\( \ell^1 \)-Analysis Minimization and Generalized (Co-)Sparsity: When Does Recovery Succeed?

Martin Genzel, Gitta Kutyniok, and Maximilian M"arz
Technische Universit"at Berlin, Department of Mathematics
Straße des 17. Juni 136, 10623 Berlin, Germany
E-Mail: [genzel, kutyniok, maerz]@math.tu-berlin.de

Abstract. This paper investigates the problem of signal estimation from undersampled noisy sub-Gaussian measurements under the assumption of a cosparse model. Based on generalized notions of sparsity, we derive novel recovery guarantees for the \( \ell^1 \)-analysis basis pursuit, enabling highly accurate predictions of its sample complexity. The corresponding bounds on the number of required measurements do explicitly depend on the Gram matrix of the analysis operator and therefore particularly account for its mutual coherence structure. Our findings defy conventional wisdom which promotes the sparsity of analysis coefficients as the crucial quantity to study. In fact, this common paradigm breaks down completely in many situations of practical interest, for instance, when applying a redundant (multilevel) frame as analysis prior. By extensive numerical experiments, we demonstrate that, in contrast, our theoretical sampling-rate bounds reliably capture the recovery capability of various examples, such as redundant Haar wavelets systems, total variation, or random frames. The proofs of our main results build upon recent achievements in the convex geometry of data mining problems. More precisely, we establish a sophisticated upper bound on the conic Gaussian mean width that is associated with the underlying \( \ell^1 \)-analysis polytope. Due to a novel localization argument, it turns out that the presented framework naturally extends to stable recovery, allowing us to incorporate compressible coefficient sequences as well.

Key words. Compressed sensing, cosparse modeling, analysis sparsity, \( \ell^1 \)-analysis basis pursuit, stable recovery, redundant frames, total variation, Gaussian mean width

1 Introduction

Initiated by the pioneering works of Candès, Donoho, Romberg, and Tao [CRT06; CT05; Don06], a considerable amount of research on compressed sensing during the last decade has dramatically changed our methodology for exploiting structure in many signal processing tasks. The classical setup in this field considers the problem of reconstructing an unknown sparse signal vector \( x^* \in \mathbb{R}^n \) from non-adaptive, linear measurements of the form

\[ y = Ax^* + e, \]

where \( A \in \mathbb{R}^{m \times n} \) is a known measurement matrix and \( e \in \mathbb{R}^m \) captures potential distortions, usually due to noise. The success of compressed sensing is based on the fundamental insight that, by explicitly incorporating a sparsity prior, this task becomes feasible even if \( m \ll n \). In particular, there exist numerous convex and greedy recovery methods that enjoy both efficient implementations and a rich theoretical foundation. Among them, probably the most popular approach is the basis pursuit [CDS98]:

\[ \min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{subject to} \quad \|Ax - y\|_2 \leq \eta, \] (BP_\eta)

where \( \eta \geq 0 \) is such that \( \|e\|_2 \leq \eta \). The crucial ingredient of \((BP_\eta)\) is the \( \ell^1 \) objective functional which promotes sparse solutions of the minimization program. Remarkably, it can be shown that \((BP_\eta)\) indeed recovers an \( S \)-sparse vector \( x^* \) with the optimal sampling rate of \( m = O(S \cdot \log(2S/n)) \) if the measurement design \( A \) is drawn according to an appropriate random distribution.

Although such types of theoretical guarantees are elegant and practically appealing, the traditional assumption of sparsity is not satisfied in most real-world applications. Fortunately, many signals-of-interest do at least exhibit a low-complexity representation with respect to a certain transformation, for instance, the wavelet or Fourier transform. In this work, we study the the so-called analysis sparsity model (also known as cosparse model), which has gained increasing attention within the past years [CENR11;
1.1 Does (Co-)Sparsity Explain the Success (or Failure) of the Analysis Basis Pursuit?

As already foreshadowed by (1.2), a central objective of the analysis formulation is to come up with an operator \( \Psi \in \mathbb{R}^{N \times n} \) that provides a coefficient vector \( \Psi x^* \) of “low complexity.” The traditional theory of compressed sensing—where the matrix \( \Psi := [\psi_1 | \ldots | \psi_N]^T \in \mathbb{R}^{N \times n} \) is called the analysis operator. If \( \Psi \) is able to reflect the underlying structure of \( x^* \), one might expect that these analysis coefficients are dominated by only a few large entries. This hypothesis of transform sparsity motivates the following generalization of (BP\( \eta \)) that is typically referred to as the analysis basis pursuit (or \( \ell^1 \)-analysis minimization):

\[
\min_{x \in \mathbb{R}^n} \| \Psi x \|_1 \quad \text{subject to} \quad \| Ax - y \|_2 \leq \eta. \tag{BP\( \eta \)}
\]

In fact, the adapted strategy of (BP\( \eta \)) has turned out to work surprisingly well for numerous problems settings, such as in total variation minimization [Cha04; ROF92] or for multiscale transforms in classical signal and image processing tasks [COS10; DPSS14; LDSP08; PM11]. Apart from that, it has become also relevant to regularizing physics-driven inverse problems [KABG16; NG12] and operator learning approaches based on test data sets [HKD13; RB11; RB13]. But despite this empirical success, many theoretical properties of the analysis basis pursuit remain unexplored and “its rigorous understanding is still in its infancy.” [KR15, p. 174]

The key idea of this approach is to test (“analyze”) the signal \( x^* \in \mathbb{R}^n \) with a collection of analysis vectors \( \psi_1, \ldots, \psi_N \in \mathbb{R}^n \), i.e., one computes

\[
\Psi x^* = ((\psi_1, x^*), \ldots, (\psi_N, x^*)) \in \mathbb{R}^N,
\]

where the matrix \( \Psi := [\psi_1 | \ldots | \psi_N]^T \in \mathbb{R}^{N \times n} \) is called the analysis operator. If \( \Psi \) is able to reflect the underlying structure of \( x^* \), one might expect that these analysis coefficients are dominated by only a few large entries. This hypothesis of transform sparsity motivates the following generalization of (BP\( \eta \)) that is typically referred to as the analysis basis pursuit (or \( \ell^1 \)-analysis minimization):

\[
\min_{x \in \mathbb{R}^n} \| \Psi x \|_1 \quad \text{subject to} \quad \| Ax - y \|_2 \leq \eta. \tag{BP\( \eta \)}
\]

Although many of those approaches rely on different proof strategies, e.g., the D-RIP [CENR11] or an adapted null space property [KR15], they eventually promote results of a very similar type: Recovery of \( x^* \in \mathbb{R}^n \) via (BP\( \eta \)) succeeds if the number of measurements obeys

\[
m \geq C \cdot S \cdot \text{PolyLog}(\frac{2N}{S}),
\]

where \( S := \| \Psi x^* \|_0 \) and \( C > 0 \) is a constant that might depend on \( \Psi \). This bound on the sampling rate clearly resembles classical guarantees for the basis pursuit (BP\( \eta \)) and therefore forms a quite natural extension towards analysis sparsity.

Another important branch of research takes a somewhat contrary perspective, and identifies the cosparse \( L := N - S \) as a crucial quantity for the success of the analysis model. A remarkable observation of [NDEG13] was that the location of vanishing coefficients in \( \Psi x^* \) is the driving force behind the analysis formulation, rather than the number of non-zero components. This viewpoint naturally leads to the so-called cosparse signal model, which is typically described by union-of-subspaces [BD09]. Following this terminology, it has turned out that successful recovery via combinatorial searching can be guaranteed if the number of observations is of the order of the signal’s manifold dimension [NDEG13, Sec. 3]. However, such a simple relationship does not seem to carry over to tractable methods like (BP\( \eta \)), see [GPV15]. Indeed, many sampling-rate bounds relying on cosparsivity can be easily translated into sparsity-based statements, simply meaning that \( S \) is replaced by \( N - L \) in (1.3), e.g., see [KR15, Thm. 1] or [GNEG14, Thm. 3.8].

Even tough the above approaches are quite appealing due to their simplicity, interpretability, and consistency, it still remains unclear whether they are sufficient for a sound foundation of (BP\( \eta \)) in its general form. The notions of analysis (co-)sparsity are merely determined by the support of \( \Psi x^* \), which in turn does not account for the coherence structure of the individual analysis vectors \( \psi_1, \ldots, \psi_N \); in other words, the underlying “geometry” of \( \Psi \) is ignored.
To get a first glimpse of this issue, let us consider a simple example: Figure 1 shows the results of a numerical simulation that reconstructs a block-signal (see Figure 1(a)) using three different analysis operators. The plot of Figure 1(b) exhibits the phase transition behavior of \((BP_{\Psi_{\eta=0}}\Psi_{rdwt})\) for a redundant, discrete Haar wavelet transform \(\Psi_{rdwt}\) and the analysis operator \(\Psi_{irdwt}\) associated with its inverse transform (i.e., a certain dual frame of \(\Psi_{rdwt}\)). Although the (co-)sparsity \((S = 906\) and \(L = 886\)) are exactly the same for both choices, their recovery capability indeed differs dramatically! In conclusion, just investigating the parameters \(S\) and \(L\) does by far not explain why the transition of \(\Psi_{irdwt}\) happens much earlier \((m \approx 85)\) than the one of \(\Psi_{rdwt}\) \((m \approx 240)\). Even more striking, the prediction of (1.3) deviates from the truth by orders of magnitudes.

The plot of Figure 1(c) reveals another insight: While (1.3) is reliable for an orthonormal Haar wavelet transform \(\Psi_{dwt} \in \mathbb{R}^{n \times n}\), a comparison with the actual recovery rate of \((BP_{\Psi_{\eta=0}}\Psi_{irdwt})\) indicates that redundancy can be beneficial in the analysis model. Finally, it is also worth mentioning that a compressibility argument does “save the day” here: Figure 1(d) demonstrates that the transition takes place far before the remaining coefficients would be negligibly small.

We emphasize that the above example is not too specific or artificial, but rather illustrates a scenario that often occurs in applications: Due to linear dependencies within \(\Psi\), the analysis sparsity oftentimes cannot become arbitrarily small. For instance, if \(\Psi\) corresponds to a highly redundant dictionary, one typically has \(S \gg n\), whereas the true sample complexity is relatively small (below the space dimension \(n\)). This important phenomenon gives rise to the following fundamental question:

> If (co-)sparsity does not entirely explain what is going on, which general principles make the analysis basis pursuit \((BP_{\Psi})\) succeed or fail?

To be a more precise about this concern, let us formulate three central issues that we will address in this paper:

(Q1) **Sampling rate.** How many measurements are needed for an accurate estimate of \(x^*\) via \((BP_{\Psi_{\eta}})\)? What crucial quantities and parameters determine the sample complexity?

(Q2) **Compressibility.** What if \(x^*\) is only compressible, i.e., its analysis coefficients \(\Psi x^*\) are not exactly sparse but only close to a sparse vector? Is \((BP_{\Psi_{\eta}})\) stable under such model inaccuracies?

(Q3) **Interpretability.** What practical guidelines can we derive from our recovery results? Which characteristics of \(\Psi\) deserve special attention in applications?

# 1.2 Main Contributions and Overview

A major concern of this work is to shed more light on the problem of \(\ell^1\)-analysis recovery and to provide a deeper understanding of its underlying mechanisms. Our first main result in Section 2 (Theorem 2.5) focuses on the setup of noiseless Gaussian measurements, i.e., \(e = 0\) in (1.1) and \(A\) is a standard Gaussian matrix. In this prototypical situation, we can even expect an exact reconstruction of \(x^*\) via \((BP_{\Psi_{\eta=0}})\), supposed that \(m\) is sufficiently large. In fact, Theorem 2.5 states a non-asymptotic and non-uniform bound on the sampling rate which intends to meet the following desiderata:

(i) **Accurate.** The number of required measurements should be close to the (optimal) sample complexity.

(ii) **Computable.** The expressions should be explicit with respect to the underlying analysis operator \(\Psi\) and numerically evaluable.

(iii) **Interpretable.** The involved parameters should have a clear meaning and resemble well-known principles.

(iv) **Generic.** The bound should not be tailored to a specific operator \(\Psi\) (e.g., wavelets or total variation), but apply in a general setting.
The key ingredients of our bound are three parameters, generalizing the notions of sparsity and cosparse (Definition 2.3). While the support of the coefficients $\Psi x^*$ still plays a big role, these novel quantities do also take account of the (mutual) coherence structure of $\Psi$ by incorporating the (off-diagonal) entries of its Gram matrix. Indeed, this refinement will allow us to satisfy the above “wish list” for many different examples of interest. Besides the well-known case of orthonormal bases, this particularly includes highly redundant analysis operators, which are only poorly understood in that context so far.

The proof of Theorem 2.5 (provided in Section 6) loosely follows a recent technique proposed by [ALMT14, Recipe 4.1]. More precisely, we establish a highly non-trivial upper bound on the Gaussian mean width$^1$ of the descent cone of $x \mapsto \|\Psi x\|_1$ at $x^*$ (Theorem 6.8), which is known to characterize the phase transition behavior of (BP) for Gaussian measurements. Noteworthy, Kabanava, Rauhut, and Zhang follow a very similar proof strategy in [KRZ15]. It is therefore not surprising that their

---

$^1$This is equivalent to bounding the so-called statistical dimension, introduced in [ALMT14]. But we will keep using the notion of ‘Gaussian mean width’, which is more common in the literature.
Let us fix some notations and conventions that will be frequently used in the remainder of this paper. Our second guarantee (Theorem 2.8) tackles the problem of stability (Q2) in the setting of noisy sub-Gaussian measurements. It is based on the simple geometric idea of approximating the ground truth $x^*$ by another signal vector $\hat{x}^* \in \mathbb{R}^n$ of lower complexity, in the sense that its associated Gaussian mean width is significantly smaller. This novel result is achieved by means of a localized variant of the (conic) Gaussian mean width [MPT07] which allows us to quantify the mismatch caused by approximating $x^*$. Let us emphasize that this approach goes beyond the “naive” reasoning for compressible vectors just suggests to determine the best $S$-term approximation of the analysis coefficient vector $\mathbf{Y}x^*$.

Section 3 presents several applications of the framework developed in Section 2. A special focus here is on the little studied example of redundant wavelet frames, but the variety of our results will be also demonstrated for other important choices of $\mathbf{Y}$, such as the total variation operator and random frames. In this course, we assess the predictive quality of our theoretical findings by means of various numerical experiments. Section 4 is then dedicated to an overview of the related literature, revisiting the initial concern of Subsection 1.1: Does (co-)sparsity explain the recovery performance of $(\mathbf{BP}_n)$? This discussion is concluded by Section 5, where we address the issues of (Q3) in greater detail: Even though the main objective of this work is a mathematical foundation, our novel perspective on (co-)sparsity might have practical implications as well, for instance, when designing or learning an analysis operator for a specific task. Finally, we point out some unresolved challenges left for future works, in particular, the question of how close to optimal our prediction bounds are.

### 1.3 General Notation

Let us fix some notations and conventions that will be frequently used in the remainder of this paper:

- For an integer $d \in \mathbb{N}$, we set $[d] := \{1, \ldots, d\}$. If $\mathcal{I} \subset [d]$, the set complement of $\mathcal{I}$ in $[d]$ is given by $\mathcal{I}^c := [d] \setminus \mathcal{I}$. Vectors and matrices are denoted by lower- and uppercase boldface letters, respectively. Unless states otherwise, their entries are indicated by subscript indices and lowercase letters, e.g., $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ for a vector and $\mathbf{B} = [b_{k,l}] \in \mathbb{R}^{d \times d^*}$ for a matrix. For $\mathcal{I} \subset [d]$, the vector $x_\mathcal{I} \in \mathbb{R}^{|\mathcal{I}|}$ is the restriction of $x \in \mathbb{R}^d$ to the components of $\mathcal{I}$. Similarly, $\mathbf{B}_\mathcal{I} \in \mathbb{R}^{|\mathcal{I}| \times d^*}$ restricts a matrix $\mathbf{B} \in \mathbb{R}^{d \times d^*}$ to the rows corresponding to $\mathcal{I}$.

- Let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The support of $x$ is defined by the set of its non-zero entries $\text{supp}(x) := \{k \in [d] \mid x_k \neq 0\}$ and the sparsity of $x$ is $\|x\|_0 := |\text{supp}(x)|$. For $1 \leq p \leq \infty$, we denote the $\ell^p$-norm on $\mathbb{R}^d$ by $\|x\|_p$. The associated unit ball is given by $B^p_\mathbb{R} := \{x \in \mathbb{R}^d \mid \|x\|_p \leq 1\}$ and the Euclidean unit sphere is $S^{d-1} := \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$. If $\Sigma^d_S := \{x \in \mathbb{R}^d \mid \|x\|_0 \leq S\}$ denotes the set of all $S$-sparse vectors in $\mathbb{R}^d$, the best $S$-term approximation error (with respect to the $\ell^p$-norm) of a vector $x' \in \mathbb{R}^d$ is

$$\sigma^S(x')_p := \min_{x \in \Sigma^d_S} \|x' - x\|_p.$$ 

- We denote the conic hull of a set $K \subset \mathbb{R}^d$ by $\text{cone}(K)$. If $U \subset \mathbb{R}^d$ is a linear subspace, the associated orthogonal projection onto $U$ is denoted by $P_U \in \mathbb{R}^{d \times d}$. Then, we have $P_{U^\perp} = I_d - P_U$, where $U^\perp \subset \mathbb{R}^d$ is the orthogonal complement of $U$ and $I_d \in \mathbb{R}^{d \times d}$ is the identity matrix.

- If $g$ is a mean-zero Gaussian random vector in $\mathbb{R}^d$ with covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, we just write $g \sim \mathcal{N}(0, \Sigma)$. Moreover, a random vector $a$ in $\mathbb{R}^d$ is sub-Gaussian if $\|a\|_{\psi_2} < \infty$, where $\|\cdot\|_{\psi_2}$ is the sub-Gaussian norm; see [Ver12, Def. 5.22].

- Let $\mathbf{Y} \in \mathbb{R}^{N \times n}$ be a matrix with row vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_N \in \mathbb{R}^n$. Then, the collection $\mathcal{F} := \{\mathbf{Y}_k\}_{k \in [N]} \subset \mathbb{R}^n$ forms a frame for $\mathbb{R}^n$ with frame bounds $0 < a \leq b < \infty$ if

$$a \cdot \|x\|_2^2 \leq \|\mathbf{Y}x\|_2^2 \leq b \cdot \|x\|_2^2$$ 

for all $x \in \mathbb{R}^n$. 

If there is no danger of confusion, we will identify the analysis operator $\Psi$ with $F$. Finally, $F$ is said to be tight if $a = b$. See [CK13] for a detailed introduction to finite-dimensional frame theory.

- For $s, t \in \mathbb{R}$, we define the clip function $\text{clip}(s; t) := \text{sign}(s) \cdot \min\{|s|, t\}$ and the positive part $[s]_+ := \max\{s, 0\}$.

- The letter $C$ is always reserved for a (generic) constant, whose value could change from time to time. We refer to $C$ as a numerical constant if its value does not depend on any other involved parameter. If an (in-)equality holds true up to a numerical constant $C$, we sometimes simply write $A \lesssim B$ instead of $A \leq C \cdot B$.

## 2 Main Results

In this section, we present the main theoretical results of this work. Starting with some notations and model assumptions in Subsection 2.1, our novel sparsity parameters are introduced in Subsection 2.2. Based on these central notions, Subsection 2.3 then establishes an exact recovery result for noiseless Gaussian observations. Finally, in Subsection 2.4, we show that the analysis basis pursuit ($\text{BP}_F^\Psi$) is also robust against noise and stable under model inaccuracies, meaning that the analysis coefficient vector is allowed to be compressible. For the sake of readability, all corresponding proofs are postponed to Section 6.

### 2.1 Model Setup and Notation

In this part, we state the standing assumptions and model hypothesis of our recovery framework; see also Table 1 for a summary. The following notations and conventions are supposed to hold true for the remainder of this section. Let us first recall the linear measurement scheme of (1.1) and set up a formal random observation model:

**Model 2.1 (Noisy Linear Measurements)** Let $x^* \in \mathbb{R}^n$ be a fixed vector, which is typically referred to as the signal (or source). The measurement vectors $a_1, \ldots, a_m \in \mathbb{R}^n$ are assumed to be independent copies of an isotropic, mean-zero, sub-Gaussian random vector $a$ in $\mathbb{R}^n$ with $\|a\|_{\psi_2} \leq \kappa$. These vectors form the rows of the measurement matrix $A := [a_1 \ldots a_m]^T \in \mathbb{R}^{m \times n}$. The actual measurements of $x^*$ are then given by

$$y := Ax^* + e \in \mathbb{R}^m,$$

where $e = (e_1, \ldots, e_m) \in \mathbb{R}^m$ models noise, which could be deterministic and systematic. We assume that the noise is $\ell^2$-bounded, i.e., $\|e\|_2 \leq \eta$ for some $\eta \geq 0$.

