Quantitative recovery conditions for tree-based compressed sensing

Coralia Cartis and Andrew Thompson*

***Abstract*—As shown in [1], [2], signals whose wavelet coefficients exhibit a rooted tree structure can be recovered using specially-adapted compressed sensing algorithms from just $n = \mathcal{O}(k)$ measurements, where k is the sparsity of the signal. Motivated by these results, we introduce a simplified proportional-dimensional asymptotic framework which enables the quantitative evaluation of recovery guarantees for tree-based compressed sensing. In the context of Gaussian matrices, we apply this framework to existing worst-case analysis of the Iterative Tree Projection (ITP) algorithm [1], [2] which makes use of the tree-based Restricted Isometry Property (RIP). Within the same framework, we then obtain quantitative results based on a new method of analysis, recently introduced in [3], which considers the fixed points of the algorithm. By exploiting the realistic average-case assumption that the measurements are statistically independent of the signal, we obtain significant quantitative improvements when compared to the tree-based RIP analysis. Our results have a refreshingly simple interpretation, explicitly determining a bound on the number of measurements that are required as a multiple of the sparsity. For example we prove that exact recovery of binary tree-based signals from noiseless Gaussian measurements is asymptotically guaranteed for ITP with constant stepsize provided $n \geq 50k$. All our results extend to the more realistic case in which measurements are corrupted by noise.**

***Index Terms*—Compressed sensing, recovery guarantees, wavelets, tree-based models, asymptotic analysis, Gaussian matrices.**

I. Introduction

Compressed sensing is motivated by the observation that many signals have an approximately sparse representation in some basis. Under this assumption, it has been proven that, to guarantee signal recovery, the sampling rate need only be proportional to the sparsity of the signal's approximation, rather than the signal dimension [4], [5]. Given an unknown signal x^* of dimension N, our aim is to recover x^* from $n < N$ undersampled linear measurements of the form $b = Ax^* + e$, where e is sampling noise. Many signals have additional structure that can be exploited in the recovery process, and one such example occurs when a wavelet basis is used to represent the signal. Wavelet representations are now widely used in a variety of signal processing contexts, most notably image processing, due to the fact that piecewise smooth signals have sparse representations in wavelet bases [6]. Wavelet representations have a multi-scale tree structure, in which signals are decomposed from coarse to fine scales, with the nested support properties of wavelets inducing a parent/child relationship between wavelet coefficients at different scales. One-dimensional wavelets, for example, have a binary tree structure, in which almost all coefficients have precisely two children. Section II-A gives a precise characterization of the tree structures we consider here.

C. Cartis is with the Mathematical Institute, University of Oxford, Oxford, UK (cartis@maths.ox.ac.uk).

A. Thompson is with the Mathematical Institute, University of Oxford, Oxford, UK (thompson@maths.ox.ac.uk).

Since wavelets essentially work as local discontinuity detectors, signal discontinuities give rise to a chain of large coefficients along a single branch [2]. For this reason, if a particular wavelet coefficient is large, its parent wavelet coefficient is also likely to be large, which means that the large wavelet coefficients of many signals can be modelled as forming a connected subset of the whole tree which is itself a *rooted tree*. This motivates an alternative model of data simplicity: assume that the image is supported on some rooted tree of cardinality k, for some sparsity parameter k.

Several algorithms have been proposed which approximately perform the Euclidean projection onto the set of vectors supported on a rooted tree of given cardinality [7], [8], [9]. Algorithms guaranteed to exactly calculate the projection were proposed in [10], [11]. Consequently, certain iterative projection algorithms for compressed sensing can be adapted to the tree-based setting. One such algorithm proposed in [2], and also in [1], is an adaptation of the well-known Iterative Hard Thresholding (IHT) algorithm [12], which we choose to call Iterative Tree Projection (ITP). Section II-B gives precise details on the ITP algorithm and associated stepsize variants.

ITP is one of several algorithms that have been proposed for the tree-based compressed sensing problem. Also relying upon tree projection, an adaptation of the CoSaMP algorithm [13] was proposed in [2]. Tree-based variants of matching pursuit algorithms were proposed in [14], [15]. Convex relaxations of the tree-based compressed sensing problem have also been considered [16], [17], [18], [19], [20].

Worst-case recovery guarantees for ITP (with exact tree projection) were obtained in [1], [2] in the case of binary trees, by extending the notion of the ubiquitous Restricted Isometry Property [21] to the tree-based setting. More recently, worst-case recovery guarantees based on tree-based RIP have been proved for approximate versions of ITP and tree-based CoSaMP in which the tree projections are computed to a given accuracy [22].

Bounds on tree-based RIP for random matrices with sub-gaussian entries were obtained in [1], [2] in terms of the ratio k/n. The bounds imply that it suffices to take only $n = C \cdot k$ measurements to guarantee recovery, for some implicitly quantified constant C. The value of the constant C is an issue of crucial importance to practitioners since it essentially determines how many measurements must be taken as a multiple of the signal sparsity. The main contribution of this paper is to determine explicit bounds on the constant C guaranteeing recovery. While our bounds are likely to be pessimistic compared to observed behaviour, they makes clear the extent of current theory in explicit quantitative terms. We obtain results in the context of one particular family of measurement matrices, the Gaussian ensemble, in which each entry of the matrix is i.i.d. Gaussian.

Since a Gaussian matrix is stochastic by nature, it is not possible to obtain deterministic results. However, by exploiting the remarkable concentration of measure properties of Gaussian matrices, it is possible to obtain limiting results as one

lets the matrix dimensions grow. In the context of simple sparsity, Donoho introduced a *proportional-dimensional asymptotic framework* as a way of quantifying results for recovery using l_1 minimization [23]. More precisely, let $(k, n, N) \to \infty$ such that $n/N \to \delta \in (0, 1]$ and $k/n \to \rho \in (0, 1]$, where δ is the undersampling ratio and ρ is the oversampling ratio. Following this framework, limiting results were obtained in [24] for three state-of-the-art greedy algorithms including IHT, the algorithm on which ITP is based. These results, which are worst-case in nature, make use of analysis in [25] which relies upon the RIP. More recently, by introducing a new method of analysis and by switching to an average-case framework, the present authors obtained improved quantitative results for IHT in [3].

We now describe the main contributions of this paper.

1) We introduce a simplified proportional growth asymptotic to enable quantitative comparison of recovery guarantees for tree-based compressed sensing. The aforementioned results from [1], [2] show that tree-based compressed sensing recovery depends only upon the ratio between n and k, and is independent of N, the signal dimension. This suggests that recovery results may be captured by a simplified asymptotic framework in which we dispense with the undersampling ratio δ and consider only the oversampling ratio ρ.

Definition I.1 (Simplified proportional-growth asymptotic) *We say that a sequence of problem sizes (k, n, N), where $0 < k \leq n \leq N$, obeys the simplified proportional-growth asymptotic if, for some $\rho \in (0, 1]$,*

$$\frac{k}{n} \to \rho \quad \text{as } (k, n, N) \to \infty.^{1}$$

While the commmon two-variable asymptotic framework leads to recovery phase transitions in the (δ, ρ)-plane, our recovery conditions take the refreshingly simple form of a threshold $\hat{\rho}$, such that stable recovery is asymptotically guaranteed provided the oversampling ratio satisfies $\rho < \hat{\rho}$. The framework allows a direct comparison of recovery conditions for different tree-based recovery algorithms, and for different methods of analysis.

A possible objection to our claim that our results are of practical relevance is that they are asymptotic in nature. However, our recovery results take the form of asymptotic bounds which hold for a sequence of increasing problem sizes, except with probability which decays exponentially in the problem dimension. We therefore believe that it is reasonable to expect recovery behaviour in practice to rapidly approach asymptotic limits, or be even better (since our asymptotic bounds are likely to be pessimistic).

2) We obtain explicit quantitative recovery guarantees for ITP algorithms with Gaussian measurement matrices in this simplified asymptotic framework. Our results are based upon a translation of the RIP analysis in [26] to the tree-based setting, and require the derivation of upper bounds on tree-based RIP constants for Gaussian matrices in the simplified proportional-growth asymptotic. We tighten the implicit bounds on tree-based RIP from [1] (see discussion in Section 3.3). We quantify oversampling thresholds for ITP and Gaussian matrices, the precise recovery values being dependent on the ITP stepsize scheme variant used (see Section II-B). In the case of zero noise, we have exact recovery of the original signal. In the case of noise, we derive *stability factors* which bound the approximation error of the output of ITP as a multiple of the noise level. The analysis in the present paper broadly follows the approach used to analyze IHT in [27], and deviates from it by tightening union bound arguments by exploiting the fact that only certain support sets (those corresponding to rooted trees) are permissible in the tree-based model. We compare our quantification with that obtainable from the existing analysis in [1], [2] for binary trees, demonstrating a dramatic improvement in the value of the constant.

3) We obtain improved quantitative recovery guarantees for ITP algorithms by exploiting average-case assumptions. We obtain results in the same framework based upon a translation of the *stable point* approach recently introduced by the present authors in [3] to the tree-based setting. Whereas the RIP is entirely worst-case, this alternative approach is more amenable to probabilistic analysis under the average-case (but realistic) assumption that the original signal and measurement matrix are statistically independent. Just as for the RIP analysis, the extension of the results in [3] involves the tightening of union bound arguments. The stable point condition is especially amenable to probabilistic analysis for Gaussian matrices under the average-case (but realistic) assumption that Central to the analysis are large deviations results for quantities related to Gaussian matrices, which are used to bound the constituent terms of the stable point condition, employing union bounds over all permissible support sets. We obtain oversampling thresholds for the same stepsize schemes, enabling a quantitative comparison with those derived from tree-based RIP analysis. For both stepsize schemes, the incorporation of average-case assumptions leads to a significant quantitative improvement in recovery guarantees for ITP and Gaussian matrices. We also extend our stable point recovery analysis to the case of noisy measurements, obtaining stability factors that show a substantial quantitative improvement over those derived from tree-based RIP analysis.

Outline of the paper. The rest of the paper is structured as follows: In Section II, we give full technical details of the tree-based compressed sensing problem, describe in more detail the generic ITP algorithm along with two possible stepsize schemes, and give a brief roadmap to the proofs. We describe our main results in Sections III and IV, first for those derived from tree-based RIP analysis (Section III), followed by the results derived from our stable point analysis (Section IV. A discussion of all our main results then follows in Section V. All proofs can be found in the appendix. We present the tree-based RIP analysis in Appendix A, and the stable point analysis in Appendix B. Both analyses rely crucially upon large deviations results for quantities related to Gaussian matrices (including bounds on tree-based RIP constants), and proofs of these subsidiary results can be found in Appendix D.

II. Problems and algorithms

A. Problem statement

Suppose we have a signal $y^* \in \mathbb{R}^N$ which has a sparse rooted-tree representation $x^* \in \mathbb{R}^N$ in some orthogonal wavelet basis, so that $x^* = \Psi y^*$ where $\Psi \in \mathbb{R}^{N \times N}$ is an orthogonal discrete wavelet transform matrix. We obtain the measurements $b = \Phi y^* + e \in \mathbb{R}^n$, where $\Phi \in \mathbb{R}^{n \times N}$, where e is sampling noise, and where we assume $n < N$. Referring to $A = \Phi \Psi^{-1} \in \mathbb{R}^{n \times N}$ from now on as the measurement matrix, we have

$$b = Ax^* + e. \tag{II.1}$$

We say that a vector x^* is *k-tree sparse* if it is supported on a rooted tree of cardinality k, and denote by $\mathcal{T}_k$ the set

[1] Note that the only restriction that the simplified proportional-growth asymptotic places upon N is that we must have $N \to \infty$ such that $N \geq n$.

of supports permitted by the model. Denoting by $\|\cdot\|$ the Euclidean norm $\|\cdot\|_2$, and defining

$$\Psi(x) := \frac{1}{2}\|b - Ax\|^2, \tag{II.2}$$

we can formulate signal recovery as the following optimization problem.

$$\min_{x\in\mathbb{R}^N} \Psi(x) \quad \text{subject to} \quad \text{supp}(x) \in \mathcal{T}_k, \tag{II.3}$$

where $\text{supp}(x)$ denotes the support of the signal x. We write $\mathcal{P}_k$ for the (exact) Euclidean projection onto the set $\{x : \text{supp}(x) \in \mathcal{T}_k\}$, namely

$$\mathcal{P}_k(z) := \underset{\text{supp}(x)\in\mathcal{T}_k}{\arg\min} \|x - z\|. \tag{II.4}$$

Our analysis will consider arbitrary tree structures, characterized only by the existence of a root coefficient (that is, a coefficient with no parents) a *tree order* d, defined to be the maximum number of children of any coefficient in the tree. We will at times refer to a tree of order d as a d-ary tree. The coefficients of one-dimensional wavelet transforms typically have a binary tree structure, that is tree order $d = 2$. The two-dimensional wavelet transforms often used in image processing typically form quad-trees ($d = 4$). Orthogonal discrete wavelet transforms often have a particular canonical tree structure, in which every coefficient essentially has the same number of children, but this condition is never enforced in our analysis.

Our challenge, then, is to recover the wavelet representation x^* (and therefore the original signal y^*) from the measurements (II.1), which we formally state as the following two problems.

