Iterative Hard Thresholding with Near Optimal Projection for Signal Recovery

Raja Giryes and Michael Elad
Computer Science Department
Technion - IIT 32000, Haifa, ISRAEL
Email: [raja,elad]@cs.technion.ac.il

***Abstract*—Recovering signals that have sparse representations under a given dictionary from a set of linear measurements got much attention in the recent decade. However, most of the work has focused on recovering the signal's representation, forcing the dictionary to be incoherent and with no linear dependencies between small sets of its columns. A series of recent papers show that such dependencies can be allowed by aiming at recovering the signal itself. However, most of these contributions focus on the analysis framework. One exception to these is the work reported in [1], proposing a variant of the CoSaMP for the synthesis model, and showing that signal recovery is possible even in high-coherence cases. In the theoretical study of this technique the existence of an efficient near optimal projection scheme is assumed. In this paper we extend the above work, showing that under very similar assumptions, a variant of IHT can recover the signal in cases where regular IHT fails.**

I. Introduction

Recovering a sparse signal from a given set of linear measurements has been a major subject of research in recent years. In the basic setup, an unknown signal $\mathbf{x}_0 \in \mathbb{R}^d$ passes through a given linear transformation $\mathbf{M} \in \mathbb{R}^{m\times d}$ with an additive noise $\mathbf{e} \in \mathbb{R}^m$ providing a set of linear measurements $\mathbf{y} = \mathbf{M}\mathbf{x}_0 + \mathbf{e}$. The signal $\mathbf{x}_0$ is assumed to have a k-sparse representation $\boldsymbol{\alpha}_0 \in \mathbb{R}^n$ under a given dictionary $\mathbf{D} \in \mathbb{R}^{d\times n}$, i.e. $\mathbf{x}_0 = \mathbf{D}\boldsymbol{\alpha}_0$, $\|\boldsymbol{\alpha}_0\|_0 \leq k$ and $k \ll d$, where $\|\cdot\|_0$ is the "ℓ_0-norm" that counts the number of non-zero entries in a vector. The sparsity prior results with the following minimization problem

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{M}\mathbf{D}\boldsymbol{\alpha}\|_2 \quad s.t. \quad \|\boldsymbol{\alpha}\|_0 \leq k, \tag{1}$$

in which we pursue the representation $\boldsymbol{\alpha}$ in order to recover the original signal $\mathbf{x}_0$ from $\mathbf{y}$. Given a reconstructed representation $\hat{\boldsymbol{\alpha}}$, the estimation for the signal is simply given by $\hat{\mathbf{x}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$.

Solving (1) is a NP-hard problem and many approximation techniques has been proposed for it [2]. One of these is the iterative hard thresholding (IHT) algorithm [3]. This approach, summarized in Algorithm 1, recovers the representation in an iterative way using two repeating steps: (i) Gradient step: moving in the optimal gradient direction for minimizing $\|\mathbf{y} - \mathbf{M}\mathbf{D}\boldsymbol{\alpha}\|_2$; (ii) Projection step: ensuring that the representation estimate is k-sparse. The operator $\mathrm{supp}(\cdot, k)$ returns the support of the largest k elements in a given vector and the subscript T for a vector/matrix means taking the entries/columns corresponding to the indices in T.

In order to evaluate the performance of IHT, the restricted isometry property (RIP) [4] of the matrix $\mathbf{MD}$ is used. A matrix $\mathbf{A} \in \mathbb{R}^{d\times n}$ satisfies the RIP with a constant δ_k if for any k sparse vector $\boldsymbol{\alpha} \in \mathbb{R}^n$

$$(1-\delta_k)\|\boldsymbol{\alpha}\|_2^2 \leq \|\mathbf{M}\mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq (1+\delta_k)\|\boldsymbol{\alpha}\|_2^2 . \tag{2}$$

With this definition in hand it has been shown that if $\delta_{2k} \leq 1/4$ or $\delta_{3k} \leq 1/\sqrt{3}$ then IHT recovers the representation stably, i.e.,

$$\|\hat{\boldsymbol{\alpha}}_{\text{IHT}} - \boldsymbol{\alpha}_0\|_2 \leq c_{\text{IHT}} \|\mathbf{e}\|_2 , \tag{3}$$

where $c_{\text{IHT}} > 2$ is a function of δ_{2k} and δ_{3k} [3], [5], [6]. Note that with no prior on the noise distribution only a stable recovery is guaranteed with no noise reduction effect. The latter can be achieved by adding an assumption on the noise distribution [7]. This work deals only with the former case where $\mathbf{e}$ is an adversarial bounded noise.