Next, we fix our notation for the analysis operator and coefficients. Note that the dimension of the analysis domain $\mathbb{R}^N$ does not necessarily have to be larger than the dimension of $\mathbb{R}^n$.

**Definition & Notation 2.2**

1. The matrix $\Psi = [\psi_k] \in \mathbb{R}^{N \times n}$ is called an analysis operator (or analysis matrix) if none of its rows equals the zero vector. The rows of $\Psi$, denoted by $\psi_1, \ldots, \psi_N \in \mathbb{R}^n$, are called the analysis vectors. Moreover, we define the Gram matrix of $\Psi$ as

$$G = [g_{k,k'}] := \Psi \Psi^T = [\langle \psi_k, \psi_{k'} \rangle] \in \mathbb{R}^{N \times N}.$$
The canonical choice of an orthonormal basis, it becomes actually clear why we speak of 'generalized'
which are both referred to as the
\[ \text{generalized cosparsity} \]
\[ \text{analysis support} \]
\[ \text{generalized sparsity} \]
\[ \text{define the} \]
\[ \text{(Generalized (Co-)Sparsity) Let} \]
\[ \text{which will form the heart of our sampling-rate bounds:} \]
\[ 2.2 \text{ Generalized (Co-)Sparsity} \]
\[ \text{terms} \]
\[ \text{and} \]
\[ \text{σ} = (\sigma_1, \ldots, \sigma_n) \]
\[ \text{where} \]
\[ \text{Note that, for the sake of readability, we have omitted the dependence on} \Psi \text{here. Considering} \]
\[ \text{the canonical choice of an orthonormal basis, it becomes actually clear why we speak of ‘generalized’} \]
\[ 1 \text{If there is no danger of confusion, we may omit the term ‘analysis’ and just speak of ‘coefficients’, ‘support’, ‘sparsity’, etc.} \]
\[ 2 \text{The label ‘<’ indicates that} L_{S^c_k} \text{only operates on the diagonal entries of} G. \]
sparsity: Since \( G = \Psi \Psi^T = I_n \) in this case, one has that \( S_{\sigma_x}^c = S = \| \Psi x \|_0 \) and \( L_{S_x}^c = L_{S_x} = L = n - S \); hence, the notions of Definition 2.3 precisely coincide with their traditional counterparts.

In general, this correspondence is more complicated. The definition of the generalized sparsity \( S_{\sigma_x}^c \) still operates on the analysis support of \( x \), but also involves a weighted sum over all Gram matrix entries associated with \( S_x \). The same holds true for the generalized cosparsity term \( L_{S_x}^c \), respectively. Such an incorporation of the off-diagonal entries of \( G \) is in fact quite appealing because, to a certain extent, it respects the mutual coherence structure of the analysis vectors \( \psi_1, \ldots, \psi_N \). Finally, let us state two basic observations, characterizing when the above sparsity parameters do not vanish (see also Lemma 6.11):

\[
S_{\sigma_x}^c = \| \Psi^T \sigma_x \|_2^2 > 0 \quad \text{if, and only if,} \quad x \notin \ker \Psi, \\
L_{S_x}^c > 0 \quad \text{if, and only if,} \quad S_x^c \neq \emptyset.
\] (2.2)

2.3 Sampling-Rate Function and Exact Recovery

To highlight the key features of our approach, we restrict our analysis to the simplified case of noiseless Gaussian observations in this part, i.e., \( \eta = 0 \) and \( a \sim \mathcal{N}(0, I_n) \) in Model 2.1. The analysis basis pursuit then takes the form

\[
\min_{x \in \mathbb{R}^n} \| \Psi x \|_1 \quad \text{subject to} \quad Ax = y, \quad (BP_{\Psi, \eta=0})
\]

and we can even hope for an exact retrieval of \( x^* \in \mathbb{R}^n \). Indeed, the recent work of [ALMT14] has made the remarkable observation that a convex program of this type typically undergoes a sharp phase transition as \( m \) varies: Recovery of \( x^* \) fails with overwhelmingly high probability if \( m \) is below a certain threshold, but once \( m \) exceeds a small transition region, recovery succeeds with overwhelmingly high probability; see Figure 1(b) for an example. This minimal number of required measurements (also depending on the desired probability of success) is often referred to as the sample complexity (or optimal sampling rate) of an estimation problem. However, computing such a quantity in an explicit way is usually a challenging task.

Our actual goal is therefore rather to come up with an upper bound on the sample complexity of \( (BP_{\Psi, \eta=0}) \) that is tight in many situations. For this purpose, let us introduce the following function, which essentially determines the sampling rate proposed by Theorem 2.5 below:
Definition 2.4 Let $\Psi \in \mathbb{R}^{N \times n}$ be an analysis operator and let $x^* \in \mathbb{R}^n$ with $x^* \notin \ker \Psi$. Then, we define the sampling-rate function of $\Psi$ and $x^*$ by

$$
\mathcal{M}(\Psi, x^*) := \begin{cases} 
\frac{n - (L_{S^c_{x^*}})^2}{L_S x^*} \cdot \Phi \left( \frac{S^c_{x^*} x^*}{L_S x^*} \right), & \text{if } S^c_{x^*} \neq \emptyset, \\
n, & \text{otherwise,}
\end{cases}
$$

where

$$
\Phi(\rho) := \text{erf} \left( \frac{\rho}{\sqrt{2}} \right), \quad \rho > 0,
$$

with

$$
h: (0, \infty) \to (0, \infty), \quad \tau \mapsto \sqrt{\frac{2}{\pi}} e^{-\tau^2/2} + \text{erf}(\frac{\tau}{\sqrt{2}}) - 1. \quad (2.3)
$$

It is not hard to verify that the univariate functions $h, h^{-1}$, and $\Phi$ —each one mapping from $(0, \infty)$ to $(0, \infty)$ —are well-defined; see first paragraph of Appendix A.3 for more details. Together with (2.2), this also implies the well-definedness of $\mathcal{M}(\Psi, x^*)$. Figure 2 shows the graph of $\Phi$, visualizing how the ratio $S^c_{x^*} x^* / L_S x^*$ affects the sampling-rate function.

With this notion at hand, we are now ready to state our first main result:

Theorem 2.5 (Exact Recovery via $(\text{BP}_{\eta=0})$) Assume that Model 2.1 is satisfied with $a \sim \mathcal{N}(0, I_n)$ and $\eta = 0$. Let $\Psi \in \mathbb{R}^{N \times n}$ be an analysis operator such that $x^* \notin \ker \Psi$. Then, for every $u > 0$, the following holds true with probability at least $1 - e^{-u^2/2}$: If the number of measurements obeys

$$
m > \left( \sqrt{\mathcal{M}(\Psi, x^*)} + u \right)^2 + 1, \quad (2.4)
$$

then $(\text{BP}_{\eta=0})$ recovers $x^*$ exactly.

Theorem 2.5 gives a first answer to (Q1): Roughly speaking, reconstruction succeeds with high probability as $m$ slightly exceeds the sampling-rate function $\mathcal{M}(\Psi, x^*)$. Since $\mathcal{M}(\Psi, x^*)$ is completely determined by our generalized sparsity parameters from Definition 2.3, we can even conclude that the bound (2.4) meets the desiderata of (ii), (iii), and (iv) requested in Subsection 1.2.

Unlike many approaches from the literature (cf. (1.3)), the statement of Theorem 2.5 is highly non-uniform. Indeed, the condition of (2.4) does not just involve the analysis sparsity $S = \|\Psi x^*\|_0$, but explicitly depends on the support $S_{x^*}$ as well as on the sign vector $\sigma_{x^*} = \text{sign}(\Psi x^*)$. We will return to this point in the course of our experiments in Section 3, demonstrating that the performance of $(\text{BP}_{\eta=0})$ is oftentimes not entirely explainable by means of $S$, even if $\Psi$ is fixed.

On the other hand, our wish for an almost tight bound on the sample complexity has not been clarified yet (see (i) in Subsection 1.2). This issue unfortunately turns out to be very challenging, and in general, we do not have a quantitative error estimate for our prediction. However, we will give at least numerical evidence in Section 3 and discuss some theoretical aspects of optimality in Subsection 6.3 (see Remark 6.15).

Remark 2.6 (1) The assumption of $x^* \notin \ker \Psi$ in Theorem 2.5 is not very restrictive and merely of technical nature. If $x^* \in \ker \Psi$, the analysis basis pursuit $(\text{BP}_{\eta=0})$ uniquely recovers $x^*$ if, and only if,

$$
\ker \Psi \cap (x^* + \ker A) = \{ x^* \}.
$$

In the setting of Gaussian measurements, where $\ker A$ can be identified almost surely with a random subspace of dimension $n - m$, the latter condition holds true if, and only if, $m \geq \dim(\ker \Psi)$. Thus, the dimension of $\ker \Psi$ precisely yields the sample complexity of $(\text{BP}_{\eta=0})$ in this case.
(2) An important feature of Theorem 2.5 is scaling invariance: Replacing $\Psi$ by a scaled version $\lambda \Psi$ with $\lambda \neq 0$ does not affect the sampling-rate function, i.e., $M(\lambda \Psi, x^\ast) = M(\Psi, x^\ast)$. This is due to the fact that the parameters $S_{S_{x^\ast}}^x$, $L_{S_{x^\ast}}$, and $L_{\tilde{S}_{x^\ast}}^\ast$ do only appear in terms of appropriate fractions.

(3) Since $M(\Psi, x^\ast) \leq n$, the condition of (2.4) does not lead to situations where $m$ needs to be much larger than $n$ in order to achieve successful recovery. Such a requirement would be overly restrictive because the equation system $y = Ax^\ast$ is almost surely uniquely solvable if $m \geq n$, no matter what operator $\Psi$ is applied. In contrast, this simple observation is not always reflected by a naive bound of the form (1.3), at least when the domain of analysis coefficients $\mathbb{R}^N$ is much higher dimensional, i.e., $N \gg n$.

The function $\Phi$ from Definition 2.4 is quite easy to understand from an analytical perspective, but it is however non-elementary. For that reason, let us conclude with a simplification of the bound (2.4) in Theorem 2.5. This obviously comes along with a certain loss of accuracy but sheds more light on the asymptotic behavior of our sampling-rate function $M(\Psi, x^\ast)$.

**Proposition 2.7** Let $\Psi \in \mathbb{R}^{N \times n}$ be an analysis operator and let $x^\ast \in \mathbb{R}^n$ with $x^\ast \notin \ker \Psi$ and $S_{x^\ast}^x \neq \emptyset$. Setting $S := S_{S_{x^\ast}}^x$, $L := L_{S_{x^\ast}}$, and $\tilde{L} := L_{\tilde{S}_{x^\ast}}^\ast$, it holds

$$M(\Psi, x^\ast) \leq \min \left\{ n - \frac{L^2}{\pi} + \left( \frac{L}{\pi} \right)^2 \cdot [2S \cdot \log(\frac{S + L}{2} + S)] + \frac{L^2}{\pi + 1}, n - \frac{L^2}{\pi} \cdot \frac{S}{2S + L} \right\}. \quad (2.5)$$

In particular, if $\Psi \in \mathbb{R}^{n \times n}$ forms an orthonormal basis, we have $S = \|\Psi x^\ast\|_0$ and $L = \tilde{L} = n = S$, so that

$$M(\Psi, x^\ast) \leq 2S \cdot \log(\frac{\pi}{2}) + 2S \lesssim S \cdot \log(\frac{2S}{S}). \quad (2.6)$$

The expression of (2.5) is indeed much more explicit than $M(\Psi, x^\ast)$. The left branch particularly reminds us of a well-known property of traditional bounds, where the sampling rate is supposed to grow linearly with $S$ up to a logarithmic factor. This intuition is also confirmed by (2.6), showing that our approach is consistent with the standard setup of compressed sensing. On the contrary, the right branch of (2.5) intends to mimics the asymptotics of $M(\Psi, x^\ast)$ if the value of $S_{S_{x^\ast}}^x$ is large compared to $L_{S_{x^\ast}}$, i.e., $S_{S_{x^\ast}}^x / L_{S_{x^\ast}} \gg 1$.

### 2.4 Stable and Robust Recovery

In this part, we turn to our wish for stability as stated in (Q2). For illustration, let us recall the above situation of exact recovery and inspect the sampling-rate function $x^\ast \mapsto M(\Psi, x^\ast)$ more closely. While this mapping actually insinuates a strong dependence on $x^\ast$, a brief look at our generalized sparsity parameters reveals that it is already determined by the sign vector $\sigma_{x^\ast}$. In particular, the latter quantity does not take account of the magnitudes of the analysis coefficients. The size of $M(\Psi, x^\ast)$ may therefore change dramatically (in a discontinuous manner) if $x^\ast$ is slightly varied in such a way that the zero entries of $\Psi x^\ast$ become negligibly small but do not vanish anymore. For a full coefficient vector $\Psi x^\ast$ with $S = N$, we would even have $M(\Psi, x^\ast) = n$, which typically occurs in real-world examples, such as Figure 3. In these situations, the “binary” statement proposed by Theorem 2.5 is too pessimistic, meaning that one cannot expect exact recovery with $m \ll n$. Instead of undergoing a sharp phase transition, the reconstruction error of $(BP_{\eta = 0})$ will then decay rather smoothly as $m$ grows.

Our approach to this problem relies on a simple geometric idea: Approximate $x^\ast \in \mathbb{R}^n$ by a vector $\hat{x}^\ast \in \mathbb{R}^n$ whose analysis coefficients $\Psi \hat{x}^\ast$ are much sparser than $\Psi x^\ast$. Such a “surrogate” of $x^\ast$ is supposed to be of lower complexity, in the sense that $M(\Psi, \hat{x}^\ast)$ is small, and if $\hat{x}^\ast$ and $x^\ast$ are not too distant, we speak of analysis compressibility. In that case, one may hope that the actual performance of $(BP_{\eta})$ is almost the same as if $\hat{x}^\ast$ would be the ground truth signal, i.e., $y = A\hat{x}^\ast$. This desirable property of a recovery method is usually referred to as stability.
A very natural, albeit naive, strategy to come up with a good low-complexity approximation of $x^*$ is to solve
\[
\min_{\hat{x} \in \mathbb{R}^n} \|x^* - \hat{x}\|_2 \quad \text{subject to} \quad |S_\hat{x}| = \|\Psi \hat{x}\|_0 \leq S
\] (2.7)
for some appropriately chosen sparsity threshold $S$. Any minimizer is then called a best $S$-analysis sparse approximation of $x^*$ (with respect to $\Psi$). Since the constraint set of (2.7) is actually a union of linear subspaces, it is also convenient to consider the following equivalent reformulation:
\[
\min_{S \subset \{1, \ldots, n\}, |S| \leq S} \|x^* - P_{U_S}x^*\|_2,
\] (2.8)
where $U_S := \ker \Psi_S^c \subset \mathbb{R}^n$. Interestingly, the same optimization program was recently studied in [GNEGD14] under the name of optimal projections, forming a crucial step in greedy-like algorithms for combinatorial cosparse modeling. Although (2.8) is tractable in some simple cases, it eventually turned out to be NP-hard in general [TGP14], due to its combinatorial nature. However, such algorithmic issues are only of minor importance to us, since we merely aim for a theoretical justification of stability in $\ell^1$-analysis recovery.

For these reasons, we shall state our second main result for Euclidean projections onto an arbitrary subspace $U \subset \mathbb{R}^n$, leaving space for applying less challenging (near-optimal) approximation methods. Note that, compared to Theorem 2.5, we now also allow for sub-Gaussian measurements as well as the presence of noise.

**Theorem 2.8** (Stable and Robust Recovery via (BP$^W$)) Suppose that Model 2.1 is satisfied and let $\Psi \in \mathbb{R}^{N \times n}$ be an analysis operator with $x^* \not\in \ker \Psi$. Let $U \subset \mathbb{R}^n$ be a linear subspace and assume that $P_U(x^*) \not\in \ker \Psi$. Then
\[
x^* := \frac{\|\Psi P_U x^*\|_1}{\|\Psi P_U x^*\|_1} \cdot P_U x^* \in \mathbb{R}^n
\] (2.9)
is well-defined and there exists a numerical constant $C > 0$ such that, for every $R > 0$ and $\eta > 0$, the following holds true with probability at least $1 - e^{-\eta^2/2}$: If the number of measurements obeys
\[
m > m_0 := C \cdot \kappa^4 \cdot \left(\frac{R+1}{R} \cdot \left[\sqrt{\mathcal{M}(\Psi, x^*)} + 1\right] + \eta\right)^2 + 1,
\] (2.10)
then any minimizer $\hat{x}$ of (BP$^W$) satisfies
\[
\|\hat{x} - x^*\|_2 \leq \max \left\{ R \cdot \|x^* - \hat{x}\|_2, \frac{2\eta}{\sqrt{m} - \sqrt{m_0 - 1}} \right\}.
\] (2.11)
In the Gaussian case of $a \sim \mathcal{N}(0, I_n)$, we particularly have $C = \kappa = 1$.

The statement of Theorem 2.8 makes the above geometric principle precise: Instead of evaluating the sampling-rate function $\mathcal{M}(\Psi, \cdot)$ at the ground truth $x^*$, we rather consider $\mathcal{M}(\Psi, \hat{x})$ in (2.10), which could be significantly smaller. The price to pay is an additional error term in (2.11), governed by the distance between $x^*$ and $\hat{x}$. Thus, one can formulate the following rule-of-thumb:

**Recovery accuracy can be exchanged with investigating less complex signals (taking fewer measurements), and vice versa.**

The benignity of this trade-off clearly depends on the ansatz space $U$. We already pointed out above that selecting $U$ is indeed very delicate and specific about the analysis operator $\Psi$. Moreover, even if one could identify an optimal support set $S$ according to (2.8), it is still not clear whether setting $U = U_S = \ker \Psi_S^c$ leads to the smallest sampling rate $\mathcal{M}(\Psi, x^*)$ among all $S$-analysis-sparse approximations of $x^*$. For these reasons, we defer a general discussion of this issue to future works. A simple

\footnote{One may also set $R = \infty$ here, with the convention that $\frac{R+1}{R} := 1$ and $R \cdot 0 := 0$.}
Figure 3: (a) Visualization of the ‘cameraman’ $x^* \in \mathbb{R}^{256 \times 256}$. (b) Sorted $\ell^1$-normalized magnitudes (in logarithmic scale) of the analysis coefficients $\Psi x^*$, where $\Psi$ corresponds to a compactly supported shearlet frame with four scales, provided by Shearlab [KLR16].

A greedy strategy is however presented in Appendix B.1, producing satisfactory results for redundant wavelets frames (see Subsection 3.1.4).

Compared to the setup of the previous subsection, the sub-Gaussianity of $A$ comes along with an extra factor of $C \cdot \kappa^4$ for the sampling rate, whereas the noise level $\eta$ affects the actual error estimate. Roughly speaking, (2.11) states that we can expect reliable outcomes as long as $\|e\|_2 / \sqrt{m} \leq \eta / \sqrt{m} = O(1)$, which is a well-known observation, e.g., see [Tro15, Cor. 3.5].

Finally, we would like to highlight that our approach is based on approximations in the signal domain $\mathbb{R}^n$. Working in the space of analysis coefficients $\mathbb{R}^N$ instead is actually more common in the literature (see Subsection 4.2), but suffers from disregarding the coherence structure of $\Psi$, which is in turn problematic in highly redundant scenarios. There is however a connection between both regimes, as the following bound for the approximation error in (2.11) shows:

$$\|x^* - \tilde{x}^*\|_2 \leq \|P_U x^*\|_2 \cdot \|\Psi P_U x^*\|_1 + \|P_U^\perp x^*\|_2,$$

see Proposition 6.7 for a proof. In fact, if $\Psi$ is an orthonormal basis and $U$ corresponds to the coordinate space of the $S$-largest entries of $\Psi x^*$ in magnitude, it is not hard to see that this estimate is closely related to traditional bounds from compressibility theory (cf. [FR13, Thm. 4.19]).

Remark 2.9 (1) The scaling factor in (2.9) ensures that $\|\Psi x^*\|_1 = \|\Psi \tilde{x}^*\|_1$. Hence, $\tilde{x}^*$ is not precisely equal to the optimal approximation of $x^*$ in $U$, unless $\|\Psi x^*\|_1 = \|\Psi P_U x^*\|_1$. This technicality is required for the proof of Theorem 2.8 (see Corollary 6.6), but we believe that it is not necessary in general.