Problem 1 *Recover exactly a k-tree sparse $x^* \in \mathbb{R}^N$ from the noiseless measurements $b = Ax^* \in \mathbb{R}^n$, where $k \leq n \leq N$.*

Problem 2 *Recover approximately a k-tree sparse $x^* \in \mathbb{R}^N$ from the noisy measurements $b = Ax^* + e \in \mathbb{R}^n$, where $k \leq n \leq N$.*

We consider the case where Φ is chosen to be a Gaussian matrix with entries distributed i.i.d. as $\{\Phi_{ij}\} \sim \mathcal{N}(0, 1/n)$. The orthogonality assumption on the wavelet transform Ψ then implies that the entries of A are also distributed i.i.d. as $\{A_{ij}\} \sim \mathcal{N}(0, 1/n)$, i.e. A is also i.i.d. Gaussian. Assuming Φ to be Gaussian is therefore equivalent to placing the same assumption on A, which we formalize as follows.

Assumption 1 *The measurement matrix A has i.i.d. $\mathcal{N}(0, 1/n)$ entries.*

It can be shown that x^* is the unique global solution to problem (II.3) whenever A is a Gaussian matrix [3, Sections 3 and 4.1].

Notation. Given some index set $\Gamma \subseteq \{1, 2, \dots N\}$, we define the complement of Γ to be $\Gamma^C = \{1, 2, \dots N\} \setminus \Gamma$. We write x_Γ for the restriction of the vector x to the coefficients indexed by the elements of Γ, and we write A_Γ for the restriction of the matrix A to those columns indexed by the elements of Γ.

B. ITP algorithms and stepsize schemes

In this section, we describe in more detail the ITP algorithm along with two possible stepsize schemes. Generically, on each iteration m, a steepest descent step, possibly with linesearch, is calculated for the objective Ψ in (II.3), namely, a move is performed from the current iterate x^m along the negative gradient of Ψ,

$$-\nabla\Psi(x^m) = -A^T(Ax^m - b).$$

Recalling the definition of $\mathcal{P}_k$ from Section II-A, the resulting step is then projected onto the (nonconvex) constraint in (II.3) which defines the set of all vectors supported on rooted trees of cardinality k.

Algorithm II.1 Generic ITP [1], [2]

Inputs: A, b, k.
Initialize $x^0 = 0;\ m = 0$.
While some termination criterion is not satisfied, do:
1) $x^{m+1} := \mathcal{P}_k\left\{x^m + \alpha^m A^T(b - Ax^m)\right\}$, where $\mathcal{P}_k(\cdot)$ is defined in (II.4) and $\alpha^m > 0$ is a stepsize.
2) $m := m + 1$

End; output $\hat{x} = x^m$.

To avoid a situation in which the support set Γ is not uniquely defined, if for instance some of the coefficients are equal in magnitude, then a support set for the identical components can be selected either randomly or according to some predefined ordering. In our analysis, we will consider the possibly infinite sequence of iterates generated by ITP, though in practice a useful termination criterion such as requiring the residual to be sufficiently small, would need to be employed.

Two stepsize choices will be addressed in this paper: *constant stepsize* $\alpha^m = \alpha \in (0, 1)$ for all m, which we will hereafter refer to simply as ITP [1], [2], and a *variable stepsize* scheme which we will call Normalised ITP (NITP), which adopts the same stepsize scheme as prescribed in the Normalised IHT variant of IHT algorithms proposed in [28]. The constant stepsize ITP variant can be summarized as follows.

Algorithm II.2 ITP [1], [2]

Given some $\alpha > 0$, on **step 1** of each iteration $m \geq 0$ of generic ITP, set

$$\alpha^m := \alpha. \tag{II.5}$$

The NITP variant defined below follows [28], having the stepsize α^m chosen according to an *exact linesearch* [29] when the support set of consecutive iterates stays the same, and using a *shrinkage* strategy when the support set changes, in order to ensure sufficient decrease in the objective of (II.3).

In practice, the choice of κ constitutes a trade-off between recovery performance and computational efficiency: for optimal performance, κ close to 1 should be chosen, while increasing κ will lead to fewer shrinkage steps, making the algorithm more computationally efficient. The shrinkage strategy ensures a potentially desirable property of the NITP algorithm, namely that, provided the measurement matrix satisfies mild linear independence assumptions, it is guaranteed to converge (see Section C2). A practical scheme similar to the one in [28] was proposed in [30] which does not employ a shrinkage strategy.

An important property of the operator $\mathcal{P}_k$ is that it preserves

Algorithm II.3 NITP

Given some $c \in (0,1)$ and $\kappa > 1/(1-c)$, on **step 1** of each iteration $m \geq 0$ of generic ITP, do:

1.1. **Exact linesearch.**

(a) $\Gamma^m := \text{supp}(x^m)$.

(b)
$$\alpha^m := \frac{\|A_{\Gamma^m}^T(b - Ax^m)\|^2}{\|A_{\Gamma^m}A_{\Gamma^m}^T(b-Ax^m)\|^2}. \tag{II.6}$$

(c) $\tilde{x}^{m+1} := \mathcal{P}_k\left\{x^m + \alpha^m A^T(b - Ax^m)\right\}$.

1.2. **Backtracking.** If $\text{supp}(\tilde{x}^{m+1}) = \text{supp}(x^m)$, end; output α^m.

Else, while $\alpha^m \geq (1-c)\frac{\|\tilde{x}^{m+1}-x^m\|^2}{\|A(\tilde{x}^{m+1}-x^m)\|^2}$, do:

(a) $\alpha^m := \alpha^m/(\kappa(1-c))$.

(b) $\tilde{x}^{m+1} := \mathcal{P}_k\left\{x^m + \alpha^m A^T(b - Ax^m)\right\}$.

End; output α^m.

the value of selected coefficients.

$$\{\mathcal{P}_k(x)\}_i := \begin{cases} x_i & i \in \Gamma \\ 0 & i \notin \Gamma \end{cases} \quad \text{where } \Gamma := \underset{\Gamma \in \mathcal{T}_k}{\arg\max} \ \|x_\Gamma\|. \tag{II.7}$$

See [27, Lemma 6.1] for a proof of (II.7) given its definition. It follows from (II.7) that $\mathcal{P}_k$ can be framed as an integer program with $\{0,1\}$ decision variables. This problem can either be solved exactly using dynamic programming [10] or approximately by solving its linear programming or Lagrangian relaxations [7], [31]. We refer the reader to [10] for further details on methods for performing the projection onto rooted trees.

III. Recovery results for tree-based RIP analysis

A. Results for deterministic matrices

Our first analysis relies upon a deterministic recovery condition originally given in [26]. Our contribution is to extend it to the tree-based setting and then obtain from it quantitative results for Gaussian matrices. We consider an extension of the ubiquitous (asymmetric) Restricted Isometry Property (RIP) [21], [32] to the tree-based setting.

Definition III.1 (Tree-based RIP [1], [2]) *For a given matrix A, define TL_s and TU_s, the lower and upper tree-based RIP constants of order s, to be, respectively,*

$$\begin{aligned} TL_s &:= 1 - \min_{\emptyset \neq \textit{supp}(y) \subseteq \Gamma \in \mathcal{T}_s} \frac{\|Ay\|^2}{\|y\|^2}; \\ TU_s &:= \max_{\emptyset \neq \textit{supp}(y) \subseteq \Gamma \in \mathcal{T}_s} \frac{\|Ay\|^2}{\|y\|^2} - 1. \end{aligned} \tag{III.8}$$

We obtain deterministic recovery results of the following form for both ITP and NITP.

Theorem III.2 *Consider Problem 2. Let μ^{ALG} and ξ^{ALG} be defined as in Definition III.3. Then, there exists functions μ^{ALG} and ξ^{ALG} such that, provided $\mu^{ALG} < 1$, the output, $\hat{x}$, at iteration m of variant ALG of ITP satisfies*

$$\|\hat{x} - x^*\| \leq \left(\mu^{ALG}\right)^m \|x^*\| + \frac{\xi^{ALG}}{1-\mu^{ALG}}\|e\|. \tag{III.9}$$

Proof: See Appendix A. □

The functions μ^{ALG} and ξ^{ALG} will play the role of a convergence factor and a factor controlling stability to noise. Though Theorem III.2 gives a limiting bound on the approximation error, it does not necessarily imply convergence of the algorithm. In the simplified noiseless case however, the result implies convergence to x^* at a linear rate.

Specifically, Theorem III.2 holds for ITP with stepsize α if $\mu^{ALG} := \mu^{ITP_\alpha}$ and $\xi^{ALG} := \xi^{ITP_\alpha}$, while Theorem III.2 holds for NITP with shrinkage parameter κ if $\mu^{ALG} := \mu^{NITP_\kappa}$ and $\xi^{ALG} := \xi^{NITP_\kappa}$, defined as follows.

Definition III.3 *Provided $3k \leq n$, define*

$$\mu^{ITP_\alpha} := \sqrt{3}\max\{\alpha(1+TU_{3k}) - 1, 1 - \alpha(1 - TL_{3k})\}, \tag{III.10}$$

$$\xi^{ITP_\alpha} := \alpha\sqrt{3(1+TU_{2k})}, \tag{III.11}$$

$$\mu^{NITP_\kappa} := \sqrt{3}\max\left\{\frac{1+TU_{3k}}{1-TL_k} - 1, 1 - \frac{1-TL_{3k}}{\kappa[1+TU_{2k}]}\right\}, \tag{III.12}$$

$$\xi^{NITP_\kappa} := \frac{\sqrt{3(1+TU_{2k})}}{1-TL_k}, \tag{III.13}$$

where TU and TL are defined in Definition III.1.

B. Asymptotic results for Gaussian matrices

We derive quantitative recovery conditions for Gaussian matrices by means of upper bounds on tree-based RIP constants in the simplified proportional-growth asymptotic of Definition I.1. We follow the broad approach used for the standard notion of RIP in [32], [33], [3], in which a union bound was performed over the maximum/minimum singular values of all $\binom{N}{k}$ submatrices of A of size $n \times k$. In the present work, however, the assumed tree structure means that the number of permissible support sets for iterates of the algorithm is much diminished, which means that union bound arguments can be tightened, leading to improved quantitative results.

The number, $|\mathcal{T}_k|$, of permissible support sets in the d-ary tree-based framework, is bounded above by $T(k)$, the total number of ordered, rooted d-ary trees of cardinality k. Fortunately, a formula for $T(k)$ is known.

Lemma III.4 (Tree counting result [34]) *The total number of ordered, rooted d-ary trees of cardinality k is*

$$T(k) = \frac{1}{(d-1)k+1}\binom{dk}{k}. \tag{III.14}$$

In particular, note that $T(k)$ depends only upon the tree order d and the sparsity k, and not upon the signal length N. It is for this reason that we are able to obtain quantitative bounds in the simplified proportional-growth asymptotic, i.e. in terms of d and the variable $\rho := \lim_{n\to\infty}\frac{k}{n}$ only.

Before defining bounds, it will be useful to define the Shannon entropy in the usual way.

Definition III.5 (Shannon entropy [32]) *Given $p \in (0,1)$, define the Shannon entropy with base e logarithms as*

$$H(p) := -p\ln(p) - (1-p)\ln(1-p). \tag{III.15}$$

We define the following bounds on tree-based RIP constants for Gaussian matrices.

Definition III.6 (Tree-based RIP bounds) *Define, for $\rho \in (0,1)$ and $\lambda > 0$,*

$$\psi_{max}(\lambda,\rho) = \frac{1}{2}\left[(1+\rho)\ln\lambda + 1 + \rho - \rho\ln\rho - \lambda\right] \tag{III.16}$$

and

$$\psi_{min}(\lambda,\rho) = H(\rho) + \frac{1}{2}\left[(1-\rho)\ln\lambda + 1 - \rho + \rho\ln\rho - \lambda\right], \tag{III.17}$$

where $H(\cdot)$ is defined in (III.15). Define $\lambda^{max}(\rho)$ and $\lambda^{min}(\rho)$ as the unique solution to (III.18) and (III.19) respectively:

$$\psi_{max}\Big(\lambda^{max}(\rho),\rho\Big) + d\rho\cdot H(d^{-1}) = 0 \;\; for \;\; \lambda^{max}(\rho) > 1+\rho; \tag{III.18}$$

$$\psi_{min}\left(\lambda^{min}(\rho),\rho\right) + d\rho\cdot H(d^{-1}) = 0 \;\; for \;\; \lambda^{min}(\rho) < 1-\rho, \tag{III.19}$$

and define $\mathcal{TU}(\rho) = \lambda^{max}(\rho) - 1$ and $\mathcal{TL}(\rho) = 1 - \lambda^{min}(\rho)$.

That there exists a unique solution to (III.18) follows since $\psi_{max}[\lambda,\rho]$ is positive for $\lambda = 1+\rho$, tends to $-\infty$ as $\lambda\to\infty$, and is strictly decreasing in λ. Similarly, that there exists a unique solution to (III.19) follows since $\psi_{min}[\lambda,\rho]$ is positive for $\lambda = 1-\rho$, tends to $-\infty$ as $\lambda\to\infty$, and is strictly decreasing in λ.

Intuition behind the form of the bounds given in Definition III.6 is as follows. For a given $n\times k$ submatrix A_Γ, the asymptotic distributions of λ^{max} and λ^{min}, the extreme eigenvalues of its corresponding Gram matrix $A_\Gamma^T A_\Gamma$, depend asymptotically upon ρ, and both decay exponentially away from 1, with exponents given by $\gamma_{max}(\lambda^{max}(\rho),\rho)$ and $\gamma_{min}(\lambda^{min}(\rho),\rho)$ respectively. To bound the extreme eigenvalues of all possible such Gram matrices requires a union bound over the number of permissible support sets. For the standard notion of RIP analyzed in [32], all $\binom{N}{k}$ support sets must be considered, which leads to an exponent which depends upon ρ and also $\delta := \lim_{n\to\infty} n/N$. In the tree-based setting, however, the number of permissible support sets is given by (III.14), which has no dependence upon the ambient dimension N, and the resulting exponent $d\rho\cdot H(d^{-1})$ depends only upon ρ (for a given tree order d). The asymptotic bounds $\mathcal{TU}(\rho)$ and $\mathcal{TL}(\rho)$ are defined in such a way that they are satisfied in the asymptotic limit when the net exponents in (III.18) and (III.19) respectively are negative.

