

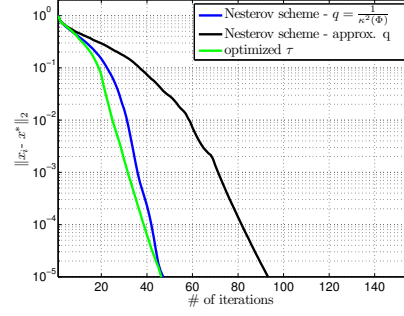


Fig. 1. Model-based IHT convergence rate example using memory; simple sparsity model assumed with  $N = 2000, M = 600, K = 120$ . Blue and black lines represent Nesterov’s  $\tau_i$  selection scheme with  $q = 1/\kappa^2(\Phi)$  and  $q \sim \mu_i^{\min}/\mu_i^{\max}$ , respectively; green line represents the proposed momentum step size selection.

Based upon the same ideas as adaptive  $\mu_i$  selection, we propose to select  $\tau_i$  as the minimizer of the objective function<sup>2</sup>:

$$\tau_i = \underset{\tau}{\text{argmin}} \|u - \Phi y_{i+1}\|_2^2 = \frac{\langle u - \Phi x_i, \Phi x_i - \Phi x_{i-1} \rangle}{\|\Phi x_i - \Phi x_{i-1}\|_2^2}, \quad (21)$$

where  $\Phi x_i, \Phi x_{i-1}$  are *previously computed*. According to (21),  $\tau_i$  requires only vector-vector inner product operations, a computationally cheaper operation than  $q$  calculation. Convergence rate performance of the above schemes is depicted in Fig. 1.

### IV. UPDATES OVER RESTRICTED SUPPORT SETS

At each iteration, the new estimate  $x_{i+1} = \mathcal{P}_{\mathcal{C}_K} \left( x_i - \frac{\mu_i}{2} \nabla f(x_i) \right)$  can be further refined by applying a single or multiple gradient descent updates with line search restricted on  $\mathcal{X}_{i+1}$  [11]:

$$x_{i+1} = x_{i+1} - \frac{\bar{\mu}_i}{2} \nabla_{\mathcal{X}_{i+1}} f(x_{i+1}), \quad \text{where } \bar{\mu}_i = \frac{\|\nabla_{\mathcal{X}_{i+1}} f(x_{i+1})\|_2^2}{\|\Phi \nabla_{\mathcal{X}_{i+1}} f(x_{i+1})\|_2^2}$$

or solving the minimization problem over  $\mathcal{X}_{i+1}$  [10]-[11]:

$$x_{i+1} = \underset{x: \text{supp}(x) \subseteq \mathcal{X}_{i+1}}{\text{argmin}} \|u - \Phi x\|_2^2. \quad (22)$$

Using the same ideas in our adaptive  $\mu_i$  selection scheme, a more accurate but computationally intensive alternative to gradient descent update in (17) is the objective minimization problem restricted on the support set  $\mathcal{S}_i$ , similar to (22).

### V. EXPERIMENTS

To set up our experiments, we incorporate these tricks into the ALPS toolbox, which is available at <http://lions.epfl.ch/ALPS>. The naming convention borrows from [13].

#### A. Experiment 1: Computational complexity and convergence rate

We generate 50 random Monte-Carlo realizations according to (1) where  $N = 5000, M = 2000$  and  $K = 700$ .  $\Phi$  is a dense random matrix with independent entries, sampled from zero-mean Gaussian distribution with variance  $1/M$ . The sparse signal  $x^*$  follows the simple sparsity model with  $K$  nonzero elements, acquired according to standard normal distribution with  $\|x^*\|_2 = 1$ . In Fig. 2, we compare five different hard thresholding methods in terms of convergence rate.

We also provide in Table 1 the matrix-vector multiplication complexity per iteration along with the total number of projections  $\mathcal{P}_{\mathcal{C}_K}(\cdot)$ .

<sup>2</sup>Similar ideas were simultaneously proposed in [8].

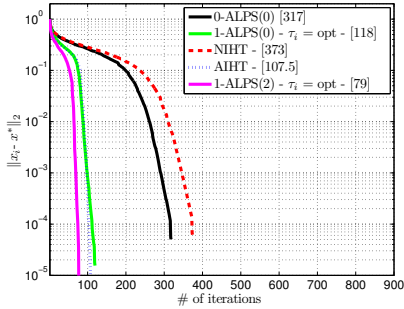


Fig. 2. Median error per iter. - [median # of iter.] - #-ALPS(0): adaptive  $\mu_i$  with # memory, NIHT: Normalized IHT [7], AIHT: NIHT with Double Relaxation [8], 1-ALPS(2): adaptive  $\mu_i$  and additional gradient update.

TABLE I

# of scalar multiplications per iter.		# of $\mathcal{P}_{C_K}(\cdot)$ per iter.	
0-ALPS(0)	$MN + 3MK$	0-ALPS(0)	2
NIHT <sup>3</sup>	$MN + 2MK$	NIHT <sup>3</sup>	2
AIHT <sup>3</sup>	$MN + 3MK$	AIHT <sup>3</sup>	3
1-ALPS(0)	$MN + 3MK$	1-ALPS(0)	2
1-ALPS(2)	$MN + 5MK$	1-ALPS(2)	2

### B. Experiment 2: Memory does not hurt

Fig. 3 illustrates the phase transition diagrams of 0-ALPS(0) and 1-ALPS(0) algorithms. The ambient dimension of the true signal is  $N = 1000$ . We observe that memory acceleration does not degrade the signal reconstruction performance compared to equivalent zero-memory schemes. As a side remark, we note that 1-ALPS(0) has a better phase transition performance as compared to AIHT [8] and NIHT [7] (not shown due to lack of space).