(2) The parameter $R$ in Theorem 2.8 allows us to fine-tune the trade-off described above: Enlarging $R$ causes a decrease in accuracy in (2.11) but improves the sampling rate of (2.10) at the same time. In this sense, the fraction $(R + 1) / R$ can be also regarded as oversampling factor. ♦
3 Applications and Numerical Experiments

The purpose of this section is to thoroughly evaluate the predictive quality of the theorems presented in Section 2. A series of experiments based on redundant Haar wavelet frames (Subsection 3.1), total variation minimization (Subsection 3.2), and tight random frames in general position (Subsection 3.3) is conducted, which covers a wide range of analysis operators in one and two spatial dimensions (Subsection 3.4). By creating phase transition plots, we shall visualize the sampling-rate bound of Theorem 2.5 and investigate in how far our wish for accuracy is fulfilled. Combined with a simple greedy approximation strategy for compressible signals, this also allows for a verification of our result on stable recovery (Theorem 2.8).

While our theoretical framework includes a more general class of sub-Gaussian random matrices, we do only consider the benchmark case of noiseless Gaussian measurements here, as it is often done in the compressed sensing literature. Let us point out again that the statement of Theorem 2.5 is non-asymptotic so that it can be directly compared with the true recovery performance of (BP_{\eta=0}). In contrast, a visualization of asymptotic results, such as stated in [CENR11; NW13], is inappropriate for this setup due to unspecified numerical constants. The only other non-asymptotic bound, that we are aware of, was established by Kabanava, Rauhut, and Zhang in [KRZ15], and will therefore often serve as an object of comparison; see also Subsection 4.4 for a more formal presentation of their findings.

The notation of this part again follows Subsection 2.1, in particular Model 2.1. As it is common in the literature (cf. [ALMT14]), only the sampling-rate function \( x^* \mapsto M(\Psi, x^*) \) is reported when analyzing the quality of Theorem 2.5, whereas the probability parameter \( u \) in (2.4) is neglected. Unless stated otherwise, we are solving the convex program (BP_{\eta=0}) using the Matlab software package cvx [GB08; GB14] with the default settings in place and the precision set to best. An outcome \( \hat{x} \in \mathbb{R}^n \) is considered to be an ‘exact’ recovery of \( x^* \in \mathbb{R}^n \) if \( \| \hat{x} - x^* \|_2 < 10^{-5} \). Indeed, this threshold has proven to produce very stable solutions and seems to reflect the numerical accuracy of cvx.

3.1 (Highly) Redundant Wavelet Frames

A major difficulty in analysis modeling is to select an appropriate operator \( \Psi \in \mathbb{R}^{N \times n} \) that enables good reconstructions for a whole class of signals. One of the most popular classes is formed by piecewise constant functions (e.g., the blocks signal of [DJ94]), oftentimes regarded as a prototype system in the literature. In this context, translation-invariant Haar wavelet frames have turned out to be very useful for various signal- and image-processing tasks, most prominently denoising [CD95; DJ94]. Depending on the number of scaling levels, such wavelet transforms are typically highly redundant, i.e., the total number of wavelet coefficients \( N \) may be orders of magnitudes larger than the signal’s dimension \( n \). However, despite their success and popularity, there exists to the best of our knowledge no sound numerical or theoretical analysis of this important use case in respect of \( \ell^1 \)-analysis recovery.

We emphasize that the situation of high redundancy is not just a flaw of this specific setup, but is ubiquitous in applied harmonic analysis. For instance, when applying Gabor frames in time-frequency analysis, or (directional) representation systems like curvelets or shearlets in imaging, one often has to face problems with \( N/n \geq 50 \) [KLR16]. Similar observations were recently also made for approaches to analysis operator learning from training data: For example, [HKD13] reported that a redundancy factor of \( N/n = 128 \) yields the best results for many classical imaging tasks. Apart from being considerably redundant, these systems usually share the following properties:

- **Linear dependency.** There are (necessarily) linear dependencies among the analysis vectors of the frame.
- **Translation-invariance.** The analysis vectors are obtained by translating a family of generators.
- **Intrinsic localization.** The frame is well-localized in the sense of [FG05; Grö04]. Roughly speaking, this means that the off-diagonal decay of the Gram matrix \( G = \Psi \Psi^T \) is fast.
3.1.1 Setup and Notation of Wavelet Frames

Inspired by these observations, we build the simulations of this subsection on a combination of piece-wise constant signals and Haar wavelet systems, serving as benchmark example of (highly) redundant frames. Our specific implementation of wavelet transforms relies on the Matlab software package spot [BF13], which is in turn adapted from the Rice Wavelet Toolbox [B+17]. The redundant wavelet transform $\text{rdwt} : \mathbb{R}^n \rightarrow \mathbb{R}^N$ of this package is based on the so-called algorithme à trous. Here, the down-sampling steps of the discrete wavelet transform are omitted, whereas instead the filter responses are upsampled by padding with zeros, which reminds us of creating holes (‘trous’ in French) in the non-zero coefficients. In the following, let $\Psi_{\text{rdwt}} \in \mathbb{R}^{N \times n}$ denote the (matricized) analysis operator associated with this frame. The redundant forward transform of [BF13] comes along with an inverse wavelet transform $\text{idwt} : \mathbb{R}^N \rightarrow \mathbb{R}^n$ satisfying

$$\text{idwt} \circ \text{rdwt} = \text{Id}_{\mathbb{R}^n}. \tag{3.1}$$

From the viewpoint of frame theory, the application of $\text{idwt}$ corresponds to the synthesis operation with a dual frame of $\Psi_{\text{rdwt}}$. More precisely, if $\Psi_{\text{idwt}} \in \mathbb{R}^{n \times N}$ denotes the analysis operator of this dual frame, the rule (3.1) translates into $\Psi_{\text{idwt}}^\top \Psi_{\text{rdwt}} = I_n$. It is worth mentioning that $\Psi_{\text{rdwt}}$ and $\Psi_{\text{idwt}}$ are not merely academic concepts that are inaccessible in practice. Indeed, the analysis and synthesis operations of both frames are computable via the algorithm à trous with a complexity of $O(n \log(n))$.

Finally, if $n$ is a power of two, the non-redundant discrete wavelet transform $\text{dwt} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is orthonormal. Its associated analysis operator is denoted by $\Psi_{\text{dwt}} \in \mathbb{R}^{n \times n}$, forming an orthonormal basis for $\mathbb{R}^n$. Unless stated otherwise, we will always choose a Haar wavelet filter and 6 decomposition levels for $\Psi_{\text{rdwt}}$, $\Psi_{\text{idwt}}$, and $\Psi_{\text{dwt}}$. For a more detailed introduction to (redundant) wavelet transforms, the interested reader is referred to [Fow05; Mal09].

3.1.2 Recovery of blocks Signal

Our first experiment revisits the simulation of Figure 1 in Subsection 1.1, where the reconstruction of the classical blocks signal [DJ94] was investigated for different choices of Haar wavelet frames. Technically, we fix the ambient dimension $n = 256$, initialize $x^* \in \mathbb{R}^n$ as blocks, and run Experiment 3.1 below for $\Psi \in \{\Psi_{\text{rdwt}}, \Psi_{\text{idwt}}, \Psi_{\text{dwt}}\}$ and $M = \{1, \ldots, n\}$. The plots of Figure 1(b) and Figure 1(c) show the empirical success rates\(^1\) for $\Psi \in \{\Psi_{\text{rdwt}}, \Psi_{\text{idwt}}\}$ and $\Psi = \Psi_{\text{dwt}}$, respectively.

---

**Experiment 3.1 (Recovery of a fixed signal)**

**Input:** Fixed ambient dimension $n$, signal vector $x^* \in \mathbb{R}^n$, analysis operator $\Psi \in \mathbb{R}^{N \times n}$, range of measurements $M \subset \{1, 2, \ldots, n\}$.

**Compute:** Repeat the following procedure 50 times for every $m \in M$:

- Draw a standard i.i.d. Gaussian random matrix $A \in \mathbb{R}^{m \times n}$ and determine the measurement vector $y = Ax^*$.
- Solve the analysis basis pursuit (BP\(_{\Psi=0}\)) to obtain an estimator $\hat{x} \in \mathbb{R}^n$.
- Compute and store the recovery error $\|x^* - \hat{x}\|_2$.

Figure 1(b) reveals that a perfect reconstruction of $x^*$ with $\Psi_{\text{rdwt}}$ requires almost $m \approx n$ measurements. This observation is remarkable, since using $\Psi_{\text{rdwt}}$ for denoising via soft thresholding has proven to be very effective [CD95], whereas it seems to be quite inappropriate for the $\ell^1$-analysis formulation. In contrast, the choice of $\Psi_{\text{idwt}}$ performs significantly better, succeeding from $m \approx 105$

---

\(^1\)As pointed out above, ‘success’ means that $x^*$ is recovered up to an accuracy of $\|x^* - \hat{x}\|_2 < 10^{-5}$. 
onward. Interestingly, the same superiority remains true for analysis basis pursuit denoising (i.e., applying \((\text{BP}_{\Psi^*})\) with \(A = I_n\)), which indicates that the analysis approach is fundamentally different from classical denoising via soft thresholding.

The subsampled orthonormal basis \(\Psi_{\text{dwt}}\) behaves slightly worse than \(\Psi_{\text{irdwt}}\) and allows for retrieval with \(m \approx 125\) onward (see Figure 1(c)). This gives further evidence that the redundancy of a sparsifying transform can help to improve the recovery capability of analysis-based priors.

A closer look at the structure of \(\Psi_{\text{rdwt}}\) and \(\Psi_{\text{irdwt}}\) reveals that they actually correspond to the same frame up to a reweighting of their analysis vectors. In fact, it holds

\[
\psi_{\text{rdwt}, k} = \frac{\psi_{\text{irdwt}, k}}{\|\psi_{\text{irdwt}, k}\|_2} \in \mathbb{R}^n
\]

for \(k = 1, \ldots, N\), particularly implying that the respective (co-)supports \(S_{x^*}\) and \(S_{c_{x^*}}\) coincide in both cases. Following traditional approaches based on analysis sparsity [CENR11; KR15] or cosparse modeling [NDEG13], one would therefore conclude that the number of samples required for exact recovery should be equal as well.

Since this is far from being true, one might be tempted to take the perspective of analysis compressibility, instead of insisting upon perfect sparsity. However, the sharp phase transition in Figure 1(b) as well as the decay of the analysis coefficients in Figure 1(d) show that this principle does not provide a valid explanation. Indeed, even the most accurate and explicit bound known from the literature [KR15, Thm. 2] predicts an error rate of

\[
\|x^* - \hat{x}\|_2 \leq \frac{2\sigma_5(\Psi x^*)_1}{\sqrt{a} \cdot S},
\]

supposed that, roughly speaking, \(m \geq 10 \cdot \frac{(b/a) \cdot S \cdot \log(N/S)}{10}\). Here, \(a\) and \(b\) are lower and upper frame bounds of \(\Psi\), respectively, and \(\sigma_5(\Psi x^*)_1\) denotes the best \(S\)-term approximation error of \(\Psi x^*\) (cf. Subsection 1.3). The visualization of (3.3) in Figure 4 indicates that a reliable reconstruction can be only expected if \(m\) exceeds the ambient dimension \(n\) by a factor of 3, contradicting the actual performance of \((\text{BP}_{\Psi^*})\). Moreover, the plots do even certify a superiority of \(\Psi_{\text{rdwt}}\) over \(\Psi_{\text{irdwt}}\), whereas the complete opposite is true. We finally would like to point out that an argumentation based on the matrix condition is also misleading in this setup because both matrices \(\Psi_{\text{rdwt}}\) and \(\Psi_{\text{irdwt}}\) actually have the same condition number (which equals the square root of the frame bound ratio \(b/a\)); see Table 2.

Turning towards the predictive quality of Theorem 2.5, we observe in Figure 1(b) that our bound on the sampling rate very precisely captures the location of the phase transitions for both \(\Psi_{\text{rdwt}}\) and \(\Psi_{\text{irdwt}}\). While \(\mathcal{M}(\Psi_{\text{rdwt}}, x^*)\) almost perfectly hits the center of transition curve (50% success rate),
Table 2: Characteristic parameters of the simulation in Figure 1. Here, $a$ and $b$ denote the lower and upper frame bound of $\Psi$, respectively. The (co-)sparsity of the analysis coefficients $\Psi x^r$ are $S = |S^\Psi|$ and $L = |S^\Psi_\sigma| = N - S$, whereas $S^\Psi_\sigma$, $L^\Psi_\sigma$, and $L^\Psi_\sigma$ correspond to the generalized sparsity terms from Definition 2.3. The sampling-rate function $M(\Psi, x^r)$ is defined according to Definition 2.4. The true sample complexity is denoted by $\delta(\Psi, x^r)$, see Appendix B.2. Note that all values are rounded.

| Ana. op. $\Psi$ | Scales | $a$ | $b$ | $b/a$ | $N$ | $S$ | $L$ | $S^\Psi_S$ | $L^\Psi_S$ | $M(\Psi, x^r)$ | $\delta(\Psi, x^r)$ |
|-----------------|--------|-----|-----|-------|-----|-----|-----|-----------|-----------|----------------|-----------------|
| $\Psi_{\text{dwt}}$ | 6      | 1   | 1   | 1     | 256 | 41  | 215 | 215       | 215       | 114            | 114             |
| $\Psi_{\text{irdwt}}$ | 6      | 2   | 64  | 32    | 1792| 906 | 886 | 28462     | 5180      | 241            | 240             |
| $\Psi_{\text{irdwt}}$ | $\frac{1}{64}$ | $\frac{1}{5}$ | 32  | 1792 | 906  | 886 | 33  | 197       | 212       | 100            | 84              |
| $\Psi_{\text{dwt}}$ | 3      | $\frac{1}{8}$ | $\frac{1}{5}$ | 4     | 1024 | 306 | 718 | 39        | 194       | 209            | 109             |

The capacity of the analysis basis pursuit using highly redundant frames is not solely captured by (co-)sparsity. In contrast, the sampling-rate bound proposed by Theorem 2.5 reliably states when recovery is possible and when not.

### 3.1.3 Asymptotic Behavior of the Sampling Rate in $n$

In the setup of the previous subsection, we have fixed both the ambient dimension $n$ and the ground truth $x^r \in \mathbb{R}^n$. Moreover, changing $n$ does also require an adaptation of the analysis operator $\Psi \in \mathbb{R}^{N \times n}$, so that the value of $M(\Psi, x^r)$ is affected in a highly non-trivial way. For these reasons, we next investigate the accuracy of $M(\Psi, x^r)$ if $n$ is varied, while $x^r$ and $\Psi$ are just adapted in terms of different resolution levels. Since the theoretical guarantee of [KRZ15, Thm. 3.3] is of a similar type, it is also reported in our simulation below.

In order to create the plots of Figure 5, we again consider a combination of $x^r \in \mathbb{R}^n$ as the blocks signal [DJ94] and $\Psi_{\text{irdwt}} \in \mathbb{R}^{N \times n}$ as analysis operator, where $n \in \{64, 128, 256, 512, 1024, 2048, 4096, 8192\}$. Although the latter range is restricted to powers of two for technical reason, Figure 5(a) indicates that both $M(\Psi_{\text{irdwt}}, x^r)$ and $M(\Psi_{\text{dwt}}, x^r) = \delta(\Psi_{\text{dwt}}, x^r)$ scale almost logarithmically in $n$. Compared to that, the prediction of [KRZ15, Thm. 3.3] rather exhibits a linear growth. To assess the mismatch of our bound with the true sample complexity, Figure 5(b) visualizes the relative error

$$\frac{M(\Psi_{\text{irdwt}}, x^r) - \delta(\Psi_{\text{irdwt}}, x^r)}{n}$$

showing that the error curve drops to $\sim 4\%$ for increasingly higher dimensions.

We emphasize that the setup of Figure 5 considers an asymptotic model: While the “resolution” of $\Psi$ gets increasingly finer as $n \to \infty$, the signal $x^r \in \mathbb{R}^n$ is always generated by discretizing a piecewise constant function consisting of exactly 11 discontinuities. Hence, one would expect that the complexity of $x^r$ is only slightly affected, and indeed, $\delta(\Psi_{\text{irdwt}}, x^r)$ and $M(\Psi_{\text{dwt}}, x^r)$ grow both much more slowly than $n$. But in certain sense, this experiment pushes our theoretical approach to its limits because it is rather designed for a non-asymptotic setting, where $n$ remains fixed. Let us briefly summarize our findings:

---

1The true sample complexity is numerically computed by means of the statistical dimension [ALMT14], see Appendix B.2 for details.
The proposed sampling-rate of Theorem 2.5 scales almost logarithmically in the ambient dimension, resembling the asymptotics of traditional bounds from compressed sensing. The relative deviation from the true sample complexity appears to remain constant for larger values of $n$.

### 3.1.4 Stable Recovery of Compressible Signals

As discussed in the course of Subsection 2.4, the significance of Theorem 2.5 relies on the assumption that sufficiently many analysis coefficients are zero. However, perfect sparsity is rarely satisfied in practice, since real-world signals are usually only compressible (cf. Figure 3). Although one cannot expect exact recovery in these cases, it still makes sense to ask for an approximate reconstruction of $x^* \in \mathbb{R}^n$ if $m$ is reasonably large, corresponding to a smooth error decay. In this part, we intend to verify this heuristic numerically and demonstrate how the stability result of Theorem 2.8 can be applied to such situations.

For this purpose, we continue to use $\Psi_{\text{irdwt}} \in \mathbb{R}^{N \times n}$ as analysis operator, whereas the `blocks` signal clearly needs to be modified in order to achieve a compressible coefficient vector with respect to $\Psi_{\text{irdwt}}$. Our adapted choice of the source signal, denoted by $x^*_{\text{c}} \in \mathbb{R}^n$, is displayed in Figure 6(a) for $n = 256$. More precisely, it is constructed by replacing one of the piecewise constant sections of `blocks` with a smoother segment, such that the number of non-zero coefficients increases from 906 to 1140, i.e., $\|\Psi_{\text{irdwt}}x^*_{\text{c}}\|_0 = 1140$. Figure 6(b) shows that this modification particularly results in a larger region of rapidly decaying analysis coefficients (between 800 and 1140). For a more appropriate comparison, we do not consider the original `blocks` signal as “sparse” counterpart, but define a simplified version $x^*_{\text{sp}} \in \mathbb{R}^n$ which does not have any discontinuities within the curvy segment of $x^*_{\text{c}}$, see Figure 6(a) for $x^*_{\text{sp}}$ and Figure 1(a) for `blocks`.

The performance of the analysis basis pursuit ($\text{BP}_{\|\cdot\|_0 = 0}^{\Psi_{\text{irdwt}}}$) is visualized in Figure 6(c): Here, we have applied the instructions of Experiment 3.1 to both $x^*_{\text{c}}$ and $x^*_{\text{sp}}$ with $M = \{80, 81, \ldots, 160\}$ and report the averaged reconstruction errors. Similar to Figure 1(b), the estimation error of $x^*_{\text{sp}}$ drops to zero at $m \approx 95$. In contrast, $x^*_{\text{c}}$ is not exactly recovered before $m \approx 150$, but starting from $m \approx 100$, the error curve however smoothly tends to zero. This underpins our intuition that one cannot expect a perfect, but still an accurate, outcome for compressible signals. For illustration, we have also plotted two exemplary reconstructions of $x^*_{\text{c}}$ via ($\text{BP}_{\|\cdot\|_0 = 0}^{\Psi_{\text{irdwt}}}$) in Figure 6(e) and Figure 6(f) for $m = 100$ and $m = 125$, respectively.

To invoke our theoretical framework from Subsection 2.4, we first need to come up with a meaning-
Figure 6: (a) Compressible signal $x_c^*$ and its piecewise constant counterpart $x_{sp}^*$. The projection $P_{U_k}x_c^*$ yields a low-complexity approximation of $x_c^*$. (b) Sorted $\ell^1$-normalized magnitudes of the analysis coefficients $\Psi_{irdwt}x_c^*$ and $\Psi_{irdwt}x_{sp}^*$. (c) Averaged recovery error via (BP$_{\Psi_{irdwt}}$)$_{\eta=0}$. (d) Relative recovery error predicted by Theorem 2.8 (see Experiment 3.2). (e) + (f) Minimizers $\hat{x}$ of (BP$_{\Psi_{irdwt}}$)$_{\eta=0}$ with input $y = Ax_c^* \in \mathbb{R}^m$. 