Counterparts of the bounds in Definition III.6 for the standard notion of asymmetric RIP constants were shown to hold asymptotically for Gaussian matrices in [32]. Following their method of proof, we obtain an analogous result for tree-based RIP constants in the simplified proportional-growth asymptotic.

Lemma III.7 (Validity of tree-based RIP bounds) *Suppose Assumption 1 holds and let $\epsilon > 0$. In the simplified proportional-growth asymptotic,*

$$\mathbb{P}\left(TU_k \geq \mathcal{TU}(\rho) + \epsilon\right) \to 0, \tag{III.20}$$

$$\mathbb{P}\left(TL_k \leq \mathcal{TL}(\rho) - \epsilon\right) \to 0, \tag{III.21}$$

both exponentially in n.

Proof: See Appendix D.

Closely following the approach in [24], we show that a naive replacement of each TL_{pk} and TU_{qk} by the tree-based RIP bounds $\mathcal{TL}(p\rho)$ and $\mathcal{TU}(q\rho)$ is valid, provided the functions $\mu_{RIP}^{ITP_\alpha}$ and $\xi_{RIP}^{ITP_\alpha}$ satisfy certain properties given in Appendix B. We finally arrive at asymptotic recovery results of the following form for both variants of ITP and Gaussian matrices.

Theorem III.8 (RIP-based recovery) *Consider Problem 2 and suppose Assumption 1 holds. Define $\hat{\rho}_{RIP}^{ALG}$ as the unique solution to $\mu_{RIP}^{ALG}(\rho) = 1$. Choose $\epsilon\in(0,1)$ and suppose that*

$$\rho < (1-\epsilon)\hat{\rho}_{RIP}^{ALG}. \tag{III.22}$$

Suppose $\hat{x}$ is the output of variant ALG of ITP at iteration m. Then

$$\mu_{RIP}^{ALG}((1+\epsilon)\rho) < 1, \tag{III.23}$$

and, in the simplified proportional-growth asymptotic[2],

$$\|\hat{x}-x^*\| \leq \left(\mu_{RIP}^{ALG}((1+\epsilon)\rho)\right)^m \|x^*\| + \frac{\xi_{RIP}^{ALG}((1+\epsilon)\rho)}{1-\mu_{RIP}^{ALG}((1+\epsilon)\rho)}\|e\|, \tag{III.24}$$

for all k-tree sparse vectors x^*, with probability tending to 1 exponentially in n.

Proof: See Appendix A. □

In the idealized case of zero measurement noise, we can deduce from Theorem III.8 guaranteed convergence of ITP variants at a linear rate.

Corollary III.9 (RIP-based recovery: noiseless case) *Consider Problem 1 and suppose Assumption 1 holds. Choose $\epsilon\in(0,1)$ and suppose that (III.22) holds, where $\hat{\rho}_{RIP}^{ALG}$ and $\mu_{RIP}^{ALG}(\rho)$ are defined as in Theorem III.8. Then, in the simplified proportional-growth asymptotic, the iterates of variant ALG of ITP converge to x^* at a linear rate, for all k-tree sparse vectors x^*, with probability tending to 1 exponentially in n.*

Proof: See Appendix A. □

Specifically, Theorem III.8 and Corollary III.9 hold for ITP with stepsize α if $\mu_{RIP}^{ALG}(\rho) := \mu_{RIP}^{ITP_\alpha}(\rho)$ and $\xi^{ALG} := \xi_{RIP}^{ITP_\alpha}(\rho)$, while Theorem III.8 and Corollary III.9 hold for NITP with shrinkage parameter κ if $\mu_{RIP}^{ALG}(\rho) := \mu_{RIP}^{NITP_\kappa}(\rho)$ and $\xi^{ALG} := \xi_{RIP}^{NITP_\kappa}(\rho)$, defined as follows (compare with Definition III.3).

Definition III.10 (Convergence and stability factors) *Define, for $\rho\in(0,1/3)$,*

$$\mu_{RIP}^{ITP_\alpha}(\rho) := \sqrt{3}\max\{\alpha[1+\mathcal{TU}(3\rho)]-1, 1-\alpha[1-\mathcal{TL}(3\rho)]\}, \tag{III.25}$$

$$\xi_{RIP}^{ITP_\alpha}(\rho) := \alpha\sqrt{3[1+\mathcal{TU}(2\rho)]}, \tag{III.26}$$

$$\mu_{RIP}^{NITP_\kappa}(\rho) := \sqrt{3}\max\left\{\frac{1+\mathcal{TU}(3\rho)}{1-\mathcal{TL}(\rho)} - 1, 1 - \frac{1-\mathcal{TL}(3\rho)}{\kappa[1+\mathcal{TU}(2\rho)]}\right\}, \tag{III.27}$$

$$\xi_{RIP}^{NITP_\kappa}(\rho) := \frac{\sqrt{3[1+\mathcal{TU}(2\rho)]}}{1-\mathcal{TL}(\rho)}, \tag{III.28}$$

where $\mathcal{TU}$ and $\mathcal{TL}$ are given in Definition III.6.

In the case of ITP with constant stepsize, Theorem III.8 and Corollary III.9 give a continuous range of oversampling thresholds for any $0<\alpha<2$. For $\alpha\geq 2$, the result gives $\hat{\rho}_{RIP}^{IHT_\alpha} = 0$ for all $\delta\in(0,1)$. It is clear that $\mu_{RIP}^{IHT_\alpha}(\rho)$ takes its minimum value when the two expressions inside the maximum in (III.25) are equal, which implies that the optimal

[2]In other words, we consider instances of the Gaussian random variables A for a sequence of triples (k,n,N) where $n\to\infty$, where n is the number of measurements, N, the signal dimension and k, the sparsity of the underlying signal.

oversampling threshold is obtained when the stepsize is taken to be

$$\hat{\alpha} := 2/[2 + \mathcal{TU}(3\rho) - \mathcal{TL}(3\rho)]. \tag{III.29}$$

We will adopt the optimal stepsize choice $\hat{\alpha}$ in all our numerical computations of oversampling thresholds[3].

C. Prior bounds on tree-based RIP

A result quantifying tree-based RIP for Gaussian matrices was proved in [1] as a special case of a more general result on restricted isometry constants for subgaussian random matrices and signals drawn from a union of linear subspaces. A symmetric notion of tree-based RIP was considered, in which no distinction is made between the upper and lower tails. For a given measurement matrix, the symmetric tree-based RIP constant TR_k is thus

$$TR_k := \max(TL_k, TU_k). \tag{III.30}$$

Theorem III.11 ([1, Corollary 4.2]) *Suppose Assumption 1 holds and choose $t > 0$ and let the tree order be $d = 2$. Then, with probability at least $1 - e^{-t}$, $TR_k \leq r$ provided*

$$n \geq \left(\frac{r^2}{144} - \frac{r^3}{1296}\right)^{-1} \left[k\left(1 + \ln\frac{72}{r}\right) - \ln\left(\frac{k+1}{2}\right) + t\right]$$

We may deduce from this result a bound on (symmetrical) tree-based RIP, $\mathcal{TR}(\rho)$, within the simplified proportional-growth asymptotic. The resulting bound is plotted in Figure 1 alongside the bounds $\mathcal{TU}(\rho)$ and $\mathcal{TL}(\rho)$ of Definition III.6.

Definition III.12 *If $\rho \in (0, 0.024)$[4], define $\mathcal{TR}(\rho)$ to be the unique solution in $r > 0$ to*

$$r^2(9 - r) = 1296\rho\left[1 + \ln\left(\frac{72}{r}\right)\right]. \tag{III.31}$$

To make the quantification of recovery results for ITP algorithms explicit, one may combine the bound $\mathcal{TR}(\rho)$ with the symmetric versions of the RIP-based recovery results for each variant. For ITP, the condition $TR_{3k} < 1/\sqrt{3}$ was proved in [26]. For NIHT, the condition $TR_{3k} < (11 - \sqrt{3})/(11 + 21\sqrt{3}) \approx 0.1956$ follows by combining (III.30) with Theorem III.2 and Definition III.3, taking $\kappa := 1.1$. A quantitative comparison in the case of binary trees between the recovery conditions obtainable from this prior analysis and those presented in Sections III-A and III-B is given in Section V-A.

[3]Note that this optimal stepsize $\hat{\alpha}$ is closely related to the (constant) maximal stepsize in gradient methods for strongly convex optimization that ensures global linear rate of convergence (see [35, Theorem 2.1.15]). In particular, the objective (II.2) restricted to a face of $\mathcal{T}_k$ is a strongly convex objective and ITP is taking a steepest descent step on this face, scaled by $\hat{\alpha}$. Then $\mu := 1 - \mathcal{TL}(3\rho)$ can be regarded as a lower bound on the smallest eigenvalue of the reduced Hessian of the objective (II.2) and $L := 1 + \mathcal{TU}(3\rho)$ as an upper bound on the largest eigenvalue of the same matrix. With this correspondence, the constant step sizes prescribed by (III.29) and the maximal one in [35, Theorem 2.1.15] coincide; see [36] for full details and similar analogies.

[4]Elementary calculus shows that (III.31) only has a solution for ρ below approximately 0.02407 for $d = 2$ (binary trees).

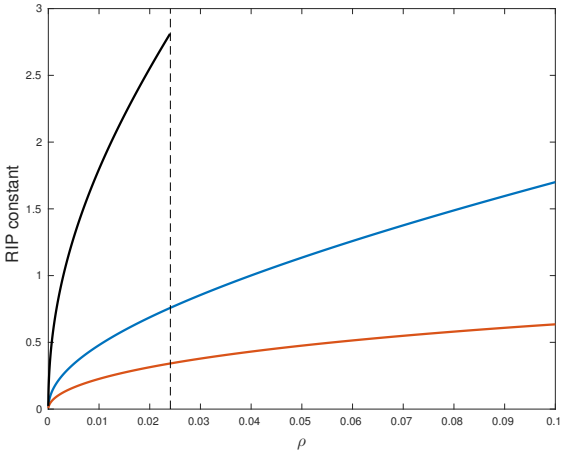

Fig. 1. A comparison of $\mathcal{TL}(\rho)$ (red), $\mathcal{TU}(\rho)$ (blue) and $\mathcal{TR}(\rho)$ (black) where defined, for $\rho \in (0, 0.1)$.

IV. Recovery results using a tree-based stable point analysis

Our second analysis, which broadly follows the approach used to analyze IHT in [3], considers the *stable points* of ITP, a concept which can be viewed as a generalization of the notion of a fixed point to accommodate variable stepsize schemes, see [3, Section 3.1].

Definition IV.1 (Stable points of generic ITP) *Given $\underline{\alpha} > 0$ and an index set $\Gamma \in \mathcal{T}_k$, we say $\bar{x} \in \mathbb{R}^N$ is an $\underline{\alpha}$-stable point of generic ITP on Γ if supp$(\bar{x}) \subseteq \Gamma$ and*

$$\left\{A^T(b - A\bar{x})\right\}_\Gamma = 0 \quad \textit{and} \tag{IV.32}$$

$$\|\bar{x}_{\Gamma\setminus\Omega}\| \geq \underline{\alpha}\|A^T_{\Omega\setminus\Gamma}(b - A\bar{x})\| \quad \forall\, \Omega \in \mathcal{T}_k. \tag{IV.33}$$

For brevity's sake, we will often drop the 'of generic ITP' label, and at times we will also drop the reference to the support set Γ. We will be interested in values of $\underline{\alpha}$ that lower bound the stepsize α^m of generic ITP.

A. Results for ITP

First considering ITP with constant stepsize α, our approach is two-stage: on the one hand, we give conditions guaranteeing convergence of ITP to *some* stable point. Meanwhile, by analysing a necessary condition for the existence of a stable point on a given support (which we refer to as the *stable point condition*), we give conditions guaranteeing that all stable points are 'close' to the original signal. Thus, if both conditions are satisfied, we ensure recovery of the original signal.

We will require the following assumption for the deterministic results given in this section.

Assumption 2 *The columns of A are in $2k$-general position, namely any collection of $2k$ of its columns are linearly independent.*

Assumption 2 is a typical (weak) assumption in compressed sensing, and which guarantees a unique solution to Problem 1. We denote by $A_\Gamma^\dagger$ the Moore-Penrose pseudoinverse $(A_\Gamma^T A_\Gamma)^{-1} A_\Gamma^T$, which is well-defined under Assumption 2. We next state a necessary condition for a stable point on a given support Γ in terms of only x^*, A and e and their restrictions to certain support sets.

We next give a deterministic condition guaranteeing convergence to some α-stable point in terms of the tree-based RIP.

Theorem IV.2 (ITP convergence) *Consider Problem 2. Suppose that Assumption 2 holds, and suppose that the stepsize in ITP satisfies*

$$\alpha < \frac{1}{1+TU_{2k}}, \tag{IV.34}$$

where TU is defined in (III.1). Then ITP with stepsize α converges to an α-stable point $\bar{x}$ of generic ITP.

We next state the *stable point condition*, that is, a necessary condition for the existence of a stable point on a given support. It will help to first define $\Lambda \in \mathcal{T}_k$ to be the support of the original signal, namely

$$\Lambda := \mathrm{supp}(x^*), \tag{IV.35}$$

so that $|\Lambda| = k$.