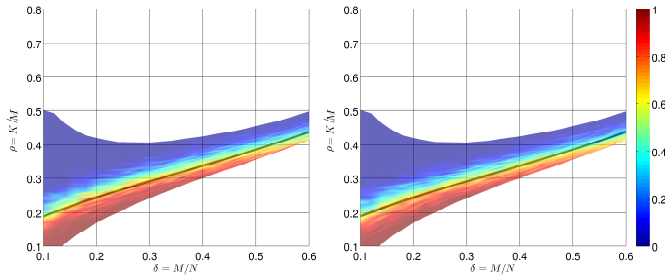


Fig. 3. Empirical phase transition performance of 0-ALPS(0) (left column) and 1-ALPS(0) (right column) algorithms. A signal recovery with solution  $\hat{x}$  is considered successful provided that  $\|\hat{x} - x^*\|_2 < 10^{-6}$ . Solid black line denotes the theoretical  $l_1$  minimization phase transition curve.

### C. Experiment 3: Phase transition performance

In this experiment, we compare the signal recovery behaviour of 0-ALPS(4) algorithm using our adaptive step size selection and HTP algorithm [11] with NIHT adaptive  $\mu_i$  selection [7]. Here, we assume  $N = 1000$ . The empirical phase transition results are depicted in Fig. 4.

## VI. CONCLUSIONS

In this paper, we present and review three building blocks of hard thresholding methods along with optimal/efficient strategies for their

<sup>3</sup>Best case scenario where no additional binary line search over  $\mu_i$  is needed.

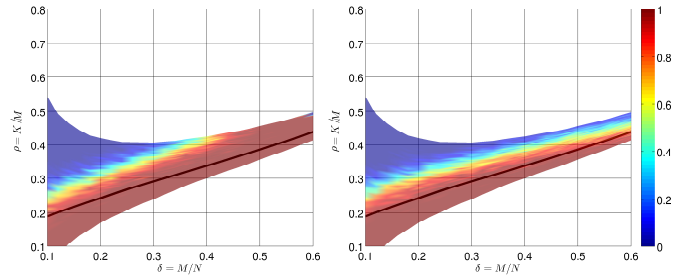


Fig. 4. Empirical phase transition performance of 0-ALPS(4) with the proposed step size selection (left column) and HTP with NIHT step size selection (right column). A signal recovery with solution  $\hat{x}$  is considered successful provided that  $\|\hat{x} - x^*\|_2 < 10^{-6}$ . Solid black line denotes the theoretical  $l_1$  minimization phase transition curve.

usage. In theory, constant  $\mu_i$  selection schemes are accompanied with strong RIP constant conditions but empirical evidence reveal signal reconstruction vulnerabilities even for small deviations from the initial problem assumptions. While convergence derivations of adaptive schemes are characterized by weaker bounds, the performance gained by this choice, both in terms of convergence rate and data recovery, is quite significant. Memory-based methods lead to convergence speed with (almost) no extra cost on the complexity of hard thresholding methods but more theoretical justification is needed; future work will likely focus on this direction. Lastly, further estimate refinement over sparse support sets using gradient update steps or pseudoinversion optimization techniques provides signal reconstruction efficacy, but more computational power is needed per iteration. In all cases, experimental results illustrate the effectiveness of the proposed schemes on different problem configurations.

## REFERENCES

- [1] A. Kyrillidis and V. Cevher, *Combinatorial selection and least absolute shrinkage via the CLASH algorithm*, Technical Report, 2011.
- [2] J. A. Tropp and S. J. Wright, *Computational methods for sparse solution of linear inverse problems*, Proceedings of the IEEE, 98(6):948-958, 2010.
- [3] R. Garg, and R. Khandekar, *Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property*, In ICML. ACM, 2009.
- [4] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, *Model-based compressive sensing*, Information Theory, IEEE Trans. on, 56(4):1982-2001, 2010.
- [5] T. Blumensath, M. E. Davies, *Iterative hard thresholding for compressed sensing*, Applied and Computational Harmonic Analysis, vol. 27, no. 3, pp. 265-274, 2009.
- [6] S. Foucart, *Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants*, In Proceedings of the 13th International Conference on Approximation Theory, 2010.
- [7] T. Blumensath, and M. E. Davies, *Normalized iterative hard thresholding: Guaranteed stability and performance*, Selected Topics in Signal Processing, IEEE Journal of, 4(2): 298-309, 2010.
- [8] T. Blumensath, *Accelerated iterative hard thresholding*, preprint, 2011.
- [9] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, Springer, 2004.
- [10] W. Dai, and O. Milenkovic, *Subspace pursuit for compressive sensing signal reconstruction*, Information Theory, IEEE Trans. on, 55(5):2230-2249, 2009.
- [11] S. Foucart, *Hard thresholding pursuit: An algorithm for compressive sensing*, preprint, 2010.
- [12] V. Cevher, *An ALPS view of sparse recovery*, In Proceedings of IEEE ICASSP, 2011.
- [13] V. Cevher, *On accelerated hard thresholding methods for sparse approximation*, Technical report, 2011.