3 APPLICATIONS AND NUMERICAL EXPERIMENTS
ful approximation strategy for $x^*_c$. More precisely, our task is to identify a subspace $U_S \subset \mathbb{R}^n$, such that the rescaled projection (cf. (2.9))

$$\bar{x}_c^* := \frac{\|\iota_{\text{irdvt}} x^*_c\|_1}{\|\iota_{\text{irdvt}} P_{U_S} x^*_c\|_1} \cdot P_{U_S} x_c^* \in \mathbb{R}^n$$

is both of lower complexity and the resulting approximation error $\|x^*_c - \bar{x}_c^*\|_2$ is small. If $U_S$ is appropriately chosen, Theorem 2.8 certifies accurate recovery via $\text{(BP}_{\Psi}=0\text{)}$, while the involved sampling-rate function $\mathcal{M}(\iota_{\text{irdvt}}, \bar{x}_c^*)$ is of moderate size. As already pointed out in the course of (2.8), finding a best $S$-analysis-sparse approximation of $x^*_c$ with respect to $\iota_{\text{irdvt}}$ is not straightforward and may lead to an NP-hard problem. Hence, we merely propose a greedy method in Appendix B.1: Roughly speaking, Algorithm B.1 provides a subspace $U_S \subset \mathbb{R}^n$ for a given $S \in [N]$ that tries to meet the above two criteria, even though this choice might be suboptimal. Figure 6(a) shows the “evolution” of our approach by plotting an intermediate approximation $P_{U_S} x_c^*$ with $S = \|\iota_{\text{irdvt}} P_{U_S} x_c^*\|_0 = 900$. The algorithm indeed first targets the area of highest curvature, approximating $x^*_c$ by piecewise constant sections here, so that the analysis coefficients of $x^*_c$ become sparser. However, we suspect that an optimal projection would rather replace the smooth segment in a more “zig-zag” like fashion.

To verify the statement of Theorem 2.8, we apply the steps of Experiment 3.2 below. Note that this template permits any subspace $U_S$ with $\|P_{U_S} x^*_c\|_0 \leq S$.

### Experiment 3.2 (Approximation of a compressible signal)

**Input:** Signal vector $x^*_c \in \mathbb{R}^{256}$, analysis operator $\iota_{\text{irdvt}} \in \mathbb{R}^{1792 \times 256}$, approximation subspace $U_S \subset \mathbb{R}^n$ for every $S \in [N]$.

**Compute:** Repeat the following procedure 50 times for every $m \in \{80, 81, \ldots, 160\}$:

- Starting with $S = 1140$, decrease $S$ until $\mathcal{M}(\iota_{\text{irdvt}}, P_{U_S} x_c^*) \leq m$ is satisfied.
- Determine $\bar{x}_c^* := \frac{\|\iota_{\text{irdvt}} x^*_c\|_1}{\|\iota_{\text{irdvt}} P_{U_S} x^*_c\|_1} \cdot P_{U_S} x_c^*$ according to (2.9).
- Compute and store the approximation error $\|x^*_c - \bar{x}_c^*\|_2$.

The results of this simulation are presented in Figure 6(d), plotting the error curve over the selected range for $m$. To illustrate the benefit of our greedy method, Algorithm B.1, we have repeated the instructions of Experiment 3.2 using a standard approximation strategy, i.e., one sets $U_S := \ker(\iota_{\text{irdvt}}|_{S^c})$ where $S$ just corresponds to the $S$ largest coefficients of $\iota_{\text{irdvt}} x^*_c$ in magnitude. Note that the plots of Figure 6(d) are only suited for a relative comparison between both approaches. In fact, we have disregarded the impact of the tuning parameter $R$ in Theorem 2.8, so that the respective curves do rather describe the qualitative behavior of the error bound (2.11). This particularly explains why the labels on the vertical axis are omitted in Figure 6(d).

For $m \geq 100$ measurements, the predicted curve based on the greedy choice of $U_S$ strongly resembles the true recovery error in Figure 6(c). This stands in contrast to the naive approach (‘largest coeff.’), which is apparently not able to exploit the compressibility of $x^*_c$. The unfavorable “jump” at $m \approx 138$ is due to the ignorance of linear dependencies within $\iota_{\text{irdvt}}$. For example, when enforcing that

$$\langle P_{U_S} x^*_c, \psi_k \rangle = 0,$$

for a high-scale wavelet frame vector $\psi_k \in \mathbb{R}^n$ of $\iota_{\text{irdvt}}$, the same orthogonality relation does automatically hold true for many more coefficients at lower scales. Such a clustering of coefficient eventually leads to poor $\ell^2$-approximations of $x^*_c$. Our greedy selection procedure avoids this drawback by implicitly respecting the multilevel structure of wavelets; see also wavelet footprints [DV03]. Let us again draw a brief conclusion:
The recovery error of analysis compressible signals typically tends smoothly to zero as m grows. The stability result of Theorem 2.8, combined with an appropriate greedy approximation scheme, adequately reflects this decay behavior.

3.1.5 Phase Transition for Piecewise Constant Signals

Up to now, we have only analyzed a specific signal (blocks) with respect to certain Haar wavelet frames. Although the resolution of $x^* \in \mathbb{R}^n$ was adapted to different dimensions $n$, its geometric features (the number of piecewise constant segments) remained untouched. On the other hand, it is also of interest to assess how well our main result, Theorem 2.5, predicts the phase transition of (BP$_{\Psi=0}$) if the “complexity” of $x^*$ is changed. In the classical situation of an orthonormal analysis operator (e.g., $\Psi_{\text{dwt}}$), the success of recovery is completely determined by the sparsity $S = \|\Psi x^*\|_0$, implying that it is quite natural to create phase transition plots over $S$, see [ALMT14]. But such a simple indicator for complexity does often not exist for (highly) redundant systems, which is in fact one of the key findings of this work. However, recalling that wavelets were specifically developed for the detection of singularities, one may argue that the number of jumps characterizes a signal much better when working, for example, with $\Psi_{\text{irdwt}}$. The following experiment shows that this heuristic of total variation sparsity (TV-sparsity)$^1$ indeed serves as an appropriate surrogate of analysis sparsity in the context of piecewise constant functions.

Our first simulation is generated according to the following template with $n = 128$, $\Psi = \Psi_{\text{irdwt}}$, and $\mathcal{S} = \{1, 2, \ldots, 128\}$:

| Experiment 3.3 (Phase transition for piecewise constant signals) |
|---------------------------------------------------------------|
| **Input:** Fixed ambient dimension $n$, analysis operator $\Psi \in \mathbb{R}^{N \times n}$, range of TV-sparsity $\mathcal{S} \subseteq \{0, 1, \ldots, n - 1\}$. |
| **Compute:** Repeat the following procedure 50 times for every $S_{\text{TV}} \in \mathcal{S}$: |
| - Select a random set $S_{\text{TV}} \subseteq [n - 1]$ with $|S_{\text{TV}}| = S_{\text{TV}}$ and determine an orthonormal basis $B$ of ker$[(\Psi_{\text{TV-1}})_{S_{\text{TV}}}^t]$. Then draw a standard Gaussian random vector $c$ and set $x^* := Bc$. |
| - Repeat the following steps 10 times for each $m \in \{1, 2, \ldots, n\}$: |
| | - Draw a standard i.i.d. Gaussian random matrix $A \in \mathbb{R}^{m \times n}$ and determine the measurement vector $y = Ax^*$. |
| | - Solve the analysis basis pursuit (BP$^\Psi_{\Psi=0}$) to obtain an estimator $\hat{x} \in \mathbb{R}^n$. |
| | - Declare success if $\|x^* - \hat{x}\|_2 < 10^{-5}$. |

Two examples of the signal-generation step in Experiment 3.3 are shown in Figure 7. The actual phase transition plot of Figure 8(a) is then created by computing the empirical mean of the success rates from Experiment 3.3. The annotated curves visualize the sampling-rate bounds proposed by Theorem 2.5 and [KRZ15, Thm. 3.3], respectively. For the second run in Figure 8(b), we have invoked Experiment 3.3 with higher dimension $n = 512$ and $\Psi = \Psi_{\text{irdwt}}$, but using a slightly coarser grid $\mathcal{S} = \{1, 5, 9, \ldots, 509\}$ to reduce the computational burden.

Interestingly, the results of Figure 8 do strongly resemble classical phase transitions, e.g., as reported in [ALMT14]. This observation is somewhat surprising because the (averaged) coefficient sparsity $S = \|\Psi_{\text{irdwt}}x^*\|_0$, displayed on the top of the plots in Figure 8, appears detached in our setting. Regarding accuracy, we can conclude that $\mathcal{M}(\Psi_{\text{irdwt}}, x^*)$ captures the location of the phase transition fairly well, whereas [KRZ15, Thm. 3.3] provides a much worse prediction. Thus, the key message of this subsection reads as follows:

$^1$The TV-sparsity (or gradient sparsity) of $x^* \in \mathbb{R}^n$ is given by $S_{\text{TV}} := \|\Psi_{\text{TV}}x^*\|_0$, where $\Psi_{\text{TV}} \in \mathbb{R}^{(n - 1) \times n}$ is the standard total variation operator in 1D. For a precise definition, see Subsection 3.2.
Figure 7: Two signal vectors $\mathbf{x}^* \in \mathbb{R}^n$ generated by the first step of Experiment 3.3 with different choices of TV-sparsity.

Figure 8: Phase transition plots for piecewise constant signals analyzed with $\Psi_{\text{irdwt}}$ ($S_{TV}$ is the number of discontinuities of $\mathbf{x}^* \in \mathbb{R}^n$). The blue curve is obtained by computing the empirical mean of the sampling-rate function $M(\Psi_{\text{irdwt}}, \mathbf{x}^*)$ within each iteration step $S_{TV} \in S$ of Experiment 3.3. The same procedure was performed for the orange curve corresponding to [KRZ15, Thm. 3.3]. The number of discontinuities governs the complexity of piecewise constant signals, enabling us to create appropriate phase transition plots for redundant Haar wavelet frames. The resulting transition curves are very accurately described by our sampling-rate bound in Theorem 2.5.

Remark 3.4 At this point, it is also worth revisiting Subsection 3.1.3: The simulation of Figure 5 investigates the behavior of the sample complexity when the TV-sparsity of $\mathbf{x}^* \in \mathbb{R}^n$ remains constant while $n$ grows. Conceptually, this corresponds to invoking Experiment 3.3 for very large values of $n$ and studying the left end of the resulting phase transition plot. The relative error in Figure 5(b) therefore particularly reflects the small deviation of our prediction from the truth in Figure 8 if $S_{TV}$ is small. In practice, however, the choice of $\Psi$ is usually adapted to the signal’s resolution level, implying that one is rather interested in those vertical cross sections of transition plots for which the ratio $S_{TV}/n$ is of moderate size. ♦
3.2 Total Variation

In this section, we consider a fundamentally different, yet classical example of an analysis operator, namely total variation in 1D (TV-1). It originates from the seminal work of [ROF92], investigating the problem of signal denoising. Although conceptually quite simple, total variation has proven to be a highly effective prior in regularizing inverse problems and therefore became one of the most popular analysis operators used in practice.

Perhaps, the most striking structural difference to the wavelet-based approach of the previous subsection is that the total variation operator

$$\Psi_{TV-1} := \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 \\ 0 & -1 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots \\ 0 & \ldots & 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1) \times n}$$

do not constitute a frame for $\mathbb{R}^n$. Note that there exist numerous variants of total variation imposing different boundary conditions. The above choice uses forward differences with von Neumann boundary conditions and is common in the field of compressed sensing.

To assess the quality of Theorem 2.5, it makes sense to follow the same strategy as in Subsection 3.1.5. Indeed, the TV-based basis pursuit (BP$\Psi_{TV-1}$) promotes piecewise constant output signals, which indicates that TV-sparsity is again the right quantity to study. Our first phase transition plot in Figure 9(a) is generated according to Experiment 3.3, with $n = 300$, $\Psi = \Psi_{TV-1}$ and $S = \{1, 4, \ldots, 298\}$. The second simulation in Figure 9(b) repeats the same procedure, but the first step of Experiment 3.3 is now replaced by

- Generate $x^* \in \mathbb{R}^n$ by setting $x^*_j = (-1)^{j-1}$ for $j = 1, 2, \ldots, S_{TV}$ and $x^*_j = 0$ otherwise.

Even though the respective TV-sparsity coincides in both signal-generation steps, the two transition curves of Figure 9 look somewhat different. This observation has a remarkable implication: While TV-sparsity is usually considered to be a heuristic measure of complexity for piecewise constant functions, it does not entirely characterize the capacity of (BP$\Psi_{TV-1}$). Similar to the example of wavelets in Subsection 3.1, we therefore suspect that a sound analysis of the total variation operator must be
non-uniform to a certain extent, depending on the specific structure of $x^*$. In fact, the sampling-rate function $x^* \mapsto M(\Psi_{TV,1}, x^*)$ seems to meet this desire for non-uniformity, since the shape of the blue curves in Figure 9 adapts to each of the cases. But we have to clearly confess that our predictions are only reliable in the regimes of mid- and high-level sparsity. In contrast, the bound of [KRZ15, Thm. 3.1] is not very significant and particularly remains unchanged for both signal classes. Let us conclude our discussion as follows:

Total variation minimization exhibits a non-uniform recovery behavior, which is also captured by the sampling-rate bound of Theorem 2.5. However, if the TV-sparsity is very small, our prediction is much less accurate than for redundant wavelet frames.

Remark 3.5 (Related Literature) Compared to redundant wavelet frames, the $\ell^1$-analysis formulation of total variation is actually much better theoretically understood. For example, the work of Cai and Xu [CX15] proves that the optimal sampling rate of $(BP_{TV})$ is given by $m = \Theta(\sqrt{\|\Psi_{TV} \cdot x\|_0})$ for Gaussian measurements, where $S_{TV} = \|\Psi_{TV} \cdot x\|_0$ and log-terms are ignored. The corresponding proofs are also based on estimating the Gaussian mean width, although using very different techniques that are specifically tailored to $\Psi_{TV}$. The guarantees of [CX15] underpin once again the fundamental role of TV-sparsity, but as pointed out above, a refined non-asymptotic analysis would eventually rely on additional geometric properties of the ground truth signal. For a more extensive discussion of total variation in compressed sensing, see [KKS17] and the references therein.

3.3 Tight Random Frames

Proposition 2.7 confirms the optimality of our sampling-rate bound for orthonormal bases, and indeed, $x^* \mapsto M(\Psi, x^*)$ precisely describes the phase transition curve of $(BP_{TV})$ in this case, see [ALMT14, Prop. 4.5]. However, the situation already becomes more complicated for a closely related class of analysis operators, namely those generated from standard Gaussian matrices. These constructions almost surely yield (tight) frames in general position\(^\dagger\) and have been widely studied in the literature as a benchmark example of the analysis formulation [KR15; NDEG13].

In this part, we roughly follow a construction of tight random frames from [NDEG13]: First, draw a matrix $\Psi_{rand} \in \mathbb{R}^{N \times n}$ with $N > n$, where $\Psi_{rand}$ is in general position if every subset of $\Psi_{TV,1}$ adapts to each of the cases. But we have to clearly confess that our predictions are not very significant and particularly remains unchanged for both signal classes. Let us conclude our discussion as follows:

Total variation minimization exhibits a non-uniform recovery behavior, which is also captured by the sampling-rate bound of Theorem 2.5. However, if the TV-sparsity is very small, our prediction is much less accurate than for redundant wavelet frames.

Remark 3.5 (Related Literature) Compared to redundant wavelet frames, the $\ell^1$-analysis formulation of total variation is actually much better theoretically understood. For example, the work of Cai and Xu [CX15] proves that the optimal sampling rate of $(BP_{TV})$ is given by $m = \Theta(\sqrt{\|\Psi_{TV} \cdot x\|_0})$ for Gaussian measurements, where $S_{TV} = \|\Psi_{TV} \cdot x\|_0$ and log-terms are ignored. The corresponding proofs are also based on estimating the Gaussian mean width, although using very different techniques that are specifically tailored to $\Psi_{TV}$. The guarantees of [CX15] underpin once again the fundamental role of TV-sparsity, but as pointed out above, a refined non-asymptotic analysis would eventually rely on additional geometric properties of the ground truth signal. For a more extensive discussion of total variation in compressed sensing, see [KKS17] and the references therein.

3.3 Tight Random Frames

Proposition 2.7 confirms the optimality of our sampling-rate bound for orthonormal bases, and indeed, $x^* \mapsto M(\Psi, x^*)$ precisely describes the phase transition curve of $(BP_{TV})$ in this case, see [ALMT14, Prop. 4.5]. However, the situation already becomes more complicated for a closely related class of analysis operators, namely those generated from standard Gaussian matrices. These constructions almost surely yield (tight) frames in general position\(^\dagger\) and have been widely studied in the literature as a benchmark example of the analysis formulation [KR15; NDEG13].

In this part, we roughly follow a construction of tight random frames from [NDEG13]: First, draw an $N \times n$ Gaussian random matrix and compute its singular value decomposition $U \Sigma V^T$. If $N \geq n$, we replace $\Sigma$ by the matrix $[I_n, 0]^T \in \mathbb{R}^{N \times n}$, which yields a tight frame

$$\Psi_{rand} := U([I_n, 0]^T)V^T \in \mathbb{R}^{N \times n}.$$

If $N < n$, we replace $\Sigma$ by $[I_N, 0] \in \mathbb{R}^{N \times n}$ and set $\Psi_{rand} := U[I_N, 0]V^T \in \mathbb{R}^{N \times n}$. But note that this operator does not form a frame, and in particular, $\Psi_{rand}$ is not in general position.

Our first phase transition plot is created according to Experiment 3.6 below with $n = 300$, $N = 350$, and

$$\mathcal{S} = \{(51, N - 51), (54, N - 54), (57, N - 57), \ldots, (297, N - 297)\}.$$

The set of tuples is chosen such that for every pair $(S, L) \in \mathcal{S}$, it holds $S + L = N = 350$. Moreover, we do only consider $S > 50$, since otherwise $\ker([\Psi_{rand}]_{S^c}) = \{0\}$ due to the general position property of $\Psi_{rand}$. The experimental result in Figure 10 shows that the sampling-rate bound of Theorem 2.5 almost perfectly hits the transition curve.

---

**Experiment 3.6 (Phase transition for random analysis operators)**

**Input:** Fixed ambient dimension $n$, range of sparsity-cosparsity tuples

$$\mathcal{S} \subset \{0, 1, \ldots, N\} \times \{0, 1, \ldots, N\}.$$  

**Compute:** Repeat the following procedure $5$ times for each pair $(S, L) \in \mathcal{S}$:

---

\(^\dagger\)A matrix $\Psi \in \mathbb{R}^{N \times n}$ with $N \geq n$ is in general position if every subset of $n$ rows is linearly independent.
Figure 10: Phase transition of $\text{BP}_{\eta=0}^{\Psi_{\text{rand}}}$ for tight random frames $\Psi_{\text{rand}} \in \mathbb{R}^{N \times n}$. The blue curve is obtained by computing the empirical mean of $M(\Psi_{\text{rand}}, x^*)$ for every iteration step $(S, L) \in \mathcal{S}$ of Experiment 3.6. The labels of the horizontal axis do only display the first component of $S$.

- Set $N = S + L$ and construct a random operator $\Psi_{\text{rand}} \in \mathbb{R}^{N \times n}$ as described above. Select a random set $S \subset [N]$ with $|S| = S$ and determine an orthonormal basis $B$ of ker$[\Psi_{\text{rand}}]_{S^c}$. Then draw a standard Gaussian random vector $c$ and set $x^* := Bc$.

- Repeat the following steps 10 times for each $m \in \{1, 2, \ldots, n\}$:
  - Draw a standard i.i.d. Gaussian random matrix $A \in \mathbb{R}^{m \times n}$ and determine the measurement vector $y = Ax^*$.
  - Solve the analysis basis pursuit $\text{BP}_{\eta=0}^{\Psi_{\text{rand}}}$ to obtain an estimator $\hat{x} \in \mathbb{R}^n$.
  - Declare success if $\|x^* - \hat{x}\|_2 < 10^{-5}$.