Theorem IV.3 (Stable point condition) *Consider Problem 2. Suppose Assumption 2 holds and suppose there exists an $\underline{\alpha}$-stable point on some Γ such that $\Gamma \neq \Lambda$. Then*

$$\begin{aligned}&\left\|A_\Gamma^\dagger A_{\Lambda\setminus\Gamma} x^*_{\Lambda\setminus\Gamma}\right\| + \left\|A_\Gamma^\dagger e\right\| \geq \\ &\underline{\alpha}\left\{\left\|A^T_{\Lambda\setminus\Gamma}(I - A_\Gamma A_\Gamma^\dagger)A_{\Lambda\setminus\Gamma} x^*_{\Lambda\setminus\Gamma}\right\| - \left\|A^T_{\Lambda\setminus\Gamma}(I - A_\Gamma A_\Gamma^\dagger)e\right\|\right\}\end{aligned} \tag{IV.36}$$

where Λ is defined in (IV.35).

Proof: See Appendix C. □

While it would be possible to analyse the stable point condition using the tree-based RIP, we take a different approach. The stable point condition is especially amenable to probabilistic analysis for Gaussian matrices under the average-case (but realistic) assumption that the original signal and measurement matrix are statistically independent.

Assumption 3 *The original signal x^* and the measurement matrix A are statistically independent.*

The crucial independence assumption will allow us to obtain better quantitative results than could be achieved through the purely worst-case RIP-based analysis of Section III. However, it is worth noting that independence is the only average-case assumption we invoke: we assume nothing further about the coefficient values of x^*. In keeping with the spirit of average-case analysis, we also assume that the noise is Gaussian and independent of both A and x^*, which we formalize as follows.

Assumption 4 *The noise vector e has i.i.d. Gaussian entries $e_i \sim N(0, \sigma^2/n)$, independently of A and x^*.*

Note that, under Assumption 4, $\mathbb{E}\|e\|^2 = \sigma^2$, so that $\|e\| \approx \sigma$.

Assumption 2 is satisfied with probability 1 by a Gaussian matrix, see [3, Section 4.1], and so may now be replaced with Assumption 1.

Under Assumptions 1, 3 and 4, each of the terms in (IV.36), viewed as a Rayleigh quotient over $\|x_{\Lambda\setminus\Gamma}\|^2$, is distributed according to either the χ^2 or the $\mathcal{F}$ distribution. We write χ^2_s for the (univariate) χ^2-distribution with $s \geq 1$ degrees of freedom. Furthermore, if $P \sim \frac{1}{s}\chi^2_s$ and $Q \sim \frac{1}{t}\chi^2_t$ are independent random variables, we say that P/Q follows the $\mathcal{F}$-distribution, and we write $P/Q \sim \mathcal{F}(s,t)$. The following lemma, which was proved in [3], gives the precise distributions.

Lemma IV.4 (Distribution results [3, Lemma 4.4]) *Suppose Assumptions 1, 3 and 4 hold, and let Γ be an index set of cardinality k, where $k < n$. Then*

$$\frac{\|A_\Gamma^\dagger A_{\Lambda\setminus\Gamma} x_{\Lambda\setminus\Gamma}\|^2}{\|x_{\Lambda\setminus\Gamma}\|^2} = F_\Gamma, \quad \text{where} \quad F_\Gamma \sim \frac{k}{n-k+1}\mathcal{F}(k, n-k+1); \tag{IV.37}$$

$$\frac{\|A^T_{\Lambda\setminus\Gamma}(I - A_\Gamma A_\Gamma^\dagger)A_{\Lambda\setminus\Gamma} x_{\Lambda\setminus\Gamma}\|^2}{\|x_{\Lambda\setminus\Gamma}\|^2} \geq \left(\frac{n-k}{n}\right)^2 \cdot R^2_\Gamma, \quad \text{where} \quad R_\Gamma \sim \frac{1}{n-k}\chi^2_{n-k}; \tag{IV.38}$$

$$\|A_\Gamma^\dagger e\| \leq \sigma\cdot\sqrt{G_\Gamma}, \quad \text{where} \quad G_\Gamma \sim \frac{k}{n-k+1}\mathcal{F}(k, n-k+1); \tag{IV.39}$$

$$\|A^T_{\Lambda\setminus\Gamma}(I - A_\Gamma A_\Gamma^\dagger)e\| \leq \sigma\sqrt{\frac{k(n-k)}{n^2}\cdot(S_\Gamma)(T_\Gamma)}, \quad \text{where} \quad S_\Gamma \sim \frac{1}{n-k}\chi^2_{n-k}, \quad T_\Gamma \sim \frac{1}{k}\chi^2_k. \tag{IV.40}$$

Recalling the stable point condition, we wish to show that all stable points are 'close' to the original signal, which can be achieved by bounding each of the constituent terms over all permissible support sets. We can make an analogy with the tree-based RIP, where upper bounds on tree-based RIP constants are obtained in the simplified porportional-growth asymptotic by union bounding the tail probabilities of extreme singular values of submatrices of A corresponding to permissible support sets. Similarly, large deviation bounds over $|\mathcal{T}_k|$ instances of χ^2 and $\mathcal{F}$ distributed random variables can be derived in the same asymptotic framework. One can view the resulting bounds as a kind of 'independent RIP', where the assumption of independence between the measurement matrix and the original signal allows the tightening of bounds on Rayleigh quotients. Such an analysis is only possible if matrix-vector independence can be assumed, which is the case for the stable point condition (IV.36). We define three tail bound functions.

Definition IV.5 (χ^2 tail bounds) *Let $\rho \in (0,1)$ and $\lambda \in (0,1]$. Let $\mathcal{TIU}(\rho,\lambda)$ be the unique solution to*

$$\nu - \ln(1+\nu) = \frac{2d\rho\cdot H(d^{-1})}{\lambda} \quad \text{for} \quad \nu > 0, \tag{IV.41}$$

and let $\mathcal{TIL}(\rho,\lambda)$ be the unique solution to

$$-\nu - \ln(1-\nu) = \frac{2d\rho\cdot H(d^{-1})}{\lambda} \quad \text{for} \quad \nu \in (0,1), \tag{IV.42}$$

where $H(\cdot)$ is defined in (III.15).

That $\mathcal{TIU}$ is well-defined follows since the left-hand side of (IV.41) is zero at $\nu = 0$, tends to infinity as $\nu \to \infty$, and is strictly increasing on $\nu > 0$. Similarly, $\mathcal{TIL}$ is well-defined since the left-hand side of (IV.42) is zero at $\nu = 0$, tends to infinity as $\nu \to 1$, and is strictly increasing on $\nu \in (0,1)$.

Definition IV.6 ($\mathcal{F}$ tail bound) *Let $\rho \in (0, 1/2]$ and let $\mathcal{TIF}(\rho)$ be the unique solution in f to*

$$\ln(1+f) - \rho\ln f = 2d\rho\cdot H(d^{-1}) + H(\rho) \quad \text{for} \quad f > \frac{\rho}{1-\rho}, \tag{IV.43}$$

where $H(\cdot)$ is defined in (III.15).

That $\mathcal{TIF}$ is well-defined follows since the left-hand side of (IV.43) is equal to $H(\rho)$ at $f = \rho/(1-\rho)$, tends to infinity as $f \to \infty$, and is strictly increasing on $f > \rho/(1-\rho)$.

The bounds given in Definitions IV.5 and IV.6 are related to those given in the context of standard sparsity in [3, Definitions 4.4 and 4.5], and their intuition is as follows. The expressions on the left-hand sides of (IV.41), (IV.42) and (IV.43) capture the rate of exponential decay of the χ^2 and F distributions, and these expressions are identical to the corresponding expressions in [3]. The difference lies in the expressions on the right-hand side, which capture the effect of the union bound over all permissible support sets. As was observed for the bounds on tree-based RIP in Section III, the number of permissible support sets in the tree-based setting is given by (III.14), which has no dependence upon the ambient dimension N, which is why the expressions on the right-hand sides of (IV.41), (IV.42) and (IV.43) depend only upon ρ (for a given tree order d).

Lemma IV.7 (Tree-based large deviations result for χ^2) *Let $l \in \{1,\dots,n\}$ and let the random variables $X_l^i \sim \frac{1}{l}\chi_l^2$ for all $i \in S_n$, where $|S_n| = T(k)$, and let $\epsilon > 0$. In the simplified proportional growth asymptotic, let $l/n \to \lambda \in (0,1]$. Then*

$$\mathbb{P}\left\{\cup_{i\in S_n}[X_l^i \geq 1 + \mathcal{TIU}(\rho,\lambda) + \epsilon]\right\} \to 0 \qquad \text{(IV.44)}$$

and

$$\mathbb{P}\left\{\cup_{i\in S_n}[X_l^i \leq 1 - \mathcal{TIL}(\rho,\lambda) - \epsilon]\right\} \to 0, \qquad \text{(IV.45)}$$

exponentially in n, where $\mathcal{TIU}(\rho,\lambda)$ and $\mathcal{TIL}(\rho,\lambda)$ are defined in (IV.41) and (IV.42) respectively.

Lemma IV.8 (Tree-based large deviations results for F) *Let the random variables $X_n^i \sim \frac{k}{n-k+1}\mathcal{F}(k, n-k+1)$ for all $i \in S_n$, where $|S_n| = T(k)$, and let $\epsilon > 0$. In the simplified proportional growth asymptotic,*

$$\mathbb{P}\left\{\cup_{i\in S_n}[X_n^i \geq \mathcal{TIF}(\rho) + \epsilon]\right\} \to 0, \qquad \text{(IV.46)}$$

exponentially in n, where $\mathcal{TIF}(\rho)$ is defined in (IV.43).

We define oversampling thresholds for ITP algorithms in terms of the above tail bounds.

Definition IV.9 (Oversampling threshold for ITP) *Define $\hat{\rho}_{SP}^{ITP}$ to be the unique solution to*

$$\frac{\sqrt{\mathcal{TIF}(\rho)}}{(1-\rho)\left[1 - \mathcal{TIL}(\rho, 1-\rho)\right]} = \frac{1}{1+\mathcal{TU}(2\rho)}, \quad \rho \in (0, 1/2], \qquad \text{(IV.47)}$$

where $\mathcal{TIF}$ is defined in (IV.43), $\mathcal{TIL}$ is defined in (IV.42) and $\mathcal{TU}$ is defined in Definition III.6.

The oversampling threshold (IV.47) is a counterpart of the phase transitions given in [27, Section 5.2] for IHT algorithms, with the only changes being the switch to tree-based tail bounds and the disappearance of the δ variable. A proof that (IV.47) admits a unique solution proceeds analogously to the one given for the counterpart phase transitions in [27, Section 5.2]. Next, we define a function $\xi_{SP}^{ITP_\alpha}(\rho)$ which will represent a stability factor in our results, bounding the approximation error of the output of ITP as a multiple of the noise level σ.

Definition IV.10 (Stability factor for ITP) *Consider Problem 2. Given $\rho \in (0, 1/2]$ and $\underline{\alpha} > 0$, provided*

$$\rho < \hat{\rho}_{SP}^{ITP}, \qquad \text{(IV.48)}$$

define

$$a(\rho) := \frac{\sqrt{\mathcal{TIF}(\rho)} + \alpha\sqrt{\rho(1-\rho)[1+\mathcal{TIU}(\rho,1-\rho)][1+\mathcal{TIU}(\rho,\rho)]}}{\alpha(1-\rho)[1-\mathcal{TIL}(\rho,1-\rho)] - \sqrt{\mathcal{TIF}(\rho)}}, \qquad \text{(IV.49)}$$

and

$$\xi_{SP}^{ITP_\alpha}(\rho) := \sqrt{\mathcal{TIF}(\rho)\left[1+a(\rho)\right]^2 + \left[a(\rho)\right]^2}, \qquad \text{(IV.50)}$$

where $\mathcal{TIF}$ is defined in (IV.43), and where $\mathcal{TIU}$ and $\mathcal{TIL}$ are defined in (IV.41) and (IV.42) respectively.

Note that (IV.48) ensures that the denominator in (IV.49) is strictly positive and that $a(\rho)$ is therefore well-defined. We proceed to our recovery result for constant stepsize ITP.

Theorem IV.11 (Stable point recovery for ITP) *Consider Problem 2 and suppose Assumptions 1, 3 and 4 hold. If (IV.48) holds and the stepsize α satisfies*

$$\frac{\sqrt{\mathcal{TIF}(\rho)}}{(1-\rho)\left[1-\mathcal{TIL}(\rho,1-\rho)\right]} < \alpha < \frac{1}{1+\mathcal{TU}(2\rho)}, \qquad \text{(IV.51)}$$

then, in the simplified proportional-growth asymptotic[5], ITP with stepsize α converges to $\bar{x}$ such that

$$\|\bar{x} - x^*\| \leq \xi_{SP}^{ITP_\alpha}(\rho)\cdot\sigma, \qquad \text{(IV.52)}$$

with probability tending to 1 exponentially in n.

Proof: See Appendix B. □

In the special case of Problem 1, the same oversampling threshold guarantees exact recovery of the underlying signal x^*.

Corollary IV.12 *Consider Problem 1. Suppose Assumptions 1 and 3 hold, suppose that (IV.48) holds, and suppose that α satisfies (IV.51). Then, in the simplified proportional-growth asymptotic, ITP with stepsize α converges to x^* with probability tending to 1 exponentially in n.*

Proof: See Appendix B. □

1) Results for NITP: We now turn our attention to NITP, and define the following oversampling threshold and stability factor in this case.