Algorithm 1 Iterative hard thresholding (IHT)

Require: $k, \mathbf{M}, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{M}\mathbf{D}\boldsymbol{\alpha}_0 + \mathbf{e}$, k is the cardinality of $\boldsymbol{\alpha}_0$ and $\mathbf{e}$ is an additive noise.
Ensure: $\hat{\boldsymbol{\alpha}}_{\text{IHT}}$: k-sparse approximation of $\boldsymbol{\alpha}_0$.
Initialize representation $\hat{\boldsymbol{\alpha}}^0 = \mathbf{0}$ and set $t = 0$.
while halting criterion is not satisfied **do**
$t = t + 1$.
Perform a gradient step: $\boldsymbol{\alpha}_g = \hat{\boldsymbol{\alpha}}^{t-1} + \mu^t \mathbf{M}\mathbf{D}^*(\mathbf{y} - \mathbf{M}\mathbf{D}\hat{\boldsymbol{\alpha}}^{t-1})$
Find a new support: $T^t = \mathrm{supp}(\boldsymbol{\alpha}_g, k)$
Calculate a new representation: $\hat{\boldsymbol{\alpha}}^t = (\boldsymbol{\alpha}_g)_{T^t}$.
end while
Form the final solution $\hat{\boldsymbol{\alpha}}_{\text{IHT}} = \hat{\boldsymbol{\alpha}}^t$.

Note that in the case where $\mathbf{D}$ contains k correlated columns we have $\delta_k \geq 1$. Then the above recovery conditions fail and (3) does not hold. The reason for this is that in the presence of linear dependencies between a small group of columns from $\mathbf{D}$, the representation is no longer unique [8] and the solution of (1) is no longer stable [4]. Though the recovery of the representation is not achievable in the presence of correlations within $\mathbf{D}$, we should keep in mind that our task is to estimate the signal and not the representation. Recovering the wrong support of $\boldsymbol{\alpha}$, but one that is closely related to the original signal may suffice for our needs.

This key point is contained in the union of subspaces literature [9], [10], [11]. However, it has been pointed out more clearly in a series of contributions for the analysis framework [12], [13], [14], [15], [16], assuming a different sparse model. As such, correlations in the analysis dictionary were found to pose no problem and it has been demonstrated that such are even an advantage [14], [15], [16].

The analysis results serve as a clue that the same may happen in the synthesis model when the signal is the objective. In particular, the condition in [12] are presented in terms of the $\mathbf{D}$-RIP, which is a property of the measurement matrix $\mathbf{M}$ for the synthesis model. However, as indicated in [15], the results in [12] essentially hold true for signals emerging from the analysis model.

The work reported in [1] is very different from all the above, in addressing the synthesis model, providing signal recovery guarantees using the $\mathbf{D}$-RIP. This work presents a modified version of CoSaMP, Signal space CoSaMP (SSCoSaMP), that aims at recovering the signal, showing empirically that unlike the regular CoSaMP, the modified version gets a good recovery even in the presence of linear dependencies in $\mathbf{D}$. The authors of [1] use a similar proof technique to the one in [15] that was derived for the analysis CoSaMP (ACoSaMP). Just like [15], the work in [1] relies on the availability of near-optimal projection (this property will be defined clearly in the

next section). Another recent paper that exploits the $\mathbf{D}$-RIP in the context of the synthesis model is the one reported in [17], studying the basic synthesis ℓ_0-minimization problem.

In this work we continue with the same assumption as in [1] – the existence of a near optimal projection scheme[1] – and use the $\mathbf{D}$-RIP too. In Section II we present notations, and the definitions of the $\mathbf{D}$-RIP and the near-optimality of a projection. In Section III we introduce the signal space IHT (SSIHT) method for signal recovery and in Section IV we propose theoretical guarantees for it, relying on ideas taken from [15]. The SSIHT emerges from IHT as SSCoSaMP emerges from CoSaMP. The novelty of this work is by its theoretical study which relies on [15] and differs from [1]. Note that the proof technique used here can be adopted to develop new theoretical results for SSCoSaMP that differ from those in [1] and resemble those of ACoSaMP [15]. Section V presents some numerical results showing the advantage of SSIHT over IHT for the task of signal recovery.

II. Preliminaries

We start with the definition of the $\mathbf{D}$-RIP. As indicated in [12], many types of random matrices satisfy this property with a small δ_k[2].