In Figure 10, we have selected the pairs $(S, L)$ such that both $n$ and $N$ remained fixed. The purpose of the second and third run of Experiment 3.6 is to highlight the non-trivial impact of both parameters by varying either one of them (and adapting $N$). Figure 11(a) is obtained by fixing $L = 250$ and applying Experiment 3.6 with $n = 300$ and

$$
\mathcal{S} = \{(1, L), (4, L), (7, L), \ldots, (100, L)\}.
$$

In Figure 11(b), we have fixed $S = 50$ and invoked Experiment 3.6 with $n = 300$ and

$$
\mathcal{S} = \{(S, 200), (S, 203), (S, 206), \ldots, (S, 299), (S, 300)\}.
$$

The transition curves of Figure 11(a) and Figure 11(b) are obviously not constant in $S$ and $L$, respectively, which verifies once again that neither sparsity nor cosparsity can thoroughly quantify the recovery performance of $\ell^1$-analysis minimization. Note that the plots of Figure 11 do actually consist of two regions for which the shapes of the curves are somewhat different. This is due to the case distinction between $N \geq n$ and $N < n$, i.e., whether $\Psi_{\text{rand}} \in \mathbb{R}^{N \times n}$ forms a frame or not. Interestingly, Figure 11(b) also reveals another drawback of the traditional bound in (1.3): If $S$ remains fixed, the analysis dimension $N = S + L$ increases with $L$, so that one would expect a logarithmic growth of the number of required measurements. But Figure 11(b) shows that the true sample complexity even
GENZEL, KUTYNIOK, AND MÄRZ: \(\ell^1\)-Analysis Minimization and Generalized (Co-)Sparsity

Figure 11: Phase transition of \((\text{BP}_{\Psi_{\text{rand}}}^\eta; \eta = 0})\) for random frames \(\Psi_{\text{rand}} \in \mathbb{R}^{N \times n}\) with either \(L\) or \(S\) fixed. (a) \(L = 250\) is fixed and the labels of the horizontal axis do only display the sparsity of \(\mathcal{S}\). (b) \(S = 50\) is fixed and the labels of the horizontal axis do only display the cosparsity of \(\mathcal{S}\).

decreases for larger values of \(L\). Regarding Theorem 2.5, we can again conclude that the sampling-rate function \(M(\Psi_{\text{rand}}, \mathbf{x}^*)\) captures the phase transition almost perfectly in both scenarios. The prediction of [KRZ15, Thm. 3.3] also reflects the shapes of the transition curves here, but is still slightly worse than ours.\(^1\)

### 3.4 Analysis Operators for 2D Signals

We have only considered one-dimensional signals up to now, in particular, the class of piecewise constant functions in Subsection 3.1 and Subsection 3.2. However, signals in higher dimensions typically exhibit a richer geometric structure, such as anisotropic features. For example, it has been observed in the literature that the sample complexity of total variation minimization in 2D scales very differently compared to its counterpart in 1D [CX15; KKS17; NW13]. Let us therefore also conduct a simple experiment in 2D, using total variation as well as a redundant Haar wavelet frame as analysis operators. In order to reduce the immense computational burden of creating a phase transition plot, we just restrict ourselves to a specific image signal here. Figure 12(a) illustrates our choice of \(\mathbf{x}^*\), which is a realistic brain phantom for magnetic resonance imaging [GLPU12] at a resolution of \(100 \times 100\).

Recovery based on total variation in 2D aims at retrieving images from compressed measurements by promoting sparse gradients, i.e., piecewise constant signals. As in 1D, there exist multiple variants of total variation minimization, depending on different boundary conditions and on how the \(\ell^1\)-norm of the discrete gradient is calculated. To keep the exposition as brief as possible, we focus on the simple case where an anisotropic total variation operator, based on periodic forward differences, is used and \(\mathbf{x}^*\) corresponds to an \(\tilde{n} \times \tilde{n}\) grayscale image. Thus, up to modifications of the boundary values, the entries of the discrete (forward) gradient in 2D are defined by

\[
\begin{align*}
(\nabla \mathbf{x}^*)_{j,j'} := & \begin{bmatrix} (\nabla_1 \mathbf{x}^*)_{j,j'} \\ (\nabla_2 \mathbf{x}^*)_{j,j'} \end{bmatrix} := \begin{bmatrix} x^*_{j+1,j'} - x^*_{j,j'} \\ x^*_{j,j'+1} - x^*_{j,j'} \end{bmatrix}.
\end{align*}
\]

For the sake of convenience, we will now identify the signal domain \(\mathbb{R}^{\tilde{n} \times \tilde{n}}\) with its canonical columnwise vectorization \(\mathbb{R}^{\tilde{n}^2}\), i.e., the ambient space is of dimension \(n := \tilde{n}^2\) and \(\mathbf{x}^* \in \mathbb{R}^n\). The associated 2D

\(^1\)Note that, although [KRZ15, Thm. 3.3] is only stated for frames, it literally holds true for any choice of \(\Psi \in \mathbb{R}^{N \times n}\) where the upper frame bound just needs to be replaced by \(\|\Psi\|_{2\to2}^2\).
APPLICATIONS AND NUMERICAL EXPERIMENTS

Figure 12: (a) Visualization of the ground truth image signal $x^* \in \mathbb{R}^{100^2} \simeq \mathbb{R}^{100 \times 100}$ (‘brain phantom’ [GLPU12]). (b)+(c) Average recovery error achieved by $(\text{BP}_{\Psi_{TV-2}}^\eta=0)$ and $(\text{BP}_{\Psi_{irdwt-2}}^\eta=0)$, respectively.

To define a redundant Haar wavelet frame in 2D, we again rely on the (two-dimensional) inverse wavelet transform with 2 scales provided by the software package spot [BF13], and denote it by $\Psi_{irdwt-2} \in \mathbb{R}^{7n \times n}$. We assess the recovery capability very similarly to Figure 6(c): Experiment 3.1 is invoked for both $\Psi = \Psi_{TV-2}$ and $\Psi = \Psi_{irdwt-2}$, where $n = \tilde{n}^2 = 10.000$ and $M = \{1, 26, 51, \ldots, 9976\}$; note only 12 repetitions were performed instead of 50. The plots of Figure 12(b) and 12(c) visualize the averaged error curves for $(\text{BP}_{\Psi_{TV-2}}^\eta=0)$ and $(\text{BP}_{\Psi_{irdwt-2}}^\eta=0)$, respectively. Unfortunately, solving these minimization problems via cvx with $n = 10.000$ is not feasible anymore, so that we have applied an inhouse implementation based on the alternating direction method of multipliers instead; see [BPCPE11] and references therein.

Similar to the 1D case, it can be observed that the coefficient sparsity $\|\Psi_{irdwt-2}x^*\|_0 = 25.897$ is substantially larger than the ambient dimension $n = 10.000$, even though $\Psi_{irdwt-2}$ is rather mildly redundant. Nevertheless, there appears to be a sharp phase transition at $m \approx 5.500$ measurements in Figure 12(c). The prediction $m \approx 5.500$ of Theorem 2.5 is also quite accurate in this situation, corresponding to a relative error of about 5%. In contrast, our bound is slightly more pessimistic for 2D total variation (relative error about 10 %) but still quite reliable. This is obviously not the case for [KRZ15], whose predictions deviate significantly from the true sample complexity. Finally, it is worth mentioning that the above experimental setup may not be suited for a comparison between total variation and
Haar wavelets in respect of analysis modeling. We suspect that redefining $\Psi_{\text{irdwt-2}}$ with more scales and a more sophisticated weighting would considerably improve the outcomes of $(BP_{\Psi_{\text{irdwt-2}}})$.

4 Related Literature

In this part, we survey existing theoretical approaches to the analysis formulation and put them into context with our findings. Our discussion starts in Subsection 4.1 with a comparison to the synthesis formulation, which is also widely used in compressed sensing and promotes a somewhat different viewpoint on sparse representations. In Subsection 4.2, we return to our initial concern from Subsection 1.1, presenting several results relying on traditional analysis sparsity (cf. (1.3)). Subsection 4.3 then points out the importance of cosparse modeling. Finally, more details on the recent work of Kabanava, Rauhut, and Zhang [KRZ15] are provided in Subsection 4.4.

4.1 Analysis versus Synthesis

One of the cornerstones in the literature on sparsity priors is [EMR07], which was the first work systematically studying the relationship between the predominant synthesis formulation and the analysis formulation. Instead of solving $(BP_{\Psi})$, the synthesis formulation rather considers the convex program

$$D \cdot \left( \arg\min_{z \in \mathbb{R}^N} \|z\|_1 \text{ subject to } \|ADz - y\|_2 \leq \eta \right), \tag{BPS^D_{\eta}}$$

where $D \in \mathbb{R}^{n \times N}$ is a (possibly redundant) dictionary in $\mathbb{R}^n$. The rationale behind this approach is that the signal vector $x^* \in \mathbb{R}^n$ possesses a sparse representation by $D$, i.e., $x^* = Dz$ for a coefficient vector $z \in \mathbb{R}^N$ with $\|z\|_0 \ll n$.

By investigating the respective polytopal geometry of $(BP_{\Psi})$ and $(BPS^D_{\eta})$, the authors of [EMR07] develop a theoretical model that describes the differences between both methods. In particular, they point out the following fundamental issue: While the synthesis approach $(BPS^D_{\eta})$ seems to benefit from a higher redundancy of the dictionary $D$, it is not clear how $(BP_{\Psi})$ is influenced by the redundancy degree of $\Psi$. Theorem 2.5 shows that this structural property indeed has a wide impact, which is however highly non-trivial. We hope that our results could give rise to further progress in this matter.

The numerical simulations of [EMR07] certify a considerable gap between both strategies and the observed recovery performances indicate that the analysis formulation outperforms its synthesis-based counterpart in many situations of interest. Although never stated as a general conclusion, the superiority of analysis-based priors is often confirmed in the literature, e.g., see [SF09]. Moreover, depending on the redundancy of $\Psi$ and $D$, one may also argue that solving $(BP_{\Psi})$ is computationally more appealing, since the actual optimization takes place in $\mathbb{R}^n$. In contrast, $(BPS^D_{\eta})$ operates on the possibly much higher dimensional coefficient space of $\mathbb{R}^N$. We refer the interested reader to [RSV08] for more information on synthesis recovery.

4.2 Approaches Based on Analysis Sparsity

The first compressed-sensing-based approach to the analysis formulation was undertaken by the seminal work of Candès, Eldar, Needell, and Randall [CENR11]. While previous results for redundant dictionaries did rather study coefficient recovery in the synthesis formulation, e.g., based on incoherent frame atoms [RSV08], a major breakthrough of [CENR11] was that such assumptions can be avoided by switching over to the analysis perspective. The theoretical analysis of [CENR11] relies on the so-called $D$-RIP, which is an adaptation of the classical restricted isometry property to sparse representations in dictionaries:
Definition 4.1 (D-RIP, [CENR11]) Let $D \in \mathbb{R}^{m \times N}$ be a tight frame\textsuperscript{1} for $\mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$. Then, $A$ satisfies the D-RIP with parameters $\delta > 0$ and $S \in [N]$ if

$$(1 - \delta)\|Dz\|_2^2 \leq \|ADz\|_2^2 \leq (1 + \delta)\|Dz\|_2^2$$

holds true for all $S$-sparse vectors $z \in \mathbb{R}^N$.

Supposed that $A$ fulfills the D-RIP with constant $\delta < 0.08$, it was shown in [CENR11] that any minimizer $\hat{x} \in \mathbb{R}^n$ of (BP$_{\Psi}^A$) with $\Psi := D^\top$ obeys

$$\|\hat{x} - x^*\|_2 \leq C_0 \cdot \eta + C_1 \cdot \frac{\sigma_S(\Psi x^*)_1}{\sqrt{S}}, \quad (4.1)$$

where $C_0, C_1 > 0$ are numerical constants and $\sigma_S(\Psi x^*)_1$ denotes the best $S$-term approximation error of the coefficient vector $\Psi x^*$. This result was the starting point of several generalizations and refinements: For instance, [Fou14] has extended the above setup to arbitrary frames and Weibull matrices, incorporating an adapted robust null space property. The work [LML12] modifies the concept of D-RIP in a such way that (BP$_{\Psi}^A$) can be applied to arbitrary dual operators of frames. Several achievement towards practical applicability were also made by [KNW15], allowing for structured measurements if a certain incoherence condition on the dictionary $D$ and the subsampled measurement matrix is satisfied. Interestingly, guarantees of a similar flavor were even proven in an infinite-dimensional setting by [Poo17], using different proof techniques.

Regarding the measurement model of this work (Model 2.1), it has turned out that a standard Gaussian matrix $A \in \mathbb{R}^{m \times n}$ fulfills the D-RIP with high probability provided that the number of observations obeys $m \gtrsim S \cdot \log(N/S)$. In this case, we again obtain a uniform error bound of the form (4.1). A similar sampling rate for Gaussian measurements was also recently achieved by Kabanova and Rauhut in [KR15]. Their proof is however based on yet another statistical tool, namely a modified version of Gordon’s Escape Through a Mesh—not that we also make use of this fundamental principle in Theorem A.1. In a nutshell, the first (non-uniform) guarantee of [KR15] reads as follows: Given $x^* \in \mathbb{R}^n$ with $\|\Psi x^*\|_0 \leq S$, a Gaussian measurement matrix $A \in \mathbb{R}^{m \times n}$ satisfying

$$\frac{m^2}{m + 1} \geq \frac{2b}{a} \cdot S \cdot \left(\sqrt{\log \left(\frac{eN}{S}\right)} + \sqrt{\frac{e \log(e^{-1})}{bS}}\right)^2 \quad (4.2)$$

enables recovery of $x^*$ via (BP$_{\Psi}^{\Psi_{\eta=0}}$) with probability at least $1 - \varepsilon$. It is worth mentioning that, compared to (4.1), this bound does not involve any unspecified numerical constants. Moreover, by introducing a generalized null space property, the statement of (4.2) can be even extended to uniform recovery [KR15, Thm. 9].

Compared to our sampling-rate bound in Theorem 2.5, a condition of the form $m \gtrsim S \cdot \log(N/S)$ is quite attractive and intuitive from an aesthetic viewpoint, since it mimics traditional results in compressed sensing. This resemblance is not very astonishing, since most proofs in the literature do actually “operate” on the space of analysis coefficients $\mathbb{R}^N_1$ and then just pull back the corresponding estimates to the signal domain $\mathbb{R}^n$. For various examples of less redundant analysis operators, e.g., tight random frames, such a strategy indeed seems to provide accurate predictions of the required number of measurements.

However, as already pointed out in the course of Subsection 1.1, the situation becomes much more delicate for highly redundant and coherent systems, in which the above approaches do by far not explain the capacity of (BP$_{\Psi}^A$). The theoretical framework of Section 2 and the preceding experiments in Section 3 demonstrate, for instance, that the mutual coherence structure of an operator $\Psi$ deserves special attention as well. Apart from that, our results do not assume that $\Psi$ forms a spanning system

\footnote{In contrast to our convention in Subsection 1.3, the columns of $D$ form a frame for $\mathbb{R}^n$ here, meaning that $D$ is actually the associated synthesis operator.}
for $\mathbb{R}^n$, which in turn was an important hypothesis in [CENR11; KR15]. This observation particularly gives rise to doubts that conditioning-related quantities, such as the ratio $b/a$ in (4.2), need to appear explicitly in a sampling-rate bound.

Let us finally turn to the issue of stability: The error estimate of (4.1) relies on the best $S$-term approximation error. Since this parameter is completely determined by the analysis coefficient sequence $\Psi x^*$, the linear dependencies within $\Psi$ are again completely disregarded. The compressibility experiment of Figure 6 shows that this methodology can be problematic for redundant systems. On the other hand, our approach in Subsection 2.4 is based on approximations in the signal space $\mathbb{R}^n$, and when combined with an appropriate greedy scheme (e.g., Algorithm B.1), it yields much more reliable outcomes.

### 4.3 Approaches Based on Cosparse Modeling

A very influential contribution to analysis-based priors in inverse problem theory is the work of Nam, Davies, Elad, and Gribonval [NDEG13], which has coined the term *cosparse analysis model*. Perhaps, this recent branch of research has grown out of the same observation as we have made above:

"The fact that this representation may contain many non-zeroes (and especially so when $N \gg n$) should be of no consequence to the efficiency of the analysis model." [NDEG13, p. 35]

In fact, a key insight of this paper was that the viewpoints of analysis and synthesis sparsity are actually based on completely different signal models and therefore also draw their strength from different features. While the synthesis approaches puts its emphasis on the number of non-zero coefficients, [NDEG13] argues that the analysis model is rather specified by the number of vanishing coefficients, that is, the cosparsity. Indeed, the class of $L$-cosparse signals can be simply written as a union of subspaces

$$
\bigcup_{\Lambda \subset [N], |\Lambda| = L} W_\Lambda,
$$

where $W_\Lambda := \ker \Psi_\Lambda$. A careful comparison of both setups reveals that one should not—as suggested by the approaches presented in Subsection 4.2—regard the analysis formulation from a synthesis perspective, thereby treating "the analysis operator as a poor man’s sparse synthesis representation" [NDEG13, p. 35]. For example, the subspace collection of (4.3) may become extremely complicated if $\Psi$ corresponds to a redundant frame with strong linear dependencies, and it is quite evident that its geometric arrangement is then not just reflected by the value of $S$ (or $L$).

Towards a theoretical analysis, [NDEG13, Thm. 7] establishes the following equivalence: Every $x^* \in \mathbb{R}^n$ with cosupport $\Lambda = S_x^c \subset [N]$ is a unique minimizer of (BP $\Psi_{\eta=0}$) if, and only if,

$$
\sup_{x \in \mathbb{R}^n} \frac{|\langle \Psi_\Lambda^c z, \text{sign}(\Psi_\Lambda^c x) \rangle|}{\|\Psi_\Lambda z\|_1} < 1 \quad \text{for all } z \in \ker(A) \setminus \{0\}.
$$

We note that similar conditions were also studied for stable recovery in [VPDF13]. Moreover, it is worth mentioning that the vector $\text{sign}(\Psi_\Lambda^c x)$ plays a central role in (4.4), which reminds us of our generalized sparsity term $S^c_x$ from Definition 2.3. In contrast, such an expression is entirely missing in the concept of $D$-RIP.

Regarding sample complexity, it has turned out that the quantity

$$\kappa_\Psi(L) := \max_{|\Lambda| \geq L} \dim W_\Lambda$$

essentially determines the (optimal) number of measurements required for noiseless recovery via a combinatorial search over all $L$-cosparse signals [NDEG13, Sec. 3]. This finding is particularly consistent with the philosophy of compressed sensing after which an estimation succeeds as long as the
sampling rate slightly exceeds the signal’s manifold dimension. Somewhat surprisingly, the recent work of Giryes, Plan, and Vershynin [GPV15] has shown that the situation is unfortunately not that simple in $\ell^1$-analysis minimization: When asking for robustness or a tractable algorithm instead of combinatorial searching, the sample complexity is not appropriately captured by $\kappa_\Psi(L)$ anymore. For that reason, we hope that our theoretical framework sheds more light on this issue, featuring the complex relationship between sparsity and cosparsity in the study of $(BP_\Psi^\eta)$.

### 4.4 Non-Asymptotic Sampling-Rate Bounds

To the best of our knowledge, the only work that allows for a direct comparison to our results is by Kabanava, Rauhut, and Zhang [KRZ15], hence often serving as a “competitor” in Section 3. In fact, it is also motivated by a similar observation as we have made in this paper: Bounds on the number of needed measurements for frame-based analysis operators (cf. (4.2)) and total variation [CX15] may render vacuous statements if the sparsity level is too high.

Focusing on the case of Gaussian measurements, Theorem 3.3 in [KRZ15] provides a non-asymptotic sampling-rate bound when $\Psi \in \mathbb{R}^{N \times n}$ is a frame: Any minimizer $\hat{x} \in \mathbb{R}^n$ of $(BP_\Psi^\eta)$ satisfies $\|\hat{x} - x^\star\|_2 \leq 2\eta/\tau$ with probability $1 - \epsilon$, provided that

$$m^2 \geq \left( \sqrt{n - \frac{2 \sum_{c \subseteq S^c} \|\psi_c\|^2}{\pi^2 N}} + \sqrt{2 \log(\epsilon^{-1}) + \tau} \right)^2,$$

where $b$ is an upper frame bound of $\Psi$. An analogous statement holds true if $\Psi$ is the total variation operator in 1D (cf. Subsection 3.2), where (4.5) is replaced by

$$m^2 \geq \left( \sqrt{n \cdot \left( 1 - \frac{1}{\pi} \left( 1 - \frac{S + 1}{N} \right)^2 \right)} + \sqrt{2 \log(\epsilon^{-1}) + \tau} \right)^2,$$

see Theorem 3.1 in [KRZ15] for details. Ignoring terms of lower order, both bounds involve the dimension $n$ of the ambient space as an additive term. This resemblance to the sampling-rate function $M(\Psi, x^\star)$ from Definition 2.4 is actually not very surprising, since the proof strategy of [KRZ15] builds upon the same principle as in this work, namely estimating the conic mean width.

However, there are several substantial differences between the approach of [KRZ15] and ours:

- The results of [KRZ15] are not consistent with classical compressed sensing theory: If $\Psi \in \mathbb{R}^{n \times n}$ is an orthonormal basis, (4.5) basically degenerates into

$$m \geq \left( 1 - \frac{2}{\pi} \right) \cdot n + \frac{4S}{\pi^2} - \frac{2S^2}{\pi^3},$$

which is clearly suboptimal for small values of $S$.