Definition IV.13 (Oversampling threshold for NITP) *Define $\hat{\rho}_{SP}^{NITP_\kappa}$ to be the unique solution to*

$$\frac{\sqrt{\mathcal{TIF}(\rho)}}{(1-\rho)\left[1-\mathcal{TIL}(\rho,1-\rho)\right]} = \frac{1}{\kappa[1+\mathcal{TU}(2\rho)]}, \quad \rho \in (0, 1/2], \qquad \text{(IV.53)}$$

where $\mathcal{TIF}$ is defined in (IV.43), $\mathcal{TIL}$ is defined in (IV.42) and $\mathcal{TU}$ is defined in Definition III.6.

A proof that (IV.53) admits a unique solution proceeds analogously to the one given for the counterpart phase transitions in [27, Section 5.2]. We define the following stability factor for NITP.

Definition IV.14 (Stability factor for NITP) *Consider Problem 1. Given $\rho \in (0, 1/2]$, provided*

$$\rho < \hat{\rho}_{SP}^{NITP_\kappa}, \qquad \text{(IV.54)}$$

[5]In other words, we consider instances of k-tree sparse vectors x^* and Gaussian random variables A and e for a sequence of triples (k, n, N) where $n \to \infty$, where n is the number of measurements, N, the signal dimension and k, the sparsity of the underlying signal.

define

$$a(\rho) := \frac{\sqrt{\mathcal{TIF}(\rho)} + \{\kappa[1+\mathcal{TU}(2\rho)]\}^{-1}\sqrt{\rho(1-\rho)[1+\mathcal{TIU}}}{(1-\rho)\{\kappa[1+\mathcal{TU}(2\rho)]\}^{-1}[1-\mathcal{TIL}(\rho,1-} \quad \text{(IV.55)}$$

and

$$\xi_{SP}^{NITP_\kappa}(\rho) := \sqrt{\mathcal{TIF}(\rho)\,[1+a(\rho)]^2 + [a(\rho)]^2}, \quad \text{(IV.56)}$$

where $\mathcal{TIF}$ is defined in (IV.43), where $\mathcal{TIU}$ and $\mathcal{TIL}$ are defined in (IV.41) and (IV.42) respectively, and where $\mathcal{TU}$ is defined in Definition III.7.

Theorem IV.15 (Stable point recovery for NITP) *Consider Problem 2, suppose Assumptions 1, 3 and 4 hold, and suppose (IV.54) holds. Then, in the simplified proportional-growth asymptotic, NITP with shrinkage parameter κ converges to $\bar{x}$ such that*

$$\|\bar{x} - x^*\| \leq \xi_{SP}^{NITP_\kappa}(\rho) \cdot \sigma, \quad \text{(IV.57)}$$

with probability tending to 1 exponentially in n.

Proof: See Appendix B. □

In the case of Problem 1, Theorem IV.15 also simplifies to an exact recovery result.

Corollary IV.16 *Consider Problem 1. Suppose Assumptions 1 and 3 hold and suppose that (IV.54) holds. Then, in the simplified proportional-growth asymptotic, NITP with shrinkage parameter κ converges to x^* with probability tending to 1 exponentially in n.*

Proof: See Appendix B. □

V. Discussion of Recovery Results

A. Tree-based RIP recovery results

Noiseless case. The oversampling thresholds for ITP and NITP given by Definition III.10 and Corollary III.9 are displayed in Figure 2(a) for different tree orders d. For binary trees, for example, we have $\hat{\rho}_{RIP}^{ITP_{\hat{\alpha}}} \approx 0.00875$ for ITP and $\hat{\rho}_{RIP}^{NITP_\kappa} \approx 0.00146$ for NITP (taking $\kappa = 1.1$ for the shrinkage parameter in NITP). In both cases, exact recovery in the noiseless case is asymptotically guaranteed provided the limiting value of the ratio ρ is less than the given threshold. We see a measured deterioration in the results for higher tree orders: the corresponding thresholds for quad-trees ($d = 4$) – which arise in image analysis using 2D wavelets – are 0.00705 and 0.00123 for ITP and NITP respectively. Figure 2(b) shows the inverse of the oversampling ratio, which indicates the number of measurements required by the analysis as a multiple of the sparsity. We find, for binary trees, that $n \geq 115k$ measurements guarantees recovery by ITP, while $n \geq 683k$ measurements guarantees recovery by NITP. Provided the oversampling thresholds are respected, convergence to the original signal is guaranteed at a linear rate. The quantities $\mu_{RIP}^{ITP_{\hat{\alpha}}}(\rho)$ and $\mu_{RIP}^{NITP_{1.1}}(\rho)$ represent guaranteed bounds on the convergence rate for each variant.

While in the present paper we have dispensed with the undersampling ratio $\delta = n/N$, we may also frame our results in the (δ, ρ) asymptotic in order to make a comparison with analogous results derived in the non-tree-based setting for IHT based upon the standard notion of RIP [27]. Since there is no dependence upon δ in our case, the phase transitions we obtain are simply horizontal lines in the (δ, ρ)-plane. Exact recovery phase transitions for binary trees are displayed in Figure 3 alongside the phase transitions derived in [3]: recovery is guaranteed asymptotically beneath the respective curves. We observe that the switch to the tree-based setting leads to significantly improved results, especially for small δ.

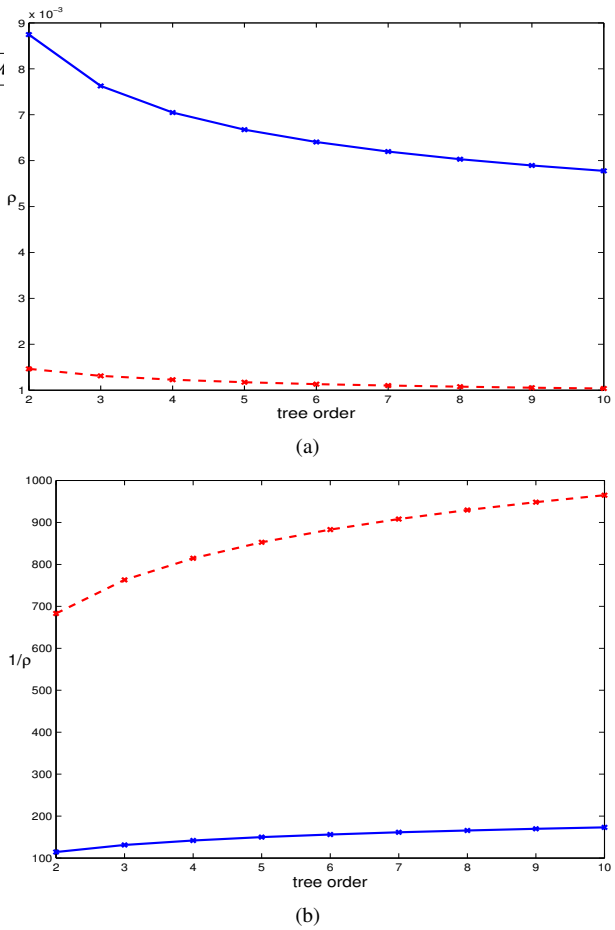

Fig. 2. (a) Critical ρ-values for different tree orders from tree-based RIP analysis: ITP – unbroken; NITP – dashed. (b) Corresponding oversampling factors (reciprocals of $\hat{\rho}$).

Comparison with prior work Analogous oversampling ratios can be explicitly obtained in the case of binary trees using the prior analysis in [1]. We observe that the oversampling thresholds given by Definition III.10 and Corollary III.9 represent a scale factor improvement of around 100 over those obtainable using the analysis in [1]. The precise thresholds for binary trees are given in Table I for comparison, along with the scale factor improvement. For the prior analysis, the optimal stepsize for ITP is $\alpha := 1$, and the parameter κ is again taken to be 1.1. The dramatic improvement in oversampling thresholds is due to the tightening of the tree-based RIP bounds in Definition III.6 over those in Definition III.12. This tightening is achieved through an asymmetric treatment of the tree-based RIP accompanied by a tighter large deviations analysis based on the PDF of the extreme singular values of the submatrices of Gaussian matrices, as opposed to the more generic sphere-covering argument relied upon in [1].

Extension to noise. In the case where measurements are contaminated by noise, exact recovery of the original signal is an unrealistic aim. However, provided the limiting value of the ratio ρ falls below the respective oversampling threshold, Theorem III.8 gives bounds on the limiting approximation error. More precisely, the results state that the limiting approximation error of the iterates of ITP/NITP is asymptotically bounded by

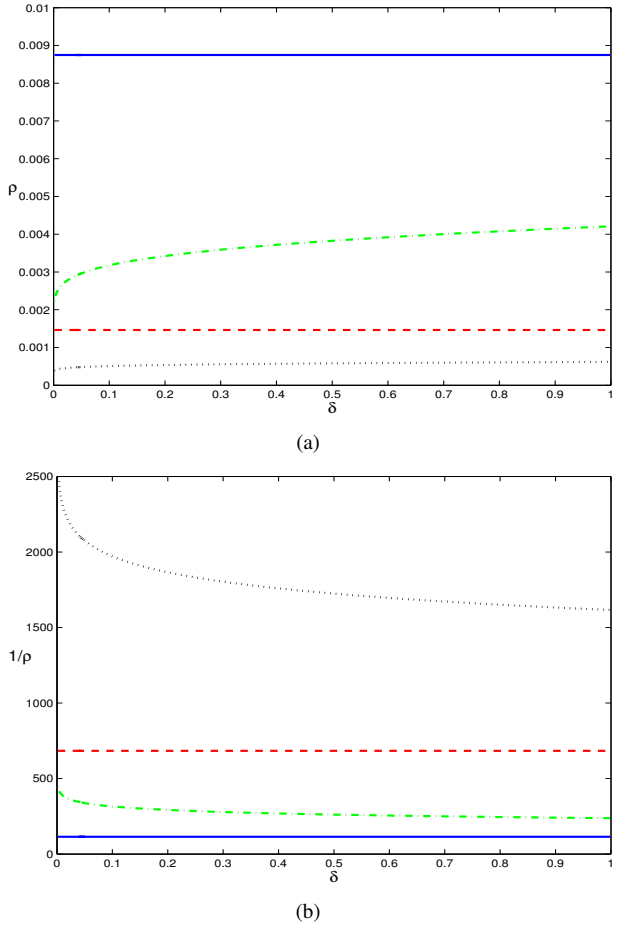

Fig. 3. (a) Phase transitions from RIP analysis in the (δ, ρ) framework for binary trees (ITP – unbroken; NITP – dashed) and non-tree-based (IHT – dash-dot; NIHT – dotted). (b) Corresponding inverses of the phase transition.

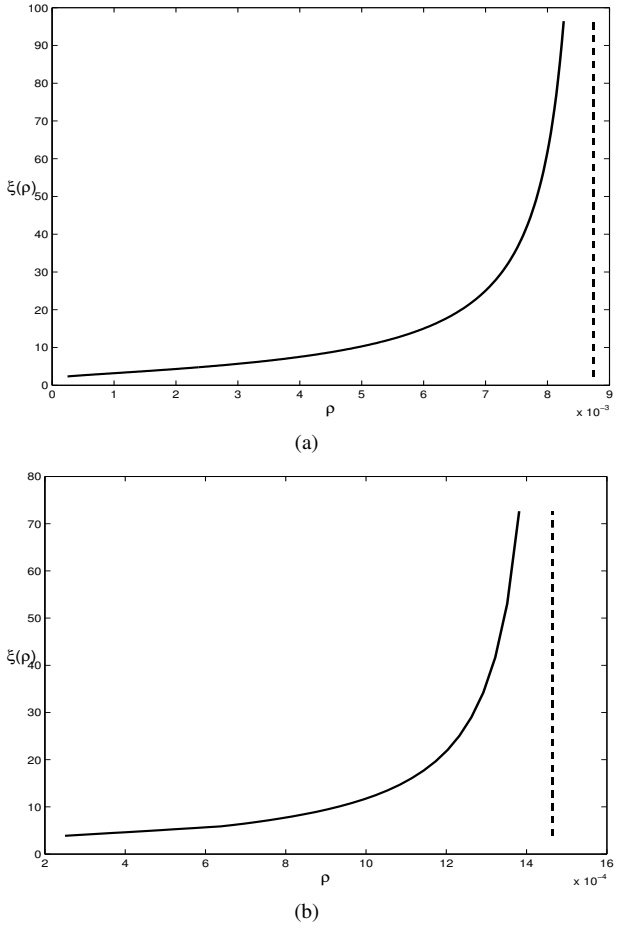

Fig. 4. Plot of the stability factor $\xi(\rho)/[1-\mu(\rho)]$ from tree-based RIP analysis for binary trees: (a) ITP; (b) NITP.

	Current paper		Analysis in [1]		Factor
	ρ	ρ^{-1}	ρ	ρ^{-1}	gain
ITP	8.75×10^{-3}	115	1.24×10^{-4}	8068	70
NITP	1.46×10^{-3}	683	1.25×10^{-5}	79705	116

TABLE I
COMPARISON OF OVERSAMPLING THRESHOLDS OBTAINED FROM THE CURRENT ANALYSIS AND THE PRIOR ANALYSIS IN [1], IN THE CASE OF BINARY TREES.

some known stability factor multiplied by the noise level σ. However, neither result necessarily implies convergence of the algorithm in the case of noise. The Figure 4 plots the noise stability factor $\xi(\rho)/[1-\mu(\rho)]$ for binary trees, for each of the two stepsize schemes considered ($\kappa = 1.1$ for NITP). In keeping with [32], [24] and [3], we observe that the stability factor tends to infinity as the transition point is reached, *i.e.* $\xi(\rho)/[1-\mu(\rho)] \to \infty$ as $\rho \to \hat{\rho}$. For both ITP and NITP, given any value of ρ for which the stability factors derived in this paper are defined, they are always lower than the corresponding stability factors derived from analysis of IHT based upon the standard RIP [26]; see [27, Section 2.4] for a comparison.