Definition 2.1: A matrix $\mathbf{M}$ obeys the $\mathbf{D}$-RIP with a constant $\delta_k^{\mathbf{D}}$, if $\delta_k^{\mathbf{D}}$ is the smallest constant that satisfies

$$(1-\delta_k^{\mathbf{D}})\left\|\mathbf{z}\right\|_2^2 \leq \left\|\mathbf{M}\mathbf{z}\right\|_2^2 \leq (1+\delta_k^{\mathbf{D}})\left\|\mathbf{z}\right\|_2^2 \tag{4}$$

for any $\mathbf{z}\in\mathbb{R}^d$ such that $\mathbf{z}=\mathbf{D}\boldsymbol{\alpha}$ and $\left\|\boldsymbol{\alpha}\right\|_0 \leq k$.

Another definition we need is the one of a near optimal projection. In SSIHT we face the following problem: Given a general vector $\mathbf{z}\in\mathbb{R}^d$, we seek the closest vector to it, in the ℓ_2-norm sense, that has a k-sparse representation. Note that given a support set T, the closest vector is computed simply by using an orthogonal projection $\mathbf{P}_T=\mathbf{D}_T\mathbf{D}_T^{\dagger}$ onto it. Thus, the problem of finding the closest vector turns into the problem of finding its support, using the scheme

$$\mathcal{S}_k^*(\mathbf{z}) = \underset{T,|T|\leq k}{\operatorname{argmin}} \left\|\mathbf{z}-\mathbf{P}_T\mathbf{z}\right\|_2^2, \tag{5}$$

where the closest vector with k-sparse representation for $\mathbf{z}$ is simply given by $\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}$. We should remark that for the task of projecting a given representation vector to the same domain (k-sparse vectors), a simple hard thresholding as done in IHT gives the ideal solution. However, finding the optimal support in the signal case seems to be a NP-hard problem as its equivalent form in analysis context is known to be so [18]. Thus an approximation procedure is needed. For this purpose we introduce the definition of a near-optimal projection [15].

Definition 2.2: A procedure $\hat{\mathcal{S}}_k$ implies a near-optimal projection $\mathbf{P}_{\hat{\mathcal{S}}_k(\cdot)}$ with a constant C_k if for any $\mathbf{z}\in\mathbb{R}^d$

$$\left\|\mathbf{z}-\mathbf{P}_{\hat{\mathcal{S}}_k(\mathbf{z})}\mathbf{z}\right\|_2^2 \leq C_k\left\|\mathbf{z}-\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}\right\|_2^2. \tag{6}$$

In [1], a slightly different definition was used:

Definition 2.3: A procedure $\hat{\mathcal{S}}_k$ implies a near-optimal projection $\mathbf{P}_{\hat{\mathcal{S}}_k(\cdot)}$ with constants $C_{k,1}$ and $C_{k,2}$ if for any $\mathbf{z}\in\mathbb{R}^d$

$$\begin{aligned}&\left\|(\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}-\mathbf{P}_{\hat{\mathcal{S}}_k(\mathbf{z})})\mathbf{z}\right\|_2 \\ &\leq \min\left\{C_{k,1}\left\|\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}\right\|_2, C_{k,2}\left\|\mathbf{z}-\mathbf{P}_{\mathcal{S}_k^*(\mathbf{z})}\mathbf{z}\right\|_2\right\}.\end{aligned} \tag{7}$$

Having these definitions we recall the problem we aim at solving:

Definition 2.4 (Problem $\mathcal{P}$): Consider a measurement vector $\mathbf{y}\in\mathbb{R}^m$ such that $\mathbf{y}=\mathbf{M}\mathbf{x}_0+\mathbf{e}$ where $\mathbf{x}_0\in\mathbb{R}^d$ has a k-sparse representation under $\mathbf{D}$, $\mathbf{M}\in\mathbb{R}^{m\times d}$ is a degradation operator and $\mathbf{e}\in\mathbb{R}^m$ is a bounded additive noise. The largest singular value of $\mathbf{M}$ is $\sigma_{\mathbf{M}}$ and its $\mathbf{D}$-RIP constant is δ_k. The dictionary $\mathbf{D}\in\mathbb{R}^{p\times d}$ is given and fixed. A procedure $\hat{\mathcal{S}}_k$ is assumed to be available. Our task is to recover $\mathbf{x}_0$ from $\mathbf{y}$. The recovery result is denoted by $\hat{\mathbf{x}}$.