- The guarantees of [KRZ15] are specific to Gaussian measurements and do not include stable recovery that allows for coping with compressible signals.

- In all cases examined in Section 3, our bounds are vastly superior in terms of predicting the phase transition behavior of $(BP_\Psi^\eta_{\eta=0})$; see Figure 8, Figure 9, and Figure 11.

Comparing the sampling-rate bound (2.4) of Theorem 2.5 with (4.5), we observe that the latter does neither incorporate the cross-correlations between the individual analysis vectors nor the sign vector $\sigma_{x^\star} = \text{sign}(\Psi x^\star)$. On the other hand, both features seem to be essential for achieving sound recovery results. A careful study of the corresponding proofs reveals the underlying reason for the superiority of our approach: While Step 1 in the proof of Theorem 6.8 in Subsection 6.3 is rather standard (cf. [ALMT14]), our argumentation in Step 2 is quite different. Indeed, the dual vector $w'$ in [KRZ15] is chosen such that it just maximizes the term $T_1$ in (6.11). Our choice of $w'$ in (6.10) is more sophisticated,
particularly taking the variable $\tau$ into account. This appears to be a minor issue at first sight but actually affects the remainder of the proofs drastically. For example, we had to come up with a highly non-trivial estimate for $T_2$ in Lemma 6.12, eventually leading to the generalized cosparsity term $L_{S_c}$ (see Definition 2.4). Furthermore, the optimization over $\tau$ is very challenging in our case, whereas it is just a simple quadratic problem in [KRZ15].

5 Discussion and Outlook

One of the key findings of this work was that the widely-used concept of analysis sparsity is often-times not capable of explaining the success (and failure) of $\ell^1$-analysis recovery. This observation defies conventional wisdom that would justify the superiority of an operator $\Psi \in \mathbb{R}^{N \times n}$ only by means of sparsity, i.e., smaller values of $S = \|\Psi x^*\|_0$ lead to better outcomes. The definition of our three generalized sparsity parameters (see Definition 2.3), in contrast, is somewhat more delicate and particularly involves the Gram matrix $G = \Psi \Psi^T \in \mathbb{R}^{N \times N}$. In that way, the mutual coherence structure of $\Psi$ is also taken into account, which is in turn a missing feature of most traditional approaches. This important refinement enabled us to prove much more reliable bounds on the sample complexity of the analysis basis pursuit $(BP_\Psi), \xi)$, even when a coefficient sequence $\Psi x^* \in \mathbb{R}^N$ is just compressible (see Subsection 2.4). Hence, we can conclude that our main results, Theorem 2.5 and Theorem 2.8, provide fairly general answers to the initial challenges of (Q1) and (Q2), as stated in Subsection 1.1.

A more technical yet crucial role in these statements is played by the sampling-rate function $M(\Psi, x^*)$, which essentially determines the minimal number of measurements needed to invoke the actual error estimates. Our numerical simulations in Section 3 have demonstrated that this quantity indeed very accurately localizes the phase transition of many different types of analysis operators, such as redundant Haar wavelet systems, total variation, or random frames. This variety certifies once again that the proposed bounds meet the desiderata of (i)–(iv) from Subsection 1.2 in many situations of interest. Thus, let us make the following closing remark:

The functional $x^* \mapsto M(\Psi, x^*)$ can be regarded as a surrogate measure of complexity, quantifying how well $\Psi$ captures the low-dimensional structure of a signal $x^*$. Compared to the number of non-zero analysis coefficients, its value has turned out to be an appropriate indicator for the success or failure of $\ell^1$-analysis minimization.

5.1 Practical Scope and Guidelines

The previous sections have rather focused on the predictive power of our theoretical framework, whereas its practical implications were only marginally addressed so far. Therefore, we shall now investigate the issues of (Q3) in greater detail. Taking the perspective of a practitioner, these problem statements are typically associated with the following tasks: How to come up with a good analysis operator $\Psi$ for my specific application? What properties and design rules are of particular relevance? Unfortunately, we are not able to answer these questions in full generality here, but our theoretical approach could at least give rise to several novel solution strategies. The above discussion in fact suggests to use $M(\Psi, x^*)$ as a measure of quality that allows for an assessment and comparison of different choices of $\Psi$. Let us recall a simple example from Subsection 3.1.2: In the course of relation (3.2), it has turned out that, although $\Psi_{rdwt}$ and $\Psi_{irdwt}$ correspond to the same frame except from a scale-wise reweighting of their analysis vectors, the predicted sampling rates differ dramatically, as illustrated in Figure 1(b). This observation indicates that the recovery performance might be greatly improved by weighting, which is a principle often referred to as weighted sparsity in the literature [FMSY12; OKH12]. The numerical results of Section 3 even show that such a rule-of-thumb could be applied to various other structural properties of analysis operators, for example:

- Intrinsic localization of the Gram matrix. The off-diagonal decay of $G = \Psi \Psi^T$ may strongly affect the recovery capability of $\Psi$ (cf. Remark 6.15).
• Conditioning. The impact of the linear dependency structure and frame bounds of $\Psi$ is expected to be highly non-trivial.

• Redundancy. Highly redundant frames are sometimes superior over orthonormal bases (see Figure 1).

While analyzing $\mathcal{M}(\Psi, x^\ast)$ with respect to certain characteristic quantities yields a first approach to (Q3), a systematic methodology for designing good operators is still missing. Interestingly, the language of statistical learning theory can be very helpful to tackle this challenge. The underlying mathematical formalism is sketched in Appendix C, but a detailed elaboration is however deferred to future research.

5.2 Open Problems

Beside the emerging challenge of the previous subsection, we would like to state some further open issues which might be studied in future works as well:

• Sharper bounds and optimality. Our analysis of the total variation operator in Subsection 3.2 has demonstrated that the sampling-rate bound of Theorem 2.5 could be somewhat inaccurate in situations of low sparsity. While there are indeed several non-sharp estimates in the proof of Theorem 6.8 (e.g., see Proposition 6.9 and Lemma 6.12), the most heuristic step is actually our choice of the dual vector $w'$ in (6.10), see also Remark 6.15. This leaves certain space for refinements, but note that, no matter how $w'$ is chosen, one eventually has to face a non-trivial minimization problem over $\tau$ in (6.8). Apart from that, it would be interesting to specify conditions on $\Psi$ and $x^\ast$ under which our bounds are already optimal, at least in an asymptotic sense.

• Greedy algorithms. The analysis basis pursuit is just a very popular example of algorithms that try to exploit analysis sparsity in signal estimation. In fact, there exist numerous alternatives to solving (BP$_\Psi$), for instance, the greedy analysis pursuit [NDEG13] or those of [GNEG14]. We believe that our proof techniques can be adapted to many of these methods, ultimately leading to similar recovery guarantees.

• Structured measurements. Our argumentation in Section 6 does strongly rely on the row-independence of the measurement matrix $A \in \mathbb{R}^{m \times n}$. This assumption is unfortunately not satisfied for many types of structured measurement schemes, so that our statistical tools from Appendix A.1 do not apply anymore. In particular, the concept of Gaussian mean width might become inappropriate, implying that most parts of our proofs break down. For such situations, one probably would have to come up with a completely new proof strategy.

• Analysis versus synthesis. We have already discussed the crucial difference between analysis- and synthesis-based priors in Subsection 4.1, but a fundamental question remains widely open: Which of the two formulations is more appropriate for a given signal class and measurement ensemble?

6 Proofs of Main Results

Before presenting all proofs in detail, let us briefly sketch the roadmap of this section: In Subsection 6.1, we first provide an abstract framework, which even applies to a generalized version of the basis pursuit. The associated recovery results rely on variants of the Gaussian mean width, which forms a key concept in analyzing the sample complexity of many convex estimation programs [ALMT14; CRPW12]. These quantities are however very implicit and hard to compute in general. For that reason, we show in Subsection 6.2 (Theorem 6.8) that the square of the Gaussian mean width of the descent cone of $x \mapsto \|\Psi x\|_1$ at $x^\ast$ is at least upper bounded by our sampling-rate function $\mathcal{M}(\Psi, x^\ast)$, introduced in Definition 2.4. Combining this key step with the guarantees from Subsection 6.1, we will immediately
obtain the statements of Theorem 2.5 and Theorem 2.8. The more technical proof of Theorem 6.8 is then postponed to Subsection 6.3.

### 6.1 Structured Signal Recovery and Gaussian Mean Width

In this part, we analyze the \textit{generalized basis pursuit}

\[
\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad \|Ax - y\|_2 \leq \eta, \tag{P_\eta^f}
\]

where \(f : \mathbb{R}^n \to \mathbb{R}\) is a convex function. Note that for our purposes, it would be sufficient to consider the \(\ell^1\)-analysis functional \(f(x) = \|\Psi x\|_1\), but in fact, each of the following arguments does literally hold true for the general case. Conceptually, the optimization program of \(\text{(P}_\eta^f\text{)}\) seeks for the most structured signal \(x \in \mathbb{R}^n\) that is consistent with the measured data \(y = Ax^* + e\). The optimal sampling rate of this approach heavily relies on how well the objective function \(f\) captures the complexity of the ground truth signal \(x^* \in \mathbb{R}^n\). In order to formalize this intuition, let us first introduce the notion of (Gaussian) mean width:

**Definition 6.1** Let \(K \subset \mathbb{R}^n\) be a set.

1. The \textit{(global) mean width} of \(K\) is defined as
   \[
   w(K) := \mathbb{E}[\sup_{h \in K} \langle g, h \rangle],
   \]
   where \(g \sim \mathcal{N}(0, I_n)\) is a standard Gaussian random vector.

2. The \textit{conic mean width} of \(K\) is given by
   \[
   w_\lambda(K) := w(\text{cone}(K) \cap S^{n-1}).
   \]

3. The \textit{local mean width} of \(K\) at scale \(t > 0\) is defined as
   \[
   w_t(K) := w(\tfrac{t}{2}K \cap S^{n-1}).
   \]

**Remark 6.2** If \(K\) is convex and \(0 \in K\), the mapping \(t \mapsto w_t(K)\) is non-increasing, and we also have \(w_t(K) \leq w_\lambda(K)\), since \(\frac{t}{2}K \subset \text{cone}(K)\) for all \(t > 0\). Moreover, it holds \(w_t(\text{cone}(K)) = w_\lambda(\text{cone}(K)) = w_\lambda(K)\).

A particularly interesting choice of \(K \subset \mathbb{R}^n\) is the set of all descent directions of \(f\) at a certain “anchor point”:

**Definition 6.3** Let \(f : \mathbb{R}^n \to \mathbb{R}\) be a convex function and let \(x^* \in \mathbb{R}^n\). The \textit{descent set of \(f\) at \(x^*\)} is given by

\[
\mathcal{D}(f, x^*) := \{h \in \mathbb{R}^n \mid f(x^* + h) \leq f(x^*)\},
\]

and its corresponding \textit{descent cone} is denoted by \(\mathcal{D}_\lambda(f, x^*) := \text{cone}(\mathcal{D}(f, x^*))\). In this context, we refer to \(w_\lambda(\mathcal{D}(f, x^*))\) as the \textit{conic mean width} of \(f\) at \(x^*\).

Note that, since \(f\) is assumed to be convex, both \(\mathcal{D}(f, x^*)\) and \(\mathcal{D}_\lambda(f, x^*)\) are convex sets. The geometric idea behind these definitions is quite simple: Every minimizer of \(\text{(P}_\eta^f\text{)}\) is close to \(x^*\) if, and only if, the intersection between the constraint set \(\{x \in \mathbb{R}^n \mid \|Ax - y\|_2 \leq \eta\}\) and \(x^* + \mathcal{D}(f, x^*)\) has a small diameter. It was therefore a remarkable observation of [CRPW12] that, for Gaussian measurements, the occurrence of this event is completely characterized by a single parameter, namely the conic mean width of \(f\) at \(x^*\). For this reason, \(w_\lambda(\mathcal{D}(f, x^*))\) may be also regarded as a measure of complexity for a signal \(x^*\) with respect to a given structure-imposing functional \(f\).
Our first theoretical guarantee actually extends this fundamental principle by incorporating the notion of local mean width. We will see below that this refinement indeed allows us to establish stable recovery, which is an essential feature of Theorem 2.8.

**Theorem 6.4** Let Model 2.1 be satisfied and let \( f: \mathbb{R}^n \to \mathbb{R} \) be a convex function. There exists a numerical constant \( C > 0 \) such that, for every fixed \( t > 0 \) and \( u > 0 \), the following holds true with probability at least \( 1 - e^{-u^2/2} \): If

\[
t \geq \frac{2\eta}{[\sqrt{m-1 - C \cdot \kappa^2 \cdot \left( w_\lambda(\mathcal{D}(f,x^*)) + u \right)]_+},
\]

then any minimizer \( \hat{x} \) of \((P_\eta^*)\) satisfies \( \|\hat{x} - x^*\|_2 \leq t \). In the Gaussian case, \( a \sim \mathcal{N}(0, I_n) \), we have \( C = \kappa = 1 \).

A proof of Theorem 6.4 and all its corollaries below can be found in Appendix A.1. Roughly speaking, the error bound of Theorem 6.4 states the following: To achieve a desired reconstruction accuracy \( t > 0 \) via \((P_\eta^*)\), increase \( m \) until (6.1) is fulfilled. While this procedure always succeeds for sufficiently large values of \( m \), the condition of (6.1) is still quite implicit, since both sides depend on \( t \). Fortunately, due to Remark 6.2, one may just estimate the local mean width by its conic counterpart. This leads to our next result, which is well-known from the literature, e.g., see [Tro15, Cor. 3.5].

**Corollary 6.5** Let Model 2.1 be satisfied and let \( f: \mathbb{R}^n \to \mathbb{R} \) be convex. There exists a numerical constant \( C > 0 \) (the same as in Theorem 6.4) such that, for every \( u > 0 \), the following holds true with probability at least \( 1 - e^{-u^2/2} \): Any minimizer \( \hat{x} \) of \((P_\eta^*)\) satisfies

\[
\|\hat{x} - x^*\|_2 \leq \frac{2\eta}{[\sqrt{m-1 - C \cdot \kappa^2 \cdot \left( w_\lambda(\mathcal{D}(f,x^*)) + u \right)]_+}.
\]

In particular, if \( \eta = 0 \) and

\[
m > C^2 \cdot \kappa^4 \cdot (w_\lambda(\mathcal{D}(f,x^*)) + u)^2 + 1,
\]

one has \( \hat{x} = x^* \).

As already briefly discussed at the beginning of Subsection 2.3, the condition of (6.2) is not only sufficient but essentially also necessary to achieve exact recovery in the Gaussian case [ALMT14].\(^1\) Hence, the conic mean width is the crucial quantity to study when determining the sample complexity of \((P_\eta^*)\).\(^0\)

While Corollary 6.5 already includes robustness against noise, the wish for stable reconstructions via \((P_\eta^*)\), as requested in Subsection 2.4, remains unfulfilled. A major difficulty is that the mapping \( x^* \mapsto w_\lambda(\mathcal{D}(f,x^*)) \) is typically discontinuous, since it involves forming the conic hull of the descent set \( \mathcal{D}(f,x^*) \). If \( w_\lambda(\mathcal{D}(f,x^*)) \approx \sqrt{n} \), the statement of Corollary 6.5 loses its significance, even when there exists a vector close to \( x^* \) whose conic mean width is much smaller. For example, such a situation might occur if \( x^* \) just possesses a compressible coefficient sequence which is not too sparse; see also Subsection 2.4 as well as [Gen17, Sec. III.D] for more details. Our second corollary of Theorem 6.4 resolves this issue by analyzing the local mean width at a specific scale \( t > 0 \). It shows that one may investigate the complexity of \( f \) with respect to a surrogate vector \( \tilde{x}^* \in \mathbb{R}^n \) by sacrificing estimation accuracy in the order of the approximation error \( \|x^* - \tilde{x}^*\|_2 \).

**Corollary 6.6** Let Model 2.1 be satisfied and let \( f: \mathbb{R}^n \to \mathbb{R} \) be convex. There exists a numerical constant \( C > 0 \) (the same as in Theorem 6.4) such that, for every \( R > 0 \) and \( u > 0 \), the following holds true with probability at least \( 1 - e^{-u^2/2} \): Let \( \tilde{x}^* \in \mathbb{R}^n \) with \( f(\tilde{x}^*) = f(x^*) \). Then, any minimizer \( \hat{x} \) of \((P_\eta^*)\) satisfies

\[
\|\hat{x} - x^*\|_2 \leq \max \left\{ R\|x^* - \tilde{x}^*\|_2, \frac{2\eta}{[\sqrt{m-1 - \sqrt{n}_0 - 1}]_+} \right\},
\]

\(^1\)The authors of [ALMT14] are focusing on the so-called statistical dimension rather than on the Gaussian mean width, but these notions are actually equivalent (see [ALMT14, Prop. 10.2]).
where
\[ \bar{m}_0 := C^2 \cdot k^4 \cdot \left( \frac{k+1}{k} \cdot \left[ w_\lambda(\mathcal{D}(f, \hat{x})) + 1 \right] + u \right)^2 + 1. \] (6.4)

In particular, if \( \eta = 0 \), we have \( \| \hat{x} - x^* \|_2 \leq R \| x^* - \hat{x} \|_2 \) provided that \( m > \bar{m}_0 \).

In a nutshell, Corollary 6.6 certifies a good performance of \( (P_\eta^f) \) if both \( \hat{x}^* \) approximates \( x^* \) sufficiently well and \( w_\lambda(\mathcal{D}(f, \hat{x})) \) is small. For a more detailed discussion of this surrogate principle, we again refer to Subsection 2.4, which considers the special case of \( \lambda \)-subspaces such that \( \| x \|_{\lambda} \leq \bar{D} \) for low-complexity vectors according to \( x \) well and \( \hat{x} \) is a good approximation. Therefore, we need to come up with a more explicit expression or at least a meaningful upper bound on \( \| x^* - x^* \|_2 \) that \( \lambda \)-subspaces onto \( U \), such that \( f(\hat{x}^*) = f(x^*) \). If \( f \) is a semi-norm, one may even invoke the following upper bound on the approximation error, leading to a more explicit expression in (6.3):

**Proposition 6.7** Assume that \( f = \| \cdot \| \) is a semi-norm on \( \mathbb{R}^n \). Let \( x^* \in \mathbb{R}^n \) and let \( U \subseteq \mathbb{R}^n \) be a linear subspace such that \( \| P_U x^* \| \neq 0 \). Setting \( \hat{x}^* := \frac{x^*}{\| P_U x^* \|} \cdot P_U x^* \), we have \( \| x^* \| = \| \hat{x}^* \| \) and
\[ \| x^* - \hat{x}^* \|_2 \leq \frac{\| P_U x^* \|_2}{\| P_U x^* \|} \cdot \| P_U x^* \| + \| P_{U^\perp} x^* \|_2. \]

**Proof.** Since \( x^* = P_U x^* + P_{U^\perp} x^* \), we have
\[ \| x^* - \hat{x}^* \|_2^2 = \left( 1 - \frac{\| x^* \|}{\| P_U x^* \|} \right) \| P_U x^* + P_{U^\perp} x^* \|_2^2 \\
= \left( 1 - \frac{\| x^* \|}{\| P_U x^* \|} \right) \| P_U x^* \|_2^2 + \| P_{U^\perp} x^* \|_2^2 \\
= \left( \frac{\| P_U x^* \| - \| P_{U^\perp} x^* \|}{\| P_U x^* \|} \right)^2 \| P_U x^* \|_2^2 + \| P_{U^\perp} x^* \|_2^2 \\
\leq \left( \frac{\| P_U x^* \|}{\| P_U x^* \|} \cdot \| P_{U^\perp} x^* \|_2^2 \right) \| P_U x^* \|_2^2 + \| P_{U^\perp} x^* \|_2^2 \\
\leq \left( \frac{\| P_U x^* \|}{\| P_U x^* \|} \cdot \| P_U x^* \|_2 + \| P_{U^\perp} x^* \|_2 \right)^2. \]

\[ \blacksquare \]

### 6.2 Proofs of Theorem 2.5 and Theorem 2.8

Let us now focus on the \( \ell^1 \)-analysis (semi-)norm \( f(\cdot) = \| \Psi(\cdot) \|_1 \). The abstract results of the previous part indicate that the conic mean width \( w_\lambda(\mathcal{D}(\Psi(\cdot), x^*)) \) plays a key role in analyzing (BP\( \eta^f \)). While this quantity may indeed yield an accurate bound on the sampling rate, our desiderata (ii)–(iv) from Subsection 1.2 are still hardly met, since the definition of the mean width is quite abstract and non-informative. Therefore, we need to come up with a more explicit expression or at least a meaningful upper bound on \( w_\lambda(\mathcal{D}(\| \Psi(\cdot) \|_1, x^*)) \). This is precisely the purpose of the following theorem, which forms the main technical ingredient of our approach. Its proof is postponed to Subsection 6.3.