B. Recovery results from the tree-based stable point analysis

Noiseless case. The oversampling thresholds for ITP and NITP defined in Corollaries IV.12 and IV.16 are displayed in Figure 5(a) for different tree orders d. For binary trees, we have $\hat{\rho}_{RIP}^{ITP_\alpha} \approx 0.0202$ for ITP and $\hat{\rho}_{RIP}^{NITP_{1.1}} \approx 0.0184$ for NITP, and the corresponding thresholds for quad-trees ($d = 4$) are 0.0147 and 0.0134 respectively. Figure 5(b) shows the inverse of the oversampling ratio: we find, for binary trees, that $n \geq 50k$ measurements guarantees recovery by ITP, while $n \geq 55k$ measurements guarantees recovery by NITP. The same exact recovery thresholds for binary trees are presented in the form of phase transitions in the (δ, ρ) asymptotic in Figure 6, alongside the phase transitions for IHT/NIHT derived in [3]. Again, we observe improved results by switching to the tree-based setting, especially for small δ.

Comparing the oversampling thresholds derived from the stable point analysis (Figure 5) with those derived from tree-based RIP analysis (Figure 2), we observe a significant quantitative improvement for both algorithm variants, by over a factor of 10 for NITP in fact for all tree orders under consideration. We have obtained improved oversampling thresholds by exploiting average-case assumptions, and we should point out the difference between the results in Sections V-A and V-B. The tree-based RIP results are worst-case in nature: given a sequence of randomly generated Gaussian matrices, it is asymptotically guaranteed that ITP/NITP will in fact recover an accurate approximation to *any* k-tree sparse signal vector. On the other hand, the results derived from our stable point analysis have a more average-case flavour: given a sequence of randomly generated Gaussian measurement matrices *along with* a sequence of signal and noise vectors which are both independent of the measurement matrix, recovery is asymp-

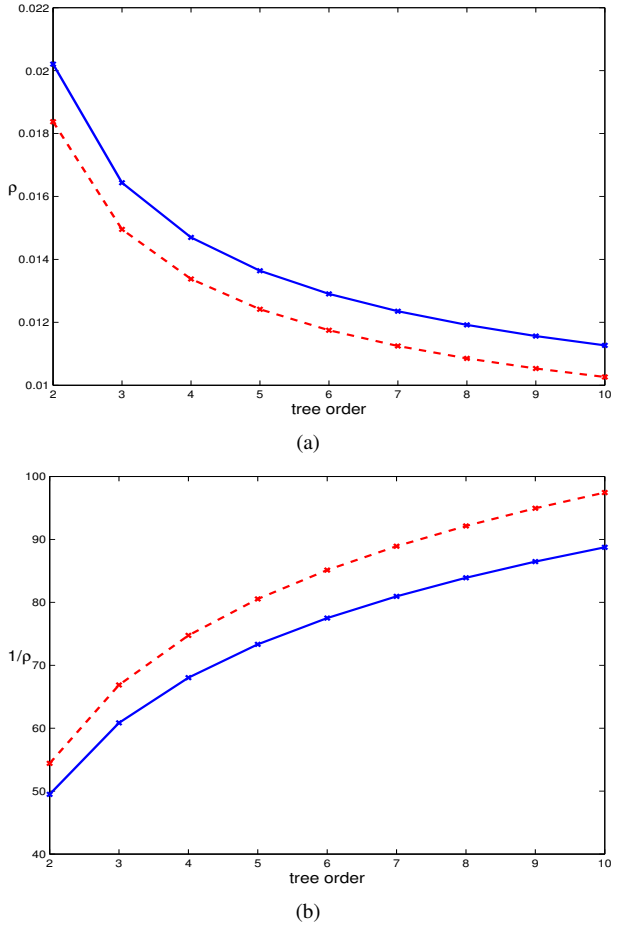

Fig. 5. (a) Critical ρ-values for different tree orders from stable point analysis: ITP – unbroken; NITP – dashed. (b) Corresponding oversampling factors (reciprocals of $\hat{\rho}$).

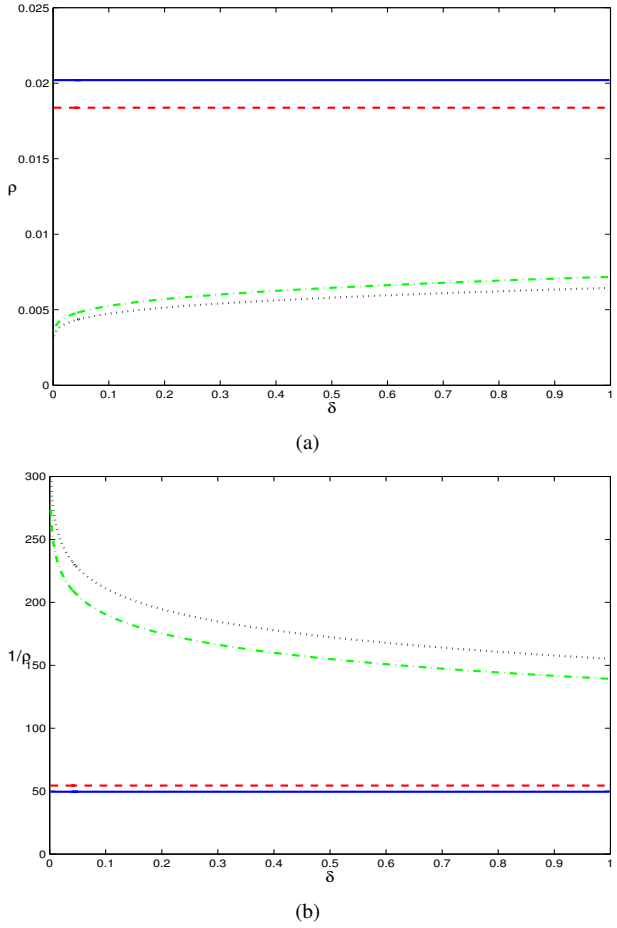

Fig. 6. (a) Phase transitions from stable point analysis in the (δ, ρ) framework for binary trees (ITP – unbroken; NITP – dashed) and non-tree-based (IHT – dash-dot; NIHT – dotted). (b) Corresponding inverses of the phase transition.

totically guaranteed in this sense. It is not surprising that our average-case framework leads to an improvement over tree-based RIP since the assumption of independence between signal and measurement matrix rules out the practically unlikely case in which one chooses the very worst possible signal for a given measurement matrix. For a comparison of phase transitions derived from both stable point and RIP analysis in the context of IHT and simple sparsity, we refer the reader to [3, Section 6].

Extension to noise. Below the same oversampling thresholds, Theorems IV.11 and IV.15 go further than the tree-based RIP analysis in proving convergence of ITP/NITP to a limit point — whose approximation error is asymptotically bounded by some known stability factor multiplied by the noise level σ. Figure 7 plots the noise stability factor $\xi(\rho)$ for binary trees, for each of the two stepsize schemes considered ($\kappa = 1.1$ for NITP). For both ITP and NITP, given any value of ρ for which the stability factors derived in this paper are defined, they are always lower than the corresponding stability factors derived from analysis of IHT based upon the standard RIP [26]; see [27, Section 2.4] for a comparison.

Comparing Figure 7 with Figure 4, we also observe a significant quantitative improvement in the stability factors for both algorithm variants compared with those achieved by means of tree-based RIP, in the case of binary trees. It should be pointed out that we have obtained improved stability results by imposing additional restrictions upon the noise, namely that the noise is Gaussian distributed and independent of the signal and measurement matrix. This assumption is in keeping with our aim of exploiting average-case assumptions. Our analysis could, however, be altered to deal with the case of non-independent noise by making more use of the RIP, though this would lead to larger stability constants.

C. Extension to tree compressible signals

While we have assumed so far in this paper that signals are exactly k-tree sparse, it is more realistic to expect that signals are *tree compressible*, meaning that they are well approximated by a k-tree sparse vector. An important consideration for any compressed sensing recovery analysis is, therefore, whether it can be extended to the tree compressible case. From the point of view of worst-case analysis, a difference emerges in this respect between standard and tree-based compressed sensing. In the case of standard compressed sensing, the extension to compressible signals can be achieved using the RIP, which can be used to bound the amplification factor of the signal tail [13]. However, it was argued in [2] that the RIP is not sufficient to control this amplification factor for more general structured sparsity models (including the tree-based model). This deficit was partially addressed by the introduction of the Restricted Amplification Property (RAmP), and the extension to model-compressible signals was established provided the sparsity model has a certain 'nested' property [2], which unfortunately is not the case for the rooted tree model.

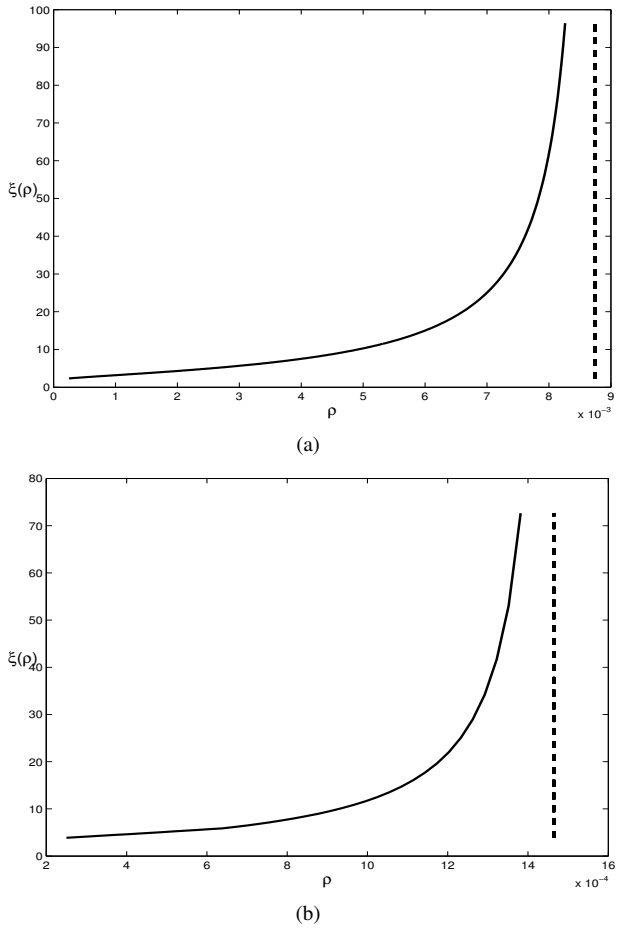

Fig. 7. Plot of the stability factor $\xi(\rho)$ from stable point analysis for binary trees: (a) ITP; (b) NITP.

On the other hand, the stable point approach in which we consider independent Gaussian noise is much more amenable to the analysis of the tree-compressible case. In [27, Chapter 7], the main results of the present paper are extended to the tree-compressible case. More precisely, the assumption that x^* is k-tree sparse is relaxed, and x_k^* is defined to be the closest k-tree sparse approximation to x^*, namely $x_k^* := \mathcal{P}_k(x^*)$. Defining Λ^k to be the support of this optimal tree-sparse approximation, that is $\Lambda^k := \text{supp}(x_k^*)$, a measure of unrecoverable energy, Σ, is defined to be

$$\Sigma := \sigma + \|x^*_{\Lambda^k}\|,$$

which represents the combined inaccuracy due to both measurement noise and signal model violation. It is shown in [27, Theorems 7.23 and 7.29] that, beneath the same oversampling thresholds given in Theorems IV.11 and IV.15 of the present paper, the approximation error of the output of ITP/NITP amplifies the unrecoverable energy by no more than some (different) stability factor. See [27, Chapter 7] for an explicit quantification of the stability factor in this case.

The observation that controlling stability to noise in tree-based compressed sensing is alleviated by switching to average-case assumptions is not new, see for example [37], [38], [39].

VI. Concluding remarks and future directions

We have introduced a simplified proportional-growth asymptotic framework, and used it to quantify recovery guarantees for ITP algorithms. Recovery guarantees in terms of tree-based RIP have also been obtained for tree-based CoSaMP [2], while recovery guarantees based on the standard notion of RIP for other greedy algorithms including Conjugate Gradient Iterative Hard Thresholding [40], Subspace Pursuit [41] and Orthogonal Matching Pursuit [42] could also be translated into the tree-based framework. Our asymptotic framework could equally well be used to quantify the existing RIP-based recovery guarantees for these algorithms and for Gaussian matrices.

Our focus in this paper has been on the tree-based model, but other variants of the model-based compressed sensing paradigm are possible, including block sparsity [2]. Many of the arguments we have presented for the tree-based model would apply equally to other union-of-subspaces model. We leave the extension of our analysis to other models as future work.

The calculation of an exact tree projection is computationally burdensome, and the more recently proposed approximate ITP algorithm [22] is attractive in practice. An interesting direction for future work is to extend our results by obtaining quantified oversampling thresholds for approximate ITP algorithms.

Appendix

The current appendix gives a proof outline for the tree-based RIP analysis, and Appendix B gives a proof outline for the stable point analysis. In both cases, we first focus on recovery conditions for deterministic matrices (in Sections A and C respectively). We then turn to the probabilistic analysis of these conditions for Gaussian matrices in the simplified proportional-growth asymptotic (in Sections B and D respectively). In both cases, the analysis for Gaussian matrices relies on large deviations results for certain quantities related to Gaussian matrices (including bounds on tree-based RIP constants). These large deviations results, which extend to the tree-based setting those given originally in [32], [3], are stated and proved in Appendix D.