The following guarantee has been proposed in [1] for SSCoSaMP.

Theorem 2.5 (Theorem 2.1 in [1]): Consider the problem $\mathcal{P}$ and assume $\hat{\mathcal{S}}_k$ implies a near optimal projection with constants $C_{k,1}$ and $C_{k,2}$. After t iterations of SSCoSaMP, its signal estimate $\hat{\mathbf{x}}^t$ obeys

$$\left\|\hat{\mathbf{x}}^t-\mathbf{x}_0\right\|_2 \leq c_1\left\|\hat{\mathbf{x}}^{t-1}-\mathbf{x}_0\right\|_2 + c_2\left\|\mathbf{e}\right\|_2, \tag{8}$$

where $c_1=((2+C_{k,1})\delta_{4k}+C_{k,1})(2+C_{k,2})\sqrt{\frac{1+\delta_{4k}}{1-\delta_{4k}}}$ and $c_2=\frac{(2+C_{k,2})((2+C_{k,1})(1+\delta_{4k})+2)}{\sqrt{1-\delta_{4k}}}$.

Assuming $C_{k,1}=0.1$ and $C_{k,2}=1$ like in Corollary 2.1 in [1], a condition for $c_1<1$ is $\delta_{4k}<0.096$ which guarantees that after a finite number of iterations we have

$$\left\|\hat{\mathbf{x}}_{\text{SSCoSaMP}}-\mathbf{x}_0\right\|_2 \leq c_{\text{SSCoSaMP}}\left\|\mathbf{e}\right\|_2, \tag{9}$$

where c_{SSCoSaMP} is a function of c_1, c_2 and δ_{4k}. The bound in (9) implies a stable recovery of SSCoSaMP.

In this paper we show that under similar assumptions on the near optimality constant C_k of definition 2.2 and the maximal singular value of $\mathbf{M}$, $\sigma_{\mathbf{M}}$, the condition $\delta_{2k}<0.289$ guarantees a stable signal reconstruction for SSIHT. Note that in the condition of SSCoSaMP, two near optimality constants are involved. The second one is related to C_k as both of them measure the projection error and it is easy to show that they obey the inequality $(C_{k,2}-1)^2\leq C_k\leq(1+C_{k,2})^2$. The first constant $C_{k,1}$ measures the energy kept in the projection. This constant's relation to the other two depends on the initial norm of the projected signal. Since there is no direct relation between C_k and $C_{k,1}$, it is natural that another constant of the system would appear in our recovery conditions and indeed $\sigma_{\mathbf{M}}$ takes this role.

The existence of a general near-optimal projection scheme for any given dictionary is still an open problem and is left for future work. It is likely that there are non-trivial examples for which an efficient procedure exists as has been shown in [15] for the analysis case. In practice, any sparse recovery algorithm can be used in order to determine the support for the projection scheme. In this work we use a simple thresholding rule: For a given signal $\mathbf{z}$ it chooses the support to be the largest entries in $\mathbf{D}^*\mathbf{z}$. We show empirically that with this scheme we recover signals using SSIHT that cannot be recovered using the regular IHT. Note that thresholding does not have any known (near) optimality guarantee except for unitary operators.

III. Signal Space Iterative Hard Thresholding

SSIHT is presented in Algorithm 2. Its main difference from the regular IHT is the projection scheme. As IHT works in the representation domain, its projection is performed also there and as mentioned in the previous section, the projection is optimal in this case. For SSIHT that works in the signal domain no general projection procedure with an optimality guarantee is known.

The stopping criterion and the step size can be selected in the same way as in the regular IHT [19]. For the step size we consider three options: (i) Constant step-size selection $\mu^t=\mu$ in all iterations; (ii) Optimal changing step-size selection μ^t in each iteration by minimizing $\left\|\mathbf{y}-\mathbf{M}\hat{\mathbf{x}}^t\right\|_2$; and (iii) Adaptive changing step-size

[1] Our projection definition follows the one in [15], which is slightly different from the one used in [1].

[2] In this paper we shall use the brief notation δ_k to denote both RIP and $\mathbf{D}$-RIP, and the meaning should be understood from the context.

selection that has a closed-form solution and uses

$$\mu^t := \underset{\mu}{\operatorname{argmin}} \left\|\mathbf{y} - \mathbf{M}(\hat{\mathbf{x}}^{t-1} + \mu \mathbf{P}_{\tilde{T}}\mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}))\right\|_2^2, \quad (10)$$

where $\tilde{T} = \hat{T}^{t-1} \cup \hat{\mathcal{S}}_k(M^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}))$. More details appears in [15], [19]. In our theoretical study we analyze the first two options. In the experimental part we use the third one as it works better than the first, and approximates the second that has no closed-form solution.