**Theorem 6.8** Let \( \Psi \in \mathbb{R}^{N \times n} \) be an analysis operator and let \( x^* \in \mathbb{R}^n \) with \( x^* \notin \ker \Psi \). Then
\[ w_\lambda^2(\mathcal{D}(\| \Psi(\cdot) \|_1, x^*)) \leq \mathcal{M}(\Psi, x^*). \]

Our two main results are now a direct consequence of this bound and the guarantees from Subsection 6.1:

**Proof of Theorem 2.5.** By Theorem 6.8 and the condition (2.4), we obtain
\[ m > \left( \sqrt{\mathcal{M}(\Psi, x^*) + u} \right)^2 + C^2 \cdot k^4 \cdot \left( w_\lambda(\mathcal{D}(\| \Psi(\cdot) \|_1, x^*)) + u \right)^2 + 1, \]
implying that in the ‘in particular’-part of Corollary 6.5 can be applied with \( f(\cdot) = \|\mathbf{y}(\cdot)\|_1 \). Note that \( C^2, \xi^4 = 1 \) here, due to the Gaussianity of \( A \). This already yields the claim of Theorem 2.5. ■

*Proof of Theorem 2.8.* By the definition of \( \bar{x}^* \in \mathbb{R}^n \) in (2.9), we have \( \bar{x}^* \not\in \ker \mathbf{y} \) and therefore Theorem 6.8 gives

\[
\omega_1^2(\mathcal{D}(\|\mathbf{y}(\cdot)\|_1, \bar{x}^*)) \leq \mathcal{M}(\mathbf{y}, \bar{x}^*).
\]

Moreover, we observe that \( \|\mathbf{y}x^*\|_1 = \|\mathbf{y}\bar{x}^*\|_1 \). Hence, Corollary 6.6 is applicable with \( f(\cdot) = \|\mathbf{y}(\cdot)\|_1 \) and we particularly have \( \bar{m}_0 \leq m_0 < m \); note that the constants \( C \) in the definitions of \( \bar{m}_0 \) (in (6.4)) and \( m_0 \) (in (2.10)) just differ in taking the square. The claim now follows from the error bound of (6.3). ■

### 6.3 Proof of Theorem 6.8 (Bounding the Conic Mean Width)

To complete the proof, it remains to verify Theorem 6.8. For this, we loosely follow [ALMT14, Recipe 4.1], but note again that we do not attempt to exactly compute \( \omega_1^2(\mathcal{D}(\|\mathbf{y}(\cdot)\|_1, x^*)) \) here, but rather establish an upper bound.

**Step 1: A polar bound on the conic mean width**

Our first proof step is based on a well-known polarity argument, providing an upper bound on the mean width of a descent cone. This result even holds true in the very general setup Subsection 6.1.

**Proposition 6.9 ([ALMT14, Prop. 4.1])** Let \( f : \mathbb{R}^n \to \mathbb{R} \) be convex and let \( x^* \in \mathbb{R}^n \). Then, the subdifferential of \( f \) at \( x^* \) is well-defined:

\[
\partial f(x^*) := \{ z \in \mathbb{R}^n \mid f(x) \geq f(x^*) + \langle z, x - x^* \rangle \text{ for all } x \in \mathbb{R}^n \}.
\]

If \( \partial f(x^*) \) is non-empty, compact, and does not contain the origin, we have\(^1\)

\[
\omega_1^2(\mathcal{D}(f, x^*)) \leq \inf_{\tau > 0} \mathbb{E} \left[ \inf_{z \in \partial f(x^*)} \| g - \tau \cdot z \|_2^2 \right],
\]

where \( g \sim \mathcal{N}(0, I_n) \).

**Remark 6.10** For several special cases, such as norms [ALMT14, Thm. 4.3] or total variation [ZXCL16], it is known that (6.5) provides a very accurate bound. For a more detailed discussion of the sharpness of Proposition 6.9, we refer to [ALMT14, Sec. 4.2].

In order to invoke Proposition 6.9, we need to specify the subdifferential of \( f(\cdot) = \|\mathbf{y}(\cdot)\|_1 \) at \( x^* \). For the sake of brevity, we now just write \( \mathcal{S} := \mathcal{S}_{x^*} = \text{supp}(\mathbf{y}x^*) \), \( \mathcal{S}^c := \mathcal{S}_{x^*}^c \) as well as \( \sigma := \sigma_{x^*} = \text{sign}(\mathbf{y}x^*) \).

**Lemma 6.11** We have

\[
\partial \|\mathbf{y}(\cdot)\|_1(x^*) = \mathbf{y}^T \sigma + \left\{ \mathbf{y}^T w \mid w \in \mathbb{R}^N, \text{supp}(w) \subset \mathcal{S}^c, \|w\|_\infty \leq 1 \right\}.
\]

In particular, \( \mathcal{O} \neq \partial \|\mathbf{y}(\cdot)\|_1(x^*) \subset \mathbf{y}^T [-1, 1]^N \) and the following equivalences hold true:

\[
0 \not\in \partial \|\mathbf{y}(\cdot)\|_1(x^*) \iff x^* \not\in \ker \mathbf{y} \iff \mathcal{S}^c = \|\mathbf{y}^T \sigma\|_2^2 > 0.
\]

**Proof.** First, we apply the chain rule for subdifferentials (see [Roc70, Thm. 23.9])

\[
\partial \|\mathbf{y}(\cdot)\|_1(x^*) = \mathbf{y}^T (\partial \| \cdot \|_1(\mathbf{y}x^*)),
\]

\(^1\)Hereafter, the expectation is always taken with respect to the Gaussian random vector \( g \).
and together with
\[ \partial \| \Psi \|_1(\Psi x^*) = \{ z \in \mathbb{R}^N \mid z_S = \text{sign}(\Psi x^*), \| z_S \|_\infty \leq 1 \}, \]
we obtain that
\[ \partial \| \Psi(\cdot) \|_1(x^*) = \Psi^T \sigma + \left\{ \Psi^T w \mid w \in \mathbb{R}^N, \text{supp}(w) \subset S^c, \| w \|_\infty \leq 1 \right\}. \]

To verify the 'in particular'-part, suppose that \( 0 \in \partial \| \Psi(\cdot) \|_1(x^*) \). Then there exists a \( w \in \mathbb{R}^N \) with \( \text{supp}(w) \subset S^c \) and \( \| w \|_\infty \leq 1 \) such that \( 0 = \Psi^T \sigma + \Psi^T w \). Thus,
\[ 0 = \langle x^*, \Psi^T \sigma + \Psi^T w \rangle = \langle \Psi x^*, \sigma \rangle + \langle \Psi x^*, w \rangle = \| \Psi x^* \|_1, \quad (6.7) \]
that means \( x^* \in \ker \Psi \). On the other hand, if \( x^* \in \ker \Psi \), we have \( \sigma = \text{sign}(\Psi x^*) = 0 \) and therefore \( \Psi^T \sigma = 0 \). Selecting \( w := 0 \) shows that \( 0 \in \partial \| \Psi(\cdot) \|_1(x^*) \). This proves the first equivalence.

By the definition of \( S^c_S \) (see Definition 2.3), it holds
\[ S^c_S = \langle \sigma, G\sigma \rangle = \langle \sigma, \Psi^T \sigma \rangle = \langle \Psi^T \sigma, \Psi^T \sigma \rangle = \| \Psi^T \sigma \|_2^2. \]
Thus, the second equivalence of (6.6) follows from
\[ S^c_S = \| \Psi^T \sigma \|_2^2 = 0 \iff \Psi^T \sigma = 0 \iff \sigma = \text{sign}(\Psi x^*) = 0 \iff \Psi x^* = 0, \]
where we have again made use of (6.7) with \( w = 0 \) in the second last equivalence.

In conclusion, if \( x^* \notin \ker \Psi \), Proposition 6.9 and Lemma 6.11 imply that
\[ w^*_N(\Psi(\cdot)_1, x^*)) \leq \inf_{\tau > 0} \mathbb{E} \left[ \inf_{w \in \mathbb{R}^N} \| g - \tau \Psi^T \sigma - \tau \Psi^T w \|_2^2 \right]. \quad (6.8) \]

**Step 2: Selecting the \( \ell^1 \)-dual vector \( w \).**

Evaluating the right-hand side of (6.8) particularly asks us to study the optimization program inside of the expected value. It can be reformulated as a least-squares problem with box-constraint:
\[ \min_{w \in \mathbb{R}^N} \| g - \tau \Psi^T \sigma - \tau \Psi^T w \|_2^2 \quad \text{subject to} \quad \begin{cases} w_k = 0 & \text{for } k \in S, \\ |w_k| \leq 1 & \text{for } k \in S^c. \end{cases} \quad (6.9) \]

Unfortunately, (6.9) does not seem to have a closed-form minimizer in general. Hence, instead of seeking for an exact solution, we rather consider an ansatz vector \( w' = (w'_1, \ldots, w'_N) \in \mathbb{R}^N \) of the form
\[ w'_k := \begin{cases} 0, & k \in S, \\ \text{clip}(\tau^{-1} a_k \psi_k, g); 1, & k \in S^c, \end{cases} \quad (6.10) \]
where \( a_k > 0 \) are flexible tuning parameters that will be specified later on. Note that \( w' \in [-1, 1]^N \) depends on both \( g \) and \( \tau \). In Remark 6.15 below, we will also state the KKT-conditions of (6.9), showing that the choice of (6.10) is at least plausible (although not always optimal) if the Gram matrix of \( \Psi \) is well-localized.
Since \( w'_k \) is a symmetric transformation of \( \langle \psi_k, g \rangle \sim \mathcal{N}(0, \| \psi_k \|_2^2) \) for every \( k \in S^c \), we observe that \( \mathbb{E}[w'] = 0 \). Now, we can proceed with (6.8) as follows:

\[
\begin{align*}
    &\mathbb{E} \left[ \inf_{w \in \mathbb{R}^N, \|w\|_\infty \leq 1} \| g - \tau \Psi^T \sigma - \tau \Psi^T w \|_2^2 \right] \\
    \leq &\mathbb{E} \left[ \| g - \tau \Psi^T \sigma - \tau \Psi^T w' \|_2^2 \right] \\
    = &\mathbb{E} \left[ \| g - \tau \Psi^T w' \|_2^2 \right] + \mathbb{E} \left[ \| \tau \Psi^T \sigma \|_2^2 \right] - \mathbb{E} \left[ 2 \langle g, \tau \Psi^T \sigma \rangle \right] + \mathbb{E} \left[ 2 \langle \tau \Psi^T w', \tau \Psi^T \sigma \rangle \right] \\
    = &\mathbb{E} \left[ \| g - \tau \Psi^T w' \|_2^2 \right] + \tau^2 \mathbb{E} \left[ \| \Psi^T \sigma \|_2^2 \right] - 2 \mathbb{E} \left[ \langle g, \tau \Psi^T \sigma \rangle \right] + 2 \mathbb{E} \left[ \tau \Psi^T w', \tau \Psi^T \sigma \rangle \right] \\
    = &\mathbb{E} \left[ \| g \|_2^2 \right] - 2 \tau \mathbb{E} \left[ \langle g, \Psi^T w' \rangle \right] + \tau^2 \mathbb{E} \left[ \| \Psi^T w' \|_2^2 \right] + \tau^2 \cdot S^c_S \\
    = &n - 2 \tau \mathbb{E} \left[ \langle g, \Psi^T g \rangle \right] + \tau^2 \mathbb{E} \left[ \| \Psi^T g \|_2^2 \right] + \tau^2 \cdot S^c_S = n - T_1 + T_2 + \tau^2 \cdot S^c_S. \quad (6.11)
\end{align*}
\]

Hence, it remains to resolve the expected values of \( T_1 \) and \( T_2 \).

**Step 3: A general bound on the mean width.**

Let us start with the term \( T_1 \): Using the definition of \( w' \), we obtain

\[
T_1 = 2 \tau \mathbb{E} \left[ \langle \Psi g, w' \rangle \right] = 2 \tau \sum_{k \in S^c} \mathbb{E} \left[ \langle \psi_k, g \rangle \cdot w'_k \right] \]

\[
= 2 \tau \sum_{k \in S^c} \mathbb{E} \left[ \langle \psi_k, g \rangle \cdot \text{sign} \left( \langle \psi_k, g \rangle \right) \cdot \min \left\{ \tau^{-1} a_k \| \langle \psi_k, g \rangle \|, 1 \right\} \right] \]

\[
= 2 \sum_{k \in S^c} \mathbb{E} \left[ \langle \psi_k, g \rangle \cdot \min \{ a_k \| \langle \psi_k, g \rangle \|, \tau \} \right],
\]

and since \( \langle \psi_k, g \rangle \sim \mathcal{N}(0, g_{kk}) \), an elementary calculation shows that

\[
T_1 = 2 \sum_{k \in S^c} a_k \cdot g_{kk} \cdot \text{erf} \left( \frac{\tau}{a_k \sqrt{g_{kk}}} \right).
\]

Handling the term \( T_2 \) turns out to be more complicated: Since \( \sqrt{g_{kk}} = \| \psi_k \|_2 \), we first rewrite \( T_2 \) such that the clipped variables are standard Gaussians:

\[
T_2 = \tau^2 \mathbb{E} \left[ \langle w', Gw' \rangle \right] \\
= \tau^2 \sum_{k,k' \in S^c} g_{kk'} \cdot \mathbb{E} \left[ w'_k \cdot w'_{k'} \right] \\
= \sum_{k,k' \in S^c} \mathbb{E} \left[ \text{clip} (a_k \langle \psi_k, g \rangle; \tau) \cdot \text{clip} (a_{k'} \langle \psi_{k'}, g \rangle; \tau) \right] \\
= \sum_{k,k' \in S^c} a_k \cdot a_{k'} \cdot \sqrt{g_{kk} \cdot g_{k'k'}} \cdot \mathbb{E} \left[ \text{clip} \left( \frac{\langle \psi_k, g \rangle}{\| \psi_k \|_2}, \tau \right) \cdot \text{clip} \left( \frac{\langle \psi_{k'}, g \rangle}{\| \psi_{k'} \|_2}, \tau \right) \right].
\]

(6.12)

In general, there does not exist a closed-form expression for the covariance between clipped Gaussian variables. However, we may invoke the following bound, which is based on Price’s Theorem and proven in Appendix A.2:
Lemma 6.12  Let $v_1, v_2 \in S^{n-1}$ and let $g \sim \mathcal{N}(0, I_n)$. Fix numbers $\beta_1, \beta_2 \geq 0$ and set $\beta_{\min} := \min\{\beta_1, \beta_2\}$, $\beta_{\max} := \max\{\beta_1, \beta_2\}$. Then, we have

$$
\mathbb{E}\left[\text{clip}((v_1, g) ; \beta_1) \cdot \text{clip}((v_2, g) ; \beta_2)\right] \\
\leq \left|\langle v_1, v_2 \rangle\right| \cdot \left[\text{erf}\left(\frac{\beta_{\min}}{\sqrt{2}}\right) + \left(1 - \text{erf}\left(\frac{\beta_{\max}}{\sqrt{2}}\right)\right) \cdot \beta_{\min} \cdot \beta_{\max}\right] \\
= \left|\langle v_1, v_2 \rangle\right| \cdot \left[\text{erf}\left(\frac{\beta_{\min}}{\sqrt{2}}\right) - h(\beta_{\max}) \cdot \beta_{\min} \cdot \beta_{\max}\right],
$$

(6.13)

where $h(\tau) = \sqrt{\frac{2}{\pi}} \cdot e^{-\frac{\tau^2}{2}} + \text{erf}(\frac{\tau}{\sqrt{2}}) - 1$; see (2.3). If $v_1 = v_2$, (6.13) holds true with equality.

Let us apply Lemma 6.12 to equation (6.12) by setting

$$
v_1 := \frac{\Psi_k}{\|\Psi_k\|_2}, \quad v_2 := \frac{\Psi_{k'}}{\|\Psi_{k'}\|_2}, \quad \beta_1 := \frac{\tau}{a_k \sqrt{S_k}}, \quad \beta_2 := \frac{\tau}{a_{k'} \sqrt{S_{k'}}}.
$$

Observing that $\frac{N_{k,k'}}{\sqrt{S_k S_{k'}}} = \langle v_1, v_2 \rangle$, we end up with

$$
T_2 \leq \sum_{k,k' \in S^c} g^2_{k,k'} \cdot a_k \cdot a_{k'} \cdot \left(\text{erf}\left(\frac{\tau \beta_{\min}^{k,k'}}{\sqrt{2}}\right) - h(\tau \beta_{\max}^{k,k'}) \cdot \beta_{\min} \cdot \beta_{\max}\right),
$$

where

$$
\beta_{\min}^{k,k'} := \min\left\{\frac{1}{a_k \sqrt{S_k}}, \frac{1}{a_{k'} \sqrt{S_{k'}}}\right\} \quad \text{and} \quad \beta_{\max}^{k,k'} := \max\left\{\frac{1}{a_k \sqrt{S_k}}, \frac{1}{a_{k'} \sqrt{S_{k'}}}\right\}.
$$

(6.14)

Combining our findings for $T_1$ and $T_2$ with the estimates of (6.8) and (6.11) yields the following interim result:

**Proposition 6.13**  Let $\Psi \in \mathbb{R}^{N \times \mu}$ be an analysis operator and let $x^* \in \mathbb{R}^n$ with $x^* \notin \ker\Psi$. For fixed numbers $\alpha_k > 0$, $k \in S^c$, it holds

$$
w^2_a(\mathcal{D}(\|\Psi(\cdot)\|_1, x^*)) \leq \inf_{\tau > 0} \left\{n + \tau^2 \cdot S^c - 2 \sum_{k \in S^c} \alpha_k \cdot g_{k,k} \cdot \text{erf}\left(\frac{\tau}{\alpha_k \sqrt{S_k}}\right) + \sum_{k,k' \in S^c} g^2_{k,k'} \cdot a_k \cdot a_{k'} \cdot \left(\text{erf}\left(\frac{\tau \beta_{\min}^{k,k'}}{\sqrt{2}}\right) - h(\tau \beta_{\max}^{k,k'}) \cdot \beta_{\min} \cdot \beta_{\max}\right)\right\},
$$

(6.15)

where $\beta_{\min}^{k,k'}$ and $\beta_{\max}^{k,k'}$ are defined according to (6.14).

The bound of Proposition 6.13 is in fact quite general and leaves much space for selecting the parameters $\alpha_k$. However, we still need to optimize over $\tau$, which seems to be a very challenging task in general. Our final step is therefore to make a specific choice of $\alpha_k$, eventually leading to the explicit bound of Theorem 6.8.

**Step 4: Selecting the tuning parameters $\alpha_k$.**

In order to simplify the optimization problem of (6.15), we would like to factor out all terms depending on $\tau$. Such a separation is easily obtained with $\alpha_k := \frac{1}{\sqrt{S_k}}$ for $k \in S^c$ and some $\lambda > 0$ which is fixed later on. With this choice of $\alpha_k$, (6.15) reads as follows:
\[ w_\lambda^2(D(\|\Psi(\cdot)\|_1, x^*)) \leq \inf_{\lambda > 0} \left\{ n + \tau^2 \cdot S_S^c - 2\lambda \cdot \left( \sum_{k \in S^c} \sqrt{8k} \cdot k \right) \cdot \text{erf} \left( \frac{\tau}{\sqrt{2}} \right) + \lambda^2 \cdot \left( \sum_{k \in S^c} \frac{\sqrt{s_k^c}}{\sqrt{8k \cdot k}} \right) \cdot \left( \text{erf} \left( \frac{\tau}{\sqrt{2}} \right) - h \left( \frac{\tau}{\sqrt{2}} \right) \right) \right\}, \tag{6.16} \]

and substituting \( \tau \) by \( \lambda \cdot \tau \) leads to

\[ w_\lambda^2(D(\|\Psi(\cdot)\|_1, x^*)) \leq \inf_{\lambda > 0} \left\{ n + \left( S_S^c - L_S^c \cdot h(\tau) \right) \tau^2 \lambda^2 + \left( \lambda^2 \cdot L_S^c - 2\lambda L_S^c \right) \cdot \text{erf} \left( \frac{\tau}{\sqrt{2}} \right) \right\}. \tag{6.16} \]

It is not hard to see that \( (\tau, \lambda) \mapsto F(\tau, \lambda) \) defines a smooth function that can be continuously extended to \([0, \infty) \times [0, \infty)\). The next lemma shows that \( F \) has in fact a unique global minimum. Its technical proof relies on basic calculus and is deferred to Appendix A.3.