A. Deterministic recovery conditions

The proof of Theorem III.2 is identical to the one given in [27, Section 2.2] for IHT and NIHT, except that bounds on the standard notion of RIP are replaced by their tree-based equivalents. A full proof is provided in Appendix A.1 of the extended technical report [43]. The proof consists of a translation of the analysis in [26] for symmetric standard RIP to the asymmetric and tree-based setting.

B. Analysis for Gaussian matrices

To prove Theorem III.8, we must show that a naive replacement of each TL_{pk} and TU_{qk} by the tree-based RIP bounds $\mathcal{TL}(p\rho)$ and $\mathcal{TU}(q\rho)$ in the deterministic recovery conditions of Theorem III.2 is valid. A full proof is given in Appendix A.2 of the accompanying extended technical report [43]. The proof largely mimics the approach of [24] for recovery conditions involving the standard RIP, and the one non-trivial step in extending the proof is to show that the functions $\mathcal{TL}(\rho)$ and $\mathcal{TU}(\rho)$ satisfy the necessary conditions of [24, Lemma 12], namely that (i) they are indeed upper bounds on TL_k and TU_k respectively, (ii) $\mathcal{TL}(\rho)$ is strictly increasing in ρ, and (iii) $\mathcal{TU}(\rho)$ is nondecreasing in ρ. The first condition holds in our case by Lemma III.7, and it is straightforward to show that the second and third properties also holds in our case. More precisely, $\mathcal{TL}(\rho)$ and $\mathcal{TU}(\rho)$ are both strictly increasing on $\rho \in (0, 1)$. It follows that the argument in [24] extends.

Corollary III.9 then follows from Theorem III.8 by setting $e := 0$.

C. Analysis for deterministic matrices

We will follow the approach first introduced by the present authors in [3], central to which is the concept of an $\underline{\alpha}$-*stable point*, defined in Definition IV.1.

We will analyse the stable points of generic ITP, and our final goal is to prove quantitative conditions that guarantee that all stable points of the algorithm are 'close' to the original signal x, in the context of Gaussian matrices. Provided we also have guaranteed convergence to some stable point, we may then conclude that ITP outputs a good approximation to x. For this reason, the results derived in this section come in two parts: a necessary condition for there to be a stable point on some support Γ, and conditions guaranteeing convergence to some stable point for our two stepsize schemes.

1) A necessary condition for the existence of a stable point: Any $\underline{\alpha}$-stable point of generic ITP may also be characterized as a minimum-norm solution on some k-subspace.

Lemma A.1 *Suppose Assumption 2 holds and suppose $\bar{x}$ is an $\underline{\alpha}$-stable point of generic ITP on Γ for some $\underline{\alpha} > 0$. Then $\bar{x}_\Gamma = A_\Gamma^\dagger b$.*

Proof: It follows from (IV.32) that $A_\Gamma^T(b - A_\Gamma \bar{x}_\Gamma) = 0$ where $\mathrm{supp}(\bar{x}) \subseteq \Gamma$ and $|\Gamma| = k$. Under Assumption 2, the pseudoinverse $A_\Gamma^\dagger$ is well-defined and we may rearrange to give $\bar{x}_\Gamma = A_\Gamma^\dagger b$. □

While this lemma tells us that any stable point is necessarily a minimum-norm solution on some k-subspace, the converse may not hold. We next prove Theorem IV.3, which gives a necessary condition for a stable point on a given support.

Proof of Theorem IV.3: Supposing that $\bar{x}$ is an $\underline{\alpha}$-stable point on Γ, choosing $\Omega := \Lambda$ in (IV.33) yields

$$\|\bar{x}_{\Gamma\setminus\Lambda}\|^2 \geq \underline{\alpha}^2 \|A_{\Lambda\setminus\Gamma}^T(b - A\bar{x})\|^2.$$

We may now follow the argument of [3, Theorem 3.2] to deduce (IV.36). □

2) Conditions guaranteeing convergence: In addition to the result of the previous section, in order to show recovery of x^*, we must also show that ITP converges to an $\underline{\alpha}$-stable point. In this section we derive convergence conditions for generic ITP used in conjunction with the two stepsize schemes introduced in Section II-B. A sufficient condition for convergence of generic ITP is given next.

Lemma A.2 (Sufficient condition for convergence) *Consider Problem 2. Suppose Assumption 2 holds, and suppose the iterates of generic ITP satisfy*

$$\|x^{m+1} - x^m\|^2 \leq c\left[\Psi(x^m) - \Psi(x^{m+1})\right] \quad \text{for all } m \geq 0, \tag{A.58}$$

for some $c > 0$ which does not depend upon m, where $\Psi(\cdot)$ is defined in (II.2). Assume that there exist $\overline{\alpha} \geq \underline{\alpha} > 0$ such that

$$\overline{\alpha} \geq \alpha^m \geq \underline{\alpha} \quad \text{for all } m \geq 0. \tag{A.59}$$

Then $x^m \to \bar{x}$ as $m \to \infty$, where $\bar{x}$ is an $\underline{\alpha}$-stable point of generic ITP.

Proof: We may follow the proof of [3, Lemma 3.5] to deduce that $x^m \to \bar{x}$, where $\bar{x}_\Gamma = A_\Gamma^\dagger b$ and $\bar{x}_{\Gamma^C} = 0$, for some Γ such that $|\Gamma| = k$. The proof still holds since all that is assumed about the hard threshold projection $\mathcal{H}_k(\cdot)$ is that it preserves the value of selected coefficients, a property which is also shared by the tree projection $\mathcal{P}_k(\cdot)$ by (II.7). Since $\Gamma = \Gamma^m$ for some $m \geq 0$, it follows that, in the case of ITP, $\Gamma \in \mathcal{T}_k$. Therefore (IV.32) holds for $\bar{x}$.
It remains to establish that $\bar{x}$ satisfies (IV.33). Defining

$$\Gamma_1 = \{i \in \Gamma \ : \ \bar{x}_i \neq 0\}, \tag{A.60}$$

it follows that $\Gamma_1 \subseteq \Gamma^m$ for all m sufficiently large. It follows from (II.7) that, for any $\Omega \in \mathcal{T}_k$,

$$\|x_{\Gamma^{m+1}}^{m+1}\|^2 \geq \|\{x^m - \alpha^m g^m\}_\Omega\|^2, \ \text{ for all } \ m \geq 0.$$

and therefore, for all m sufficiently large,

$$\|x_{\Gamma_1}^{m+1}\|^2 + \|x_{\Gamma^{m+1}\setminus\Gamma_1}^{m+1}\|^2 \geq \|x_{\Omega\cap\Gamma_1}^{m+1}\|^2 + \|\{x^m - \alpha^m g^m\}_{\Omega\setminus\Gamma_1}\|^2,$$

which cancels to

$$\|x_{\Gamma_1\setminus\Omega}^{m+1}\|^2 + \|x_{\Gamma^{m+1}\setminus\Gamma_1}^{m+1}\|^2 \geq \|\{x^m + \alpha^m g^m\}_{\Omega\setminus\Gamma_1}\|^2. \tag{A.61}$$

Furthermore, it follows from (A.60) that

$$\|x_{\Gamma^{m+1}\setminus\Gamma_1}^{m+1}\|^2 \to 0. \tag{A.62}$$

By (A.59), there exists a convergent subsequence of stepsizes,

$$\alpha^{m_r} \to \tilde{\alpha} \geq \underline{\alpha} \ \text{ as } \ r \to \infty \tag{A.63}$$

Passing to the limit in (A.61) on the subsequence m_r for which (A.63) holds, we deduce that $\|\bar{x}_{\Gamma_1\setminus\Omega}\| \geq \underline{\alpha}\|\{A^T(b - A\bar{x})\}_{\Omega\setminus\Gamma_1}\|$, from which it follows trivially that

$$\|\bar{x}_{\Gamma\setminus\Omega}\| \geq \underline{\alpha}\|\{A^T(b - A\bar{x})\}_{\Omega\setminus\Gamma}\|. \tag{A.64}$$

Since (A.64) holds for any $\Omega \in \mathcal{T}_k$, $\bar{x}$ satisfies (IV.33), and the result is proved. □

Proof of Theorem IV.2: We may follow the proof of [3, Theorem 3.6], replacing U_{2k} with TU_{2k}, to deduce that (A.58) holds with $c := 2\alpha/[1 - \alpha(1 + TU_{2k})]$. Due to (IV.34), (A.59) trivially holds with $\overline{\alpha} = \underline{\alpha} = \alpha$. Thus Lemma A.2 applies, and the ITP iterates x^m converge to an α-stable point. □

We next obtain a convergence result for NITP. In this case, there is no explicit requirement for a tree-based RIP condition to be satisfied; however, the tree-based RIP this time appears in the choice of $\underline{\alpha}$.

Theorem A.3 (NITP convergence) *Suppose Assumption 2 holds. Then NITP with shrinkage parameter κ converges to a $[\kappa(1 + TU_{2k})]^{-1}$-stable point $\bar{x}$ of generic ITP.*

Proof: By replacing L_{2k} with TL_{2k}, the proof given for [3, Theorem 3.7] holds. □

D. Analysis for Gaussian matrices

In this section, we present an outline of the proofs of the recovery results in Section IV, which build upon the results for arbitrary matrices in Section C and give quantitative oversampling thresholds for ITP algorithms of the form $\rho < \hat{\rho}$ in the case of Gaussian measurement matrices. The method of proof in [3] can be followed for both stepsize schemes, with two important changes. First, we switch to using the tree-based tail bounds defined in Appendix D. Second, since there is now

no dependence upon δ, we can prove results in the simplified proportional-growth asymptotic (Definition I.1). A full proof is given in Appendix B.2 of the accompanying extended technical report [43], and we here give a brief outline of the method of proof.

Central to the proof is a support set partition.

Definition A.4 (Support set partition) *Consider Problem 2 and suppose $\rho \in (0, 1/2]$ and $\alpha > 0$. Given $\zeta > 0$, let us write*

$$a^*(\rho;\zeta) := a(\rho) + \zeta, \tag{A.65}$$

let us write $\{\Gamma_i : i \in \mathcal{T}_k\}$ for the set of all possible support sets which form a rooted tree of cardinality k, and let us disjointly partition $\mathcal{T}_k := \Theta_n^1 \cup \Theta_n^2$ such that

$$\begin{aligned} \Theta_n^1 &:= \left\{ i \in \mathcal{T}_k \ : \ \|x^*_{\Lambda\setminus\Gamma_i}\| > \sigma \cdot a^*(\rho;\zeta) \right\} \quad \text{and} \\ \Theta_n^2 &:= \left\{ i \in \mathcal{T}_k \ : \ \|x^*_{\Lambda\setminus\Gamma_i}\| \leq \sigma \cdot a^*(\rho;\zeta) \right\}, \end{aligned} \tag{A.66}$$

where Λ is defined in (IV.35).

For ITP, the partition in (A.66), with $a(\rho)$ given by (IV.49), is defined in such a way that, provided the oversampling threshold defined in (IV.47) is satisfied, an analysis of the stable point condition (IV.36), using the large deviations results of Appendix D, shows that ITP must necessarily converge to some α-stable point on Γ_i such that $i \in \Theta_n^2$. On the other hand, it is also possible to use the large deviations results of Appendix D to bound the error in approximating x^* by any α-stable point on Γ_i such that $i \in \Theta_n^2$. It follows that, for any $\alpha > 0$, all α-stable points have bounded approximation error, which is the result of Theorem IV.11. Corollary IV.12 then follows by setting $e := 0$.

For NITP, the proof of Theorem IV.15 follows broadly the same lines, except that the support set partition in (A.66) is defined with $a(\rho)$ given by (IV.55), and α-stable points are replaced by $\underline{\alpha}(\rho;\epsilon)$-stable points, where

$$\underline{\alpha}(\rho;\epsilon) := \{\kappa[1 + \mathcal{TU}(2\rho) + \epsilon]\}^{-1}, \tag{A.67}$$

for some $\epsilon > 0$. Corollary IV.16 then follows by setting $e := 0$.

This appendix develops large deviation bounds in the simplified proportional-growth asymptotic of Definition I.1 for various quantities related to Gaussian matrices, which are required to hold for all permissible support sets.

In what follows, let the tree order d to be some fixed integer with $d \geq 2$. We need to count $|\mathcal{T}_k|$, the number of permissible support sets in the d-ary tree-based framework, which is bounded above by $T(k)$, the total number of ordered, rooted d-ary trees of cardinality k. Recalling Lemma III.4, we have

$$T(k) = \frac{1}{(d-1)k+1}\binom{dk}{k}.$$

A similar result was proved in [2, Proposition 1] for the case of binary trees ($d = 2$), though the result given above represents a generalization to any $d > 2$, and in fact also gives a tightening of the result in [2] in the case where $\log_2(N) > k$. Note also that we have an upper bound on $|\mathcal{T}_k|$ which is independent of N. This is in contrast to the total number of supports, which is $\binom{N}{k}$. However, $|\mathcal{T}_k|$ may not attain this upper bound if additional structure is imposed. In a typical wavelet tree model, for example, the root node has only $d-1$ children [10]). In addition, the number of levels in a wavelet tree structure is typically limited to $J = \log_d(N)$, which represents a further restriction if $\log_d(N) < k$. It follows that, while it is possible to give an upper bound on $|\mathcal{T}_k|$ which is valid for any N, $|\mathcal{T}_k|$ does in general depend on both k and N.

We will make use of the following limiting result for $T(k)$.

Lemma A.5 (Tree counting limit)

$$\lim_{k\to\infty} \frac{1}{k} \ln T(k) = d \cdot H(d^{-1}), \tag{A.68}$$

where $H(\cdot)$ is defined in (III.15).