Algorithm 2 Signal space iterative hard thresholding (SSIHT)

Require: $k, \mathbf{M}, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{M}\mathbf{D}\boldsymbol{\alpha}_0 + \mathbf{e}$, k is the cardinality of $\boldsymbol{\alpha}_0$ and $\mathbf{e}$ is an additive noise.

Ensure: $\hat{\mathbf{x}}_{\text{SSIHT}}$: k-sparse approximation of $\mathbf{x}_0$.

Initialize estimate $\hat{\mathbf{x}}^0 = \mathbf{0}$ and set $t = 0$.

while halting criterion is not satisfied **do**

$t = t + 1$.

Perform a gradient step: $\mathbf{x}_g = \hat{\mathbf{x}}^{t-1} + \mu^t \mathbf{M}^*(\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1})$

Find a new support: $T^t = \hat{\mathcal{S}}_k(\mathbf{x}_g)$

Project to get a new estimate: $\hat{\mathbf{x}}^t = \mathbf{D}_{T^t}\mathbf{D}^\dagger_{T^t}\mathbf{x}_g$.

end while

Form the final solution $\hat{\mathbf{x}}_{\text{SSIHT}} = \hat{\mathbf{x}}^t$.

IV. Algorithms Guarantees

A uniform guarantee for the idealized version of SSIHT that has an access to the optimal projection and uses a constant step size $\mu^t = \mu$, is presented in [11]. The work in [11] deals with a general union of subspaces, $\mathcal{A}$, where in our case $\mathcal{A} = \{\mathbf{x} | \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}, \|\boldsymbol{\alpha}\|_0 \leq k\}$. Using our notation Theorem 2 from [11] reads[3]:

Theorem 4.1 (Theorem 2 in [11]): Consider the problem $\mathcal{P}$ with $\hat{\mathcal{S}}_k = \mathcal{S}_k^*$ and apply SSIHT with a constant step size μ. If $1 + \delta_{2k} \leq \frac{1}{\mu} < 1.5(1 - \delta_{2k})$ then after a finite number of iterations t^*

$$\left\|\hat{\mathbf{x}}^{t^*} - \mathbf{x}_0\right\|_2 \leq c_3 \|\mathbf{e}\|_2, \quad (11)$$

where the constant c_3 is a function of δ_{2k} and μ.

In our work we extend the above in several ways: First, we refer to the case where an optimal projection is not known, and show that the same flavor guarantees apply for a near-optimal projection[4]. The price we seemingly have to pay is that $\sigma_{\mathbf{M}}$ enters the game. Second, we also consider the optimal step size and show that the same performance guarantees hold true in that case.

Theorem 4.2: Consider the problem $\mathcal{P}$ and apply SSIHT with a constant step size μ or an optimal changing step size. For any positive constant $\eta > 0$, let $b_1 := \frac{\eta}{1+\eta}$ and $b_2 := \frac{(C_k - 1)\sigma_{\mathbf{M}}^2 b_1^2}{C_k(1-\delta_{2k})}$. Suppose $\frac{b_2}{b_1^2} < 1$, $\frac{1}{\mu} \leq \sigma_{\mathbf{M}}^2$ and $1 + \delta_{2k} \leq \frac{1}{\mu} < \left(1 + \sqrt{1 - \frac{b_2}{b_1^2}}\right) b_1(1 - \delta_{2k})$. Then for $t \geq t^* \triangleq \frac{\log\left(\frac{\eta\|\mathbf{e}\|_2^2}{\|\mathbf{y}\|_2^2}\right)}{\log\left((1+\frac{1}{\eta})^2(\frac{1}{\mu(1-\delta_{2k})} - 1)C_k + (C_k - 1)(\mu\sigma_{\mathbf{M}}^2 - 1) + \frac{C_k}{\eta^2}\right)}$,[5]

$$\left\|\hat{\mathbf{x}}^t - \mathbf{x}_0\right\|_2^2 \leq \frac{(1+\eta)^2}{1 - \delta_{2k}} \|\mathbf{e}\|_2^2. \quad (12)$$

[3]Theorem 2 in [11] is more general and deals also with the case where $\hat{\mathcal{S}}_k$ is near-optimal up to an additive constant factor (in our definitions the factor is multiplicative). The error bound in the theorem has an additional constant factor that depends on the projection's near-optimality additive constant.