Proof:

$$\begin{aligned} & \lim_{k\to\infty} \frac{1}{k} \ln \left[\frac{1}{(d-1)k+1} \binom{dk}{k} \right] \\ = \ & \lim_{k\to\infty} \frac{1}{k} \ln \left[\frac{1}{(d-1)k+1} \right] + \lim_{k\to\infty} \frac{1}{k} \ln \binom{dk}{k} \\ = \ & 0 + \lim_{k\to\infty} d \cdot \frac{1}{dk} \ln \binom{dk}{k} \\ = \ & d \cdot H(d^{-1}), \end{aligned}$$

where the last step follows from Stirling's formula. □

We proceed to proving the validity of the bounds on tree-based RIP constants for Gaussian matrices given in Definition III.1.

Proof of Lemma III.7: We may follow the proof of [32, Proposition 2.6], replacing $\mathcal{U}(\delta,\rho)$ with $\mathcal{TU}(\rho)$, and replacing $\lambda^{max}(\delta,\rho)$ with $\lambda^{max}(\rho)$, obtaining

$$\begin{aligned} &\mathbb{P}\left[TU_k \geq \mathcal{TU}(\rho_n) + \epsilon\right] \leq \\ &\quad 2|\mathcal{T}_k| \left[\lambda^{max}(\rho_n) + \epsilon\right] g_{max}\left[k, n; \lambda^{max}(\rho_n) + \epsilon\right], \end{aligned}$$

and we may furthermore apply [32, Lemma 2.5] to give

$$\begin{aligned} &\mathbb{P}\left[TU_k \geq \mathcal{TU}(\rho_n) + \epsilon\right] \leq \\ &\quad 2|\mathcal{T}_k| \left[\lambda^{max}(\rho_n) + \epsilon\right] p_{max}\left[n, \lambda^{max}(\rho_n) + \epsilon\right] \\ &\qquad \cdot \exp\left\{n \cdot \psi_{max}\left(\lambda^{max}(\rho_n) + \epsilon, \rho\right)\right\}, \end{aligned} \tag{A.69}$$

where $p_{max}(n,\lambda)$ is a polynomial in n and λ. Now we may take limits of both sides of (A.69), using (A.68), to deduce

$$\begin{aligned} &\lim_{n\to\infty} \tfrac{1}{n} \ln \mathbb{P}\left[TU_k \geq \mathcal{TU}(\rho_n) + \epsilon\right] \leq \\ &\quad d\rho \cdot H(d^{-1}) + \psi_{max}\Big(\lambda^{max}(\rho) + \epsilon, \rho\Big), \end{aligned}$$

which is strictly negative by Definition III.6, from which (III.20) now follows using the same argument as in the proof of lemma IV.7. An analogous argument may be followed to prove (III.21) which we omit for the sake of brevity. □

The large deviations bounds for the χ^2 and $\mathcal{F}$ distributions follow.

Proof of Lemma IV.7: By [3, Lemma A.2], we have for all $i \in S_n$,

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}(X_l^i \geq 1 + \nu) \leq -\frac{\gamma}{2}[\nu - \ln(1+\nu)]. \tag{A.70}$$

Union bounding $\mathbb{P}(X_l^i \geq 1+\nu)$ over all $i \in S_n$ gives

$$\begin{aligned} &\mathbb{P}\{\cup_{i\in S_n}(X_l^i \geq 1+\nu)\} \leq \\ &\quad \textstyle\sum_i \mathbb{P}\left(X_l^i \geq 1+\nu\right) \leq T(k) \cdot \mathbb{P}(X_l^1 \geq 1+\nu). \end{aligned} \tag{A.71}$$

Taking logarithms and limits of the right-hand side of (A.71), using (A.70) and (A.68), we have

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \ln\left[T(k) \cdot \mathbb{P}(X_l^1 \geq 1+\nu)\right] = \\ &\quad d\rho \cdot H(d^{-1}) - \frac{\lambda}{2}[\nu - \ln(1+\nu)], \end{aligned}$$

and so (A.71) implies that, for any $\eta > 0$,

$$\frac{1}{n}\ln \mathbb{P}\left\{\cup_{i\in S_n}(X_l^i \geq 1+\nu)\right\} \leq d\rho \cdot H(d^{-1}) - \frac{\lambda}{2}[\nu - \ln(1+\nu)] + \eta, \quad \text{(A.72)}$$

for all n sufficiently large. By the definition of $\mathcal{TIU}(\rho,\lambda)$ in (IV.41), and since $[\nu - \ln(1+\nu)]$ is strictly increasing on $\nu > 0$, then, for any $\epsilon > 0$, setting $\nu := \nu^* = \mathcal{TIU}(\rho,\lambda) + \epsilon$ and choosing η sufficiently small in (A.72) ensures

$$\frac{1}{n}\ln \mathbb{P}\left\{\cup_{i\in S_n}(X_l^i \geq 1+\nu^*)\right\} \leq -c_Q$$

for all n sufficiently large, where c_Q is some positive constant, from which it follows that

$$\mathbb{P}\left\{\cup_{i\in S_n}(X_l^i \geq 1+\nu^*)\right\} \leq e^{-c_Q\cdot n}$$

for all n sufficiently large, and (IV.44) follows. Combining the same union bound argument with the lower tail result of [3, Lemma A.2] shows that, if we take $\nu^* = \mathcal{TIL}(\rho,\lambda) + \epsilon$ for some $\epsilon > 0$, then

$$\frac{1}{n}\ln \mathbb{P}\left\{\cup_{i\in S_n}(X_l^i \leq 1-\nu^*)\right\} \leq -c_P$$

for all n sufficiently large, where c_P is some positive constant, and (IV.45) follows similarly to (IV.44). □

Proof of Lemma IV.8: By [3, Lemma A.5], we have for all $i \in S_n$,

$$\lim_{n\to\infty}\frac{1}{n}\ln \mathbb{P}(X_n^i \geq f) \leq -\frac{1}{2}\left[\ln(1+f) - \rho\ln f - H(\rho)\right]. \quad \text{(A.73)}$$

Union bounding $\mathbb{P}(X_n^i \geq f)$ over all $i \in S_n$ gives

$$\begin{aligned}\mathbb{P}\left\{\cup_{i\in S_n}(X_n^i \geq f)\right\} &\leq \sum_{i\in S_n}\mathbb{P}\left(X_n^i \geq f\right)\\ &= |S_n|\cdot\mathbb{P}(X_n^1 \geq f), \quad \text{(A.74)}\end{aligned}$$

Taking logarithms and limits of the right-hand side of (A.74), using (A.73) and (A.68), we have

$$\lim_{n\to\infty}\frac{1}{n}\ln\left[|S_n|\cdot\mathbb{P}(X_n^1 \geq f)\right] = d\rho\cdot H(d^{-1}) - \frac{1}{2}\left[\ln(1+f) - \rho\ln f - H(\rho)\right],$$

which combines with (A.74) to imply that, for any $\eta > 0$,

$$\frac{1}{n}\ln \mathbb{P}\left\{\cup_{i\in S_n}(X_n^i \geq f)\right\} \leq d\rho\cdot H(d^{-1}) - \frac{1}{2}\left[\ln(1+f) - \rho\ln f - H(\rho)\right] + \eta, \quad \text{(A.75)}$$

for all n sufficiently large. By the definition of $\mathcal{TIF}(\rho)$ in (IV.43), and since the left-hand side of (IV.43) on $f > \frac{\rho}{1-\rho}$ is strictly increasing in f, then, for any $\epsilon > 0$, setting $f := f^* = \mathcal{TIF}(\rho) + \epsilon$ and choosing η sufficiently small in (A.75) ensures

$$\frac{1}{n}\ln \mathbb{P}\left\{\cup_{i\in S_n}(X_n^i \geq f^*)\right\} \leq -c_I$$

for all n sufficiently large, where c_I is some positive constant, from which the result follows using the same argument as in the proof of lemma IV.7. □

REFERENCES

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

[2] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, pp. 1982–2001, 2010.

[3] C. Cartis and A. Thompson, "A new and improved quantitative recovery analysis for iterative hard thresholding algorithms in compressed sensing," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1–24, 2015.

[4] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction form highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[6] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Academic Press, 2009.

[7] R. Baraniuk and D. Jones, "A signal-dependent time-frequency representation: Fast algorithm for optimal kernel design," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3530–3535, 1994.

[8] C. Hegde, P. Indyk, and L. Schmidt, "A fast approximation algorithm for tree-sparse recovery," in *IEEE International Symposium on Information Theory*, June 2014, pp. 1842–1846.

[9] ——, "Nearly linear-time model-based compressive sensing," in *International Colloquium on Automata, Languages and Programming*, July 2014, pp. 588–599.

[10] C. Cartis and A. Thompson, "An exact tree projection algorithm for wavelets," *IEEE Signal Processing Letters*, vol. 20, no. 11, pp. 1028–1031, 2013.

[11] B. Bhan, L. Baldassare, and V. Cevher, "Tractability of interpretability via selection of group-sparse models," in *International Symposium on Information Theory*, July 2013, pp. 1037–1041.

[12] T. Blumensath and M. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and its Applications*, vol. 14, no. 5, pp. 629–654, 2008.

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

[14] M. Duarte, M. Wakin, and R. Baraniuk, "Fast reconstruction of piecewise smooth signals from random projections," in *Signal Processing with Adaptive Sparse Structured Representations*, November 2005.

[15] C. La and M. Do, "Signal reconstruction using sparse tree representation," in *SPIE Optics and Photonics*, vol. Wavelets XI, July 2005.

[16] M. Duarte, M. Wakin, and R. Baraniuk, "Wavelet-domain compressive signal reconstruction using a hidden markov tree model," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 5137–5140.

[17] A. Kyrillidis and V. Cevher, "Combinatorial selection and least absolute shrinkage via the CLASH algorithm," in *International Symposium on Information Theory*, July 2012, pp. 2216–2220.

[18] F. Bach, "Learning with submodular functions: a convex optimization perspective," *Foundations and Trends in Machine Learning*, vol. 6, no. 2-3, pp. 145–373, 2013.

[19] A. Kyrillidis, L. Baldasarre, M. El Halabi, Q. Tran-Dinh, and V. Cevher, "Structured sparsity: discrete and convex approaches," in *Compressed sensing and its applications*, ser. Applied and Numerical Harmonic Analysis, H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, Eds. Springer, 2015.

[20] P. Jain, N. Rao, and I. Dhillon, "Structured sparse regression via greedy hard-thresholding," 2016, http://arxiv.org/abs/1602.06042.

[21] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[22] C. Hegde, P. Indyk, and L. Schmidt, "Approximation algorithms for model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5129–5147, 2015.

[23] D. Donoho, "High-dimensional centrosymmetric polytopes with neighborliness proportional to dimension," *Discrete and Computational Geometry*, vol. 35, no. 4, pp. 617–652, 2006.

[24] J. Blanchard, C. Cartis, J. Tanner, and A. Thompson, "Phase transitions for greedy sparse approximation algorithms," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 188–203, 2011.

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

[26] S. Foucart, "Hard thresholding pursuit: an algorithm for compressive sensing," *SIAM Journal on Numerical Analysis*, vol. 49, no. 6, pp. 2543–2563, 2011.

[27] A. Thompson, "Quantitative analysis of algorithms for compressed signal recovery," Ph.D. dissertation, School of Mathematics, University of Edinburgh, 2012.
[28] T. Blumensath and M. Davies, "Normalised iterative hard thresholding: guaranteed stability and performance," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 298–309, 2010.
[29] J. Nocedal and S. Wright, *Numerical Optimization*. Springer, 1999.
[30] A. Kyrillidis and V. Cevher, "Recipes on hard thresholding methods," in *Computational Advances in Multi-Sensor Adaptive Processing*, December 2011.
[31] D. Donoho, "CART and best ortho-basis: A connection," *Annals of Statistics*, vol. 25, no. 5, pp. 1870–1911, 1997.
[32] J. Blanchard, C. Cartis, and J. Tanner, "Compressed sensing: How sharp is the restricted isometry property?" *SIAM Review*, vol. 53, no. 1, pp. 105–125, 2011.
[33] J. Blanchard and A. Thompson, "On support sizes of restricted isometry constants," *Applied and Computational Harmonic Analysis*, vol. 29, no. 3, pp. 382–390, 2010.
[34] R. Graham, D. Knuth, and O. Patashnik, *Concrete mathematics: a foundation for computer science*. Addison-Wesley, 1994.
[35] Y. Nesterov, *Introductory lectures on convex optimization : a basic course*, ser. Applied optimization. Boston, Dordrecht, London: Kluwer Academic, 2004.
[36] A. Agarwal, S. Negahban, and M. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *Neural Information Processing Systems*, 2010, pp. 37–45.
[37] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American Mathematical Society*, vol. 22, pp. 211–231, 2009.
[38] P. Indyk and E. Price, "K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance," in *43rd Annual ACM Symposium on Theory of Computing*, June 2011, pp. 627–636.
[39] C. Hegde, P. Indyk, and L. Schmidt, "Nearly linear-time model-based compressive sensing," in *Symposium on Discrete Algorithms*, January 2014, pp. 1544–1561.
[40] J. Blanchard, J. Tanner, and K. Wei, "CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion," *Information and Inference*, vol. 4, no. 4, pp. 289–327, 2015.
[41] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2008.
[42] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Birkhäuser, 2013.
[43] C. Cartis and A. Thompson, "Quantitative recovery guarantees for tree-based compressed sensing," Mathematical Institute, University of Oxford, Tech. Rep., August 2016, extended technical report.