[4]Our work in fact improves the condition of the idealized case in [11] to be $\delta_{2k} \leq \frac{1}{3}$ instead of $\delta_{2k} \leq \frac{1}{5}$.

[5]For an optimal changing step-size the theorem conditions turn to be $\frac{b_2}{b_1^2} < 1$ and $1 + \delta_{2k} < (1 + \sqrt{1 - \frac{b_2}{b_1^2}}) b_1(1 - \delta_{2k})$ and we set $\mu = \frac{1}{1+\delta_{2k}}$ in t^*.

This theorem is a variant of Theorem 6.5 in [15] for AIHT and Theorem 2.1 in [20] for IHT. If, for example, $\sigma_{\mathbf{M}}^2 = 5$ and $C_k = 1.05$ then the conditions of Theorem 4.2 turn to be $\delta_{2k} \leq 0.289$ as mentioned before. For a better understanding of the nature of the theorem we refer the reader to the remarks after Theorems 6.2 and 6.5 in [15]. Briefly we comment on the selection of μ and η. For the step-size selection, note that an optimal changing step-size has the same theoretical guarantees as the optimal constant step-size $\mu = \frac{1}{1+\delta_{2k}}$. The advantage of the changing step-size method is that it does not need to compute (or estimate) the value of δ_{2k}. However, this comes at the cost of an additional complexity. Regarding the constant η, it gives a trade-off between satisfying the theorem conditions and the amplification of the noise. In particular, one may consider that the above theorem proves the convergence result for the noiseless case by taking η to infinity. This result is included in Lemma 4.4, which we present later, that guarantees in the case $\mathbf{e} = 0$ that $\mathbf{M}\hat{\mathbf{x}}^t$ converges geometrically to $\mathbf{M}\mathbf{x}_0$. Due to the uniqueness property that appears in [17], this implies that $\hat{\mathbf{x}}^t$ converges to $\mathbf{x}_0$.

We prove the theorem by presenting two key lemmas. The proofs rely on the ones in [15] that adopted ideas from [20] and [11]. Recall that the iterative algorithm tries to reduce the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ over iterations t. Thus, the progress of the algorithm can be indirectly measured by how much the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ is reduced at each iteration t. The two lemmas that we present capture this idea. The first lemma relates $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ to $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ and similar quantities at iteration $t-1$. We remark that the constraint $\frac{1}{\mu} \leq \sigma_{\mathbf{M}}^2$ in Theorem 4.2 may not be necessary and it is added only for having a simpler derivation of the results in this theorem. Furthermore, this is a very mild condition compared to $\frac{1}{\mu} < \left(1 + \sqrt{1 - \frac{b_2}{b_1^2}}\right) b_1(1 - \delta_{2k})$ and can only limit the range of values that can be used with the constant step size version of the algorithm.

Lemma 4.3: Consider the problem $\mathcal{P}$ and apply SSIHT with a constant step size μ satisfying $\frac{1}{\mu} \geq 1 + \delta_{2k}$ or an optimal step size[6]. Then, at the t-th iteration, the following holds:

$$\begin{aligned}\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 - \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 &\leq C_k \|\mathbf{y} - \mathbf{M}\mathbf{x}_0\|_2^2 \quad (13)\\ -C_k \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 &+ (C_k - 1)\mu\sigma_{\mathbf{M}}^2 \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2\\ +C_k \left(\frac{1}{\mu(1-\delta_{2k})} - 1\right) &\left\|\mathbf{M}(\hat{\mathbf{x}}^{t-1} - \mathbf{x}_0)\right\|_2^2.\end{aligned}$$

The proof of the above lemma is exactly the same as the proof of Lemma 6.6 in [15] with the change that here we use the $\mathbf{D}$-RIP instead of the $\boldsymbol{\Omega}$-RIP and the near-optimal projection scheme for synthesis instead of the one for analysis. The second lemma shows that once the objective $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ at iteration $t-1$ is small enough, then we are guaranteed to have small $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2$ as well. Given the presence of noise, this is quite natural; one cannot expect it to approach 0 but may expect it not to become worse. Moreover, the lemma also shows that if $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2$ is not small, then the objective in iteration t is necessarily reduced by a constant factor.

Lemma 4.4: Suppose that the same conditions of Theorem 4.2 holds true. If $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$, then $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$. Furthermore, if $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^{t-1}\right\|_2^2 > \eta^2 \|\mathbf{e}\|_2^2$, then

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq c_4 \left\|\mathbf{y} - \mathbf{M}\mathbf{x}^{t-1}\right\|_2^2, \quad (14)$$

where $c_4 < 1$ and

$$c_4 := \left(1 + \frac{1}{\eta}\right)^2 \left(\frac{1}{\mu(1-\delta_{2k})} - 1\right) C_k + (C_k - 1)(\mu\sigma_{\mathbf{M}}^2 - 1) + \frac{C_k}{\eta^2}.$$

[6]For an optimal step size the bound is achieved with the value $\mu = \frac{1}{1+\delta_{2k}}$.

The Lemma's proof is similar to the one of Lemma 6.7 in [15]. The needed adaptations are similar to those done for Lemma 4.3. Having the two lemmas above, the proof of the theorem is straightforward.

Proof of Theorem 4.2: Since $\hat{\mathbf{x}}^0 = \mathbf{0}$, $\|\mathbf{y}\|_2^2 = \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^0\right\|_2^2$. Assuming that $\|\mathbf{y}\|_2 > \eta \|\mathbf{e}\|_2$ and applying Lemma 4.4 repeatedly, we obtain $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \max(c_4^t \|\mathbf{y}\|_2^2, \eta^2 \|\mathbf{e}\|_2^2)$. Since $c_4^t \|\mathbf{y}\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2$ for $t \geq t^*$, we have

$$\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2^2 \leq \eta^2 \|\mathbf{e}\|_2^2 \tag{15}$$

for $t \geq t^*$. If $\left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^0\right\|_2 = \|\mathbf{y}\|_2 \leq \eta \|\mathbf{e}\|_2$ then according to Lemma 4.4, (15) holds for every $t > 0$. Finally, we observe

$$\left\|\hat{\mathbf{x}}^t - \mathbf{x}_0\right\|_2^2 \leq \frac{1}{1-\delta_{2k}} \left\|\mathbf{M}(\hat{\mathbf{x}}^t - \mathbf{x}_0)\right\|_2^2 \tag{16}$$

and by the triangle inequality,

$$\left\|\mathbf{M}(\hat{\mathbf{x}}^t - \mathbf{x}_0)\right\|_2 \leq \left\|\mathbf{y} - \mathbf{M}\hat{\mathbf{x}}^t\right\|_2 + \|\mathbf{e}\|_2 . \tag{17}$$

By plugging (15) into (17) and then the resulting inequality into (16), the claim of the Theorem follows. □

V. Numerical Performance

We turn to check numerically whether SSIHT can recover signals in scenarios where IHT cannot. We perform a synthetic test similar to the one in [17] for signals that are sparse under a dictionary which is highly coherent and with linear dependencies between its columns. We generate a dictionary $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2]$ where $\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{d \times d}$, $d = 200$, $\mathbf{D}_1$ contains sparse columns with 2 non-zero entries which are 1 or -1 with probability 0.5 and $\mathbf{D_2}$ contains columns which are linear combinations of random 3 columns from $\mathbf{D}_1$ with random zero-mean white Gaussian weights. Each entry of the measurement matrix $\mathbf{M} \in \mathbb{R}^{m \times d}$ is distributed according to a normal Gaussian distribution, where $m = \lfloor \gamma d \rfloor$ and γ is the sampling rate – a value in the range $(0, 1]$. We set k to be $\lfloor \rho m \rfloor$ ($\rho \ll 1$) and measure the recovery rate of the representation $\boldsymbol{\alpha}$ and the signal $\mathbf{x}$ for various values of $\gamma \in \{0.1, 0.2, \ldots, 0.9\}$ and $\rho \in \{0.01, 0.02, \ldots, 0.05\}$. We compare SSIHT also to SSCoSaMP, where both use projection by thresholding. The adaptive changing step-size selection rule is used for IHT and SSIHT. Similar to what is done in [15], by uniqueness conditions it is better to apply the algorithms with sparsity $\tilde{k} = \max(k, m/2)$.

Figure 1 presents the recovery performance over 100 realizations per each parameter setting. As expected, IHT fails almost always in recovering the signal since it focuses on the representation, while SSIHT and SSCoSaMP succeed in several cases and their performance are similar. At a first glance, some would think that the SSIHT phase diagram implies that for a fixed k/m (e.g. 0.03) one may improve the recovery result if he uses less samples, i.e. smaller m/d. However, this observation misses the fact that for a fixed k/m, k is reduced together with m. Note that the recovery results of SSIHT and SSCoSaP can be improved by using other techniques for the projection, rather than thresholding, as done in [1] for SSCoSaMP.

VI. Conclusion

In this paper we have proposed a variant of the IHT algorithm – the Signal-Space IHT (SSIHT) – for recovering signals with sparse representations under highly coherent dictionaries. We have shown that IHT fails in recovering such signals, as it operates in the representation domain. SSIHT, on the other hand, targets the signal. A uniform recovery guarantee has been derived for the SSIHT, assuming the availability of a near optimal projection. Numerical simulations show that SSIHT succeeds in recovering signals for which IHT fails, even when the projection is not near-optimal.

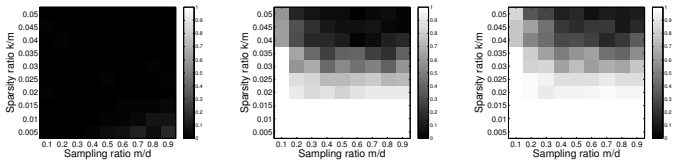

Fig. 1. IHT (left), SSIHT (middle) and SSCoSaMP (right) recovery rates for the synthetic experiment described in Section V. Color attribute: fraction of realizations in which a perfect recovery is achieved.

Acknowledgment

R. Giryes thanks the Azrieli Foundation for the Azrieli Fellowship. This research was supported by European Community's FP7- ERC program, grant agreement no. 320649.

References

[1] M. A. Davenport, D. Needell, and M. B. Wakin, "Signal space CoSaMP for sparse recovery with redundant dictionaries," *CoRR*, vol. abs/1208.0353, 2012.

[2] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[3] T. Blumensath and M. Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Anal*, vol. 27, no. 3, pp. 265 – 274, 2009.

[4] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory.*, vol. 52, no. 12, pp. 5406 –5425, Dec. 2006.

[5] S. Foucart, "Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants," in *Approximation Theory XIII*. Springer Proceedings in Mathematics, 2010, pp. 65–77.

[6] ——, "Hard thresholding pursuit: an algorithm for compressive sensing," *SIAM J. Numer. Anal.*, vol. 49, no. 6, pp. 2543–2563, 2011.

[7] R. Giryes and M. Elad, "RIP-based near-oracle performance guarantees for SP, CoSaMP, and IHT," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1465–1468, March 2012.

[8] D. L. Donoho and M. Elad, "Optimal sparse representation in general (nonorthogoinal) dictionaries via l1 minimization," *Proceedings of the National Academy of Science*, vol. 100, pp. 2197–2202, Mar 2003.

[9] Y. Lu and M. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2334 –2345, Jun. 2008.

[10] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inf. Theory.*, vol. 55, no. 4, pp. 1872 –1882, april 2009.

[11] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Trans. Inf. Theory.*, vol. 57, no. 7, pp. 4660–4671, 2011.

[12] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Anal*, vol. 31, no. 1, pp. 59 – 73, 2011.

[13] S. Nam, M. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 1, pp. 30 – 56, 2013.

[14] T. Peleg and M. Elad, "Performance guarantees of the thresholding algorithm for the cosparse analysis model," *IEEE Trans. on Information Theory*, vol. 59, no. 3, pp. 1832–1845, Mar. 2013.

[15] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies, "Greedy-like algorithms for the cosparse analysis model," *to appear in the Special Issue in Linear Algebra and its Applications on Sparse Approximate Solution of Linear Systems*, 2013.

[16] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Trans. on Signal Processing*, vol. 61, no. 3, pp. 661–677, Feb. 2013.

[17] R. Giryes and M. Elad, "Can we allow linear dependencies in the dictionary in the synthesis framework?" in *ICASSP 2013.*

[18] R. Gribonval, M. E. Pfetsch, and A. M. Tillmann, "Projection onto the k-cosparse set is NP-hard," *submitted to IEEE Trans. Inf. Theory.*, 2013.

[19] A. Kyrillidis and V. Cevher, "Recipes on hard thresholding methods," in *CAMSAP, 2011*, Dec. 2011, pp. 353 –356.

[20] R. Garg and R. Khandekar, "Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property," in *ICML '09*. New York, NY, USA: ACM, 2009, pp. 337–344.