**Lemma 6.14**  Assuming that \( S_S^c, L_S^c, L_S^c > 0 \), the function \( F: [0, \infty) \times [0, \infty) \rightarrow \mathbb{R} \), defined according to (6.16), attains its unique global minimum at

\[(\tau', \lambda') := \left( h^{-1} \left( \frac{S_S^c}{L_S^c}, \frac{L_S^c}{L_S^c} \right) \right) \).

Now, we can easily derive the statement of Theorem 6.8.

**Proof of Theorem 6.8.** First, suppose that \( S^c \neq \emptyset \). This implies that \( L_S^c, L_S^c > 0 \), and since \( x^* \notin \ker \Psi \), we also have \( S_S^c > 0 \) by Lemma 6.11. Lemma 6.14 and (6.16) then yield

\[
w_\lambda^2(D(\|\Psi(\cdot)\|_1, x^*)) \leq F(\tau', \lambda')
= n + \left( S_S^c - L_S^c \cdot h(\tau') \right) \cdot (\tau')^2 \cdot (\lambda')^2 + \left( \lambda^2 \cdot L_S^c - 2\lambda \cdot L_S^c \right) \cdot \left( \text{erf} \left( \frac{\tau'}{\sqrt{2}} \right) - \frac{1}{\sqrt{2}} \cdot h^{-1} \left( \frac{S_S^c}{L_S^c}, \frac{L_S^c}{L_S^c} \right) \right)
= n - \frac{\left( L_S^c \right)^2}{L_S^c} \cdot \left( \frac{1}{\sqrt{2}} \cdot h^{-1} \left( \frac{S_S^c}{L_S^c}, \frac{L_S^c}{L_S^c} \right) \right)
= n - \frac{\left( L_S^c \right)^2}{L_S^c} \cdot \Phi \left( \frac{S_S^c}{L_S^c} \right) = M(\Psi, x^*). \tag{6.17} \]

On the other hand, if \( S^c = \emptyset \), we simply observe that

\[
w_\lambda^2(D(\|\Psi(\cdot)\|_1, x^*)) \leq \lim_{\tau, \lambda \downarrow 0} F(\tau, \lambda) = F(0, 0) = n = M(\Psi, x^*). \]

**Remark 6.15** (KKT-Conditions and Optimality) A key step in the above proof was to select the dual vector \( w' \) in (6.10), which eventually enabled us to derive a closed-form bound on the conic mean width \( w_\lambda^2(D(\|\Psi(\cdot)\|_1, x^*)) \). Such a choice is however not always optimal. Indeed, any minimizer \( w \in \mathbb{R}^N \) of (6.9) fulfills the following KKT-conditions (cf. [MR10]):

\[ \left( \tau^2 \Psi_S^c (\Psi_S^c)^T + \text{diag}(\mu_S^c) \right) w_S^c = \tau \Psi_S^c g - \tau^2 \Psi_S^c \Psi_S^c^T c, \]

\[
\begin{align*}
\mu_k &\geq 0, \quad k \in [N], \\
\mu_k (1 - w_k) &= 0, \quad k \in S^c, \\
\mu_k^2 &\leq 1, \quad k \in S^c, \\
w_k &= 0, \quad k \in S,
\end{align*}
\]
for some $\mu = (\mu_1, \ldots, \mu_N) \in \mathbb{R}^N$. Supposed that $G = \Psi^T = \text{diag}(a_1^{-1}, \ldots, a_N^{-1})$ for certain numbers $a_k > 0$, it is not hard to see that $w = w'$ from (6.10) precisely satisfies these conditions. This observation particularly explains why our sampling-rate bound is optimal in the case of orthonormal bases (cf. Proposition 2.7), where $G = I_n$.

On the other hand, if the operator $\Psi$ is redundant, i.e., $N > n$, the Gram matrix cannot be diagonal anymore and the solution of (6.9) may take a different form. But as long as $G$ is well-localized, in the sense that $G = \Psi^T \approx \text{diag}(a_1^{-1}, \ldots, a_N^{-1})$, one can expect that (6.10) is at least relatively close to the true minimizer. While this heuristic actually suggests to set $a_k := 1/\gamma_{kk}$ for $k \in S^c$, it is not clear how to proceed with (6.15) in that situation. Therefore, we have made a slightly different choice of $a_k$ in Step 4 above. But as our numerical simulations in Section 3 indicate, this strategy still works surprisingly well for many different classes of analysis operators.

\[ \therefore \]

### 6.4 Proof of Proposition 2.7 (Simplified Sample Complexity Bound)

We prove that $\mathcal{M}(\Psi, x^*)$ is upper bounded by each of the two expressions on the right-hand side of (2.5). Recall that the left branch of the minimum is supposed to mimic the (asymptotic) behavior of $\mathcal{M}(\Psi, x^*)$ if the ratio $S_{S^c_x}/L_{S^c_x}$ is small, whereas the right branch is appropriate as $S_{S^c_x}/L_{S^c_x}$ becomes large.

**Proof of Proposition 2.7.** By Lemma 6.11 and $x^* \not\in \ker \Psi$, we know that $S = S_{S^c_x}/L_{S^c_x} > 0$, while the assumption $S_{S^c_x} \neq \emptyset$ implies that $L = L_{S^c_x} > 0$ and $L = L_{S^c_x} > 0$. Therefore, by Lemma 6.14 and (6.17), we have

\[ \mathcal{M}(\Psi, x^*) = F(\tau', \lambda') \leq F(\tau, \lambda') \]

for all $\tau > 0$. Since $h(\tau) \geq 0$ and $\text{erf}(\tau/\sqrt{2}) \geq 1 - e^{-\tau^2/2}$ for all $\tau > 0$ (see [FR13, Prop. 7.5]), one can bound $F(\tau, \lambda')$ as follows:

\[ F(\tau, \lambda') = n + (S - L \cdot h(\tau)) \cdot \tau^2 \cdot (\frac{1}{L})^2 - \frac{L^2}{L} \cdot \text{erf}(\frac{\tau}{\sqrt{2}}) \]

\[ \leq n + S \cdot (\frac{1}{L})^2 \cdot \tau^2 - \frac{L^2}{L} \cdot (1 - e^{-\tau^2/2}) \]

\[ = n - \frac{L^2}{L} + (\frac{1}{L})^2 \cdot (S \cdot \tau^2 + L \cdot e^{-\tau^2/2}). \]

Finally, setting $\tau = \tau'' := \sqrt{2 \log \left( \frac{S + L}{S} \right)} > 0$, we obtain

\[ \mathcal{M}(\Psi, x^*) \leq F(\tau'', \lambda') = n - \frac{L^2}{L} + (\frac{1}{L})^2 \cdot \left( 2S \cdot \log \left( \frac{S + L}{S} \right) + \frac{lS}{S} \right) \leq n - \frac{L^2}{L} + (\frac{1}{L})^2 \cdot \left( 2S \cdot \log \left( \frac{S + L}{S} \right) + S \right). \]

It remains to verify that $\mathcal{M}(\Psi, x^*) \leq n - \frac{2}{\pi} \cdot \frac{L^2}{S + L}$ holds true as well. By the definition of $\mathcal{M}(\Psi, x^*)$, this is equivalent to

\[ \Phi \left( \frac{S}{L} \right) = \text{erf} \left( \frac{\sqrt{2}}{L} \cdot h^{-1} \left( \frac{S}{L} \right) \right) \geq \frac{2}{\pi} \cdot \frac{L}{S + L} = \frac{2}{\pi} \cdot \frac{1}{\frac{S}{L} + 1} \]

\[ \xi := h^{-1}(S/L) \iff \text{erf}(\frac{\sqrt{2}}{L}) \geq \frac{2}{\pi} \cdot \frac{1}{h(\xi) + 1}, \quad \xi > 0 \]

\[ \iff G(\xi) := \text{erf}(\frac{\sqrt{2}}{L}) \cdot (h(\xi) + 1) \geq \frac{2}{\pi}, \quad \xi > 0. \]
First, we observe that
\[ \lim_{\xi \searrow 0} G(\xi) = \lim_{\xi \searrow 0} \left[ \text{erf}(\frac{\xi}{\sqrt{2}}) \cdot \sqrt{\frac{2}{\pi} e^{-\xi^2/2}} \right] = \frac{2}{\pi}. \]

The claim then follows from the fact that \( G \) is monotonically increasing, which is in turn a consequence of \( G'(\xi) > 0 \) for all \( \xi > 0 \): indeed, we have
\[ 0 < G'(\xi) = \sqrt{\frac{2}{\pi} e^{-\xi^2/2}} \cdot \left( (\xi^2 - 1) \cdot \text{erf}(\frac{\xi}{\sqrt{2}}) + \xi \cdot \sqrt{\frac{2}{\pi} e^{-\xi^2/2}} \right) \]
\[ \Leftrightarrow 0 < \tilde{G}(\xi) := (\xi^2 - 1) \cdot \text{erf}(\frac{\xi}{\sqrt{2}}) + \xi \cdot \sqrt{\frac{2}{\pi} e^{-\xi^2/2}}, \]
while the latter statement holds true due to \( \tilde{G}(0) = 0 \) and
\[ \tilde{G}'(\xi) = 2\xi \cdot \text{erf}(\frac{\xi}{\sqrt{2}}) > 0, \quad \xi > 0. \]

## A Proofs of Auxiliary Results

### A.1 Proofs of Recovery Results From Subsection 6.1

First, let us state an important preliminary result from random matrix theory.

**Theorem A.1** (Minimum Spheric Singular Value) Assume that Model 2.1 is satisfied and let \( K \subset S^{n-1} \).
There exists a numerical constant \( C > 0 \) such that, for every \( u > 0 \), we have
\[ \sup_{h \in K} \|Ah\|_2 > \sqrt{m - 1} - C \cdot \kappa^2 \cdot (w(K) + u) \]
with probability at least \( 1 - e^{-u^2/2} \). If \( a \sim \mathcal{N}(0, I_n) \), we have \( C = \kappa = 1 \).

**Proof.** This statement is a consequence of [LMPV17, Thm. 1.3]. To obtain the optimal constants in the Gaussian case, one can use [Tro15, Prop. 3.3] instead. \hfill \(|)

The Gaussian case of Theorem A.1 is well-known as *Gordon’s Escape Through a Mesh Theorem*, originating from [Gor88]. The above formulation is typically interpreted as a lower bound on the minimum singular value of the random matrix \( A \in \mathbb{R}^{m \times n} \) restricted to \( \text{cone}(K) \).

With this result at hand, we are now ready to prove Theorem 6.4:

**Proof of Theorem 6.4.** Any minimizer \( \hat{x} \) of \( (P_0') \) satisfies \( \hat{x} - x^* \in \mathcal{D}(f, x^*) \) as well as \( \|Ax - y\|_2 \leq \eta \). Our goal is therefore to show that every vector \( x \in \mathbb{R}^n \) with \( \|x - x^*\|_2 > t \) cannot be a minimizer, if both \( x - x^* \in \mathcal{D}(f, x^*) \) and \( \|Ax - y\|_2 \leq \eta \) are fulfilled.

First, we apply Theorem A.1 for \( K = \frac{1}{t} \mathcal{D}(f, x^*) \cap S^{n-1} \): Let \( h \in tK \) and assume \( \|Ax - y\|_2 \leq \eta \) where \( x := x^* + h \). Without loss of generality, suppose that
\[ \sqrt{m - 1} > C \cdot \kappa^2 \cdot (w_t(\mathcal{D}(f, x^*)) + u), \]
since otherwise, the condition (6.1) is not satisfied anyway. On the event of Theorem A.1, we have
\[ \delta := \left[ \sqrt{m - 1} - C \cdot \kappa^2 \cdot (w_t(\mathcal{D}(f, x^*)) + u) \right]_+ = \sqrt{m - 1} - C \cdot \kappa^2 \cdot (w(K) + u) \]
\[ < \frac{1}{t} \|Ah\|_2 \leq \frac{1}{t} (\|Ah - e\|_2 + \|e\|_2) = \frac{1}{t} (\|Ax - y\|_2 + \|e\|_2) \leq \frac{2\eta}{t}. \]
Rearranging the latter inequality gives \( t < 2\eta / \delta \) which contradicts the assumption (6.1). This implies that \( x = x^* + h \) cannot be a minimizer of \( (P_0') \).
Now, assume that $\hat{x} \in \mathcal{D}(f, x^*) + x^*$ is a minimizer of $\mathcal{P}_0^t$ with $\|\hat{x} - x^*\|_2 > t$. Setting $\hat{h} := \|\hat{x} - x^*\|$, we have $\delta := \|\hat{h}\|_2 > t$. The convexity of $\mathcal{D}(f, x^*)$ and $0 = x^* - x^* \in \mathcal{D}(f, x^*)$ imply that there exists $h \in tK = \mathcal{D}(f, x^*) \cap tS^{n-1}$ such that $\hat{h} = \frac{1}{t}h$. In other words, $h$ is the intersection point between $tS^{n-1}$ and the line that connects $\hat{h}$ with $0$. Hence, on the event of Theorem A.1, we also observe that

$$\delta < \frac{1}{t} \|Ah\|_2 = \frac{1}{t} \|A\hat{h}\|_2 \leq \frac{2\eta}{t} < \frac{2\eta}{T},$$

which is again a contradiction to (6.1). From this, we can conclude that any minimizer $\hat{x}$ of $\mathcal{P}_0^t$ satisfies $\|\hat{x} - x^*\|_2 \leq t$. \hfill \qed

**Proof of Corollary 6.5.** First, assume that $\eta > 0$. We would like to apply Theorem 6.4 with

$$t := \frac{2\eta}{\sqrt{m - 1 - C \cdot \kappa^2 \cdot (w_\Lambda(\mathcal{D}(f, x^*)) + u)}_+}.$$ 

Since $w_\Lambda(\mathcal{D}(f, x^*)) \geq w_1(\mathcal{D}(f, x^*))$ by Remark 6.2, the assumption (6.1) is satisfied and the claim now follows from Theorem 6.4.

Now, let us consider the noiseless case of $\eta = 0$. Similar to the proof of Theorem 6.4, we apply Theorem A.1 with $K = \mathcal{D}_\Lambda(f, x^*) \cap S^{n-1}$. Let $h \in K$ and $x := x^* + h$ with $\|Ax - y\|_2 = \|Ah\|_2 = 0$. On the event of Theorem A.1, we have

$$0 < \sqrt{m - 1 - C \cdot \kappa^2 \cdot (w_\Lambda(\mathcal{D}(f, x^*)) + u)} < \|Ah\|_2 = 0,$$ 

(6.2)

which is a contradiction. If $h \in \mathcal{D}_\Lambda(f, x^*) \setminus \{0\}$ with $\|Ah\|_2 = 0$, we just invoke (6.1) for $h/\|h\|_2$. Hence, the only feasible point of $\mathcal{P}_0^t$ is $x^*$, so that $\hat{x} = x^*$. \hfill \qed

To prove Corollary 6.6, we need the following lemma. It shows how to select a scale $t > 0$ for the local mean width such that it can be bounded by the conic mean width of a surrogate signal:

**Lemma A.2** Let $x^* \in \mathbb{R}^n$ and let $K \subset \mathbb{R}^n$ be convex. For every $t > 0$ and $\hat{x}^* \in K$, we have

$$w_t(K - x^*) \leq \frac{t + \|\hat{x}^* - x^*\|_2}{t} \cdot (w_\Lambda(K - \hat{x}^*) + 1).$$

In particular, $w_{R\|\hat{x}^* - x^*\|_2}(K - x^*) \leq \frac{R + 1}{R} \cdot (w_\Lambda(K - \hat{x}^*) + 1)$ for every $R > 0$.

**Proof.** We set $t' := \|\hat{x}^* - x^*\|_2$. For every $x \in K \cap (x^* + tS^{n-1})$, we have

$$\|x - x^*\|_2 \leq \|x - x^*\|_2 + \|x^* - \hat{x}^*\|_2 \leq t + t',$$

and therefore, $x \in K_{x^* + t + t'} := K \cap (x^* + (t + t')B_2^n)$. This implies

$$w_t(K - x^*) = \mathbb{E} \left[ \sup_{h \in \left( K - x^* \right) \cap S^{n-1}} (g, h) \right] = \frac{1}{t} \mathbb{E} \left[ \sup_{x \in K \cap (x^* + tS^{n-1})} (g, x - x^*) \right]$$

$$\leq \frac{1}{t} \mathbb{E} \left[ \sup_{x \in K_{x^* + t + t'}} (g, x - \hat{x}^*) \right] + \frac{1}{t} \mathbb{E} \left[ \sup_{x \in K \cap (x^* + tS^{n-1})} (g, \hat{x}^* - x^*) \right]$$

$$= \frac{t + t'}{t} \cdot \frac{1}{t + t'} \mathbb{E} \left[ \sup_{x \in K_{x^* + t + t'}} (g, x - \hat{x}^*) \right]$$

$$= \frac{t + t'}{t} \cdot \mathbb{E} \left[ \sup_{h \in \left( K - x^* \right) \cap B_2^n} (g, h) \right].$$
\[ x^* \in K \leq \frac{t + \ell}{t} \cdot \mathbb{E}[\sup_{h \in \text{cone}(K - x^*) \cap B_2^n} (g, h)] \]

\[ \text{CSI} \leq \frac{t + \ell}{t} \cdot \left( \mathbb{E}\left[\left(\sup_{h \in \text{cone}(K - x^*) \cap B_2^n} (g, h)\right)^2\right]\right)^{1/2} \]

\[ \leq \frac{t + \ell}{t} \cdot (w^2_{\lambda}(K - x^*) + 1)^{1/2} \leq \frac{t + \ell}{t} \cdot (w_{\lambda}(K - x^*) + 1), \]

where (*) follows from \cite{ALMT14, Prop. 3.1(5) and Prop. 10.2].

**Proof of Corollary 6.6.** We would like to apply Theorem 6.4 with

\[ t = \max \left\{ R\|x^* - \bar{x}^*\|_2, \frac{2\eta}{\sqrt{m - 1} - C \cdot \kappa^2 \cdot (\frac{R + 1}{\kappa} \cdot (w_{\lambda}(\mathcal{D}(f, \bar{x}^*)) + 1) + u)} \right\}. \]

Thus, it suffices verify that the assumption (6.1) is satisfied. For this, we first observe that

\[ w_t(\mathcal{D}(f, x^*)) = w_t(\mathcal{D}(f, x^*) + x^* - x^*) \leq w_R(\|x^* - x^*\|_2) \leq (\mathcal{D}(f, x^*) + x^*) - x^*) \]

\[ \leq (w_{\lambda}(\mathcal{D}(f, x^*)) + 1), \]

where we have used the monotony of the local mean width in the second step, and the last step follows from

\[ \mathcal{D}(f, x^*) + x^* - x^* \subseteq \{ h \in \mathbb{R}^n \mid f(x^* + h) \leq f(x^*) \} = \mathcal{D}(f, \bar{x}^*). \]

This implies \(^1\)

\[ t \geq \frac{2\eta}{\sqrt{m - 1} - C \cdot \kappa^2 \cdot (\frac{R + 1}{\kappa} \cdot (w_{\lambda}(\mathcal{D}(f, \bar{x}^*)) + 1) + u)} \]

\[ \geq \left( \frac{2\eta}{\sqrt{m - 1} - C \cdot \kappa^2 \cdot (w_t(\mathcal{D}(f, x^*)) + u)} \right)_+, \]

so that the claim again follows from Theorem 6.4.

**A.2 Proof of Lemma 6.12**

The main idea behind the proof of Lemma 6.12 is to regard the “clipped covariance”

\[ \mathbb{E}[\text{clip}(\langle v_1, g \rangle; \beta_1) \cdot \text{clip}(\langle v_2, g \rangle; \beta_2)] \]

as a function in the covariance between \(\langle v_1, g \rangle\) and \(\langle v_2, g \rangle\), which is given by \(\mathbb{E}[\langle v_1, g \rangle \cdot \langle v_2, g \rangle] = \langle v_1, v_2 \rangle\). In fact, we will even see that it is monotone and (6.13) holds with equality if \(|\langle v_1, v_2 \rangle| = 1\). To make this argument precise, we need the following

\(^1\)This argument does actually not cover the case of \(\eta = 0\), but here we can use that \(t = R\|x^* - \bar{x}^*\|_2 > 0\) if \(x^* \neq \bar{x}^*\).
Lemma A.3 Let $\beta_1, \beta_2 \geq 0$ and define the function

$$\Theta: [-1, 1] \to \mathbb{R}, \rho \mapsto E_{(g_1, g_2) \sim N\left(0, \left[\begin{array}{c} 1 \\ \rho \end{array}\right] \right)} \left[ \text{clip}(g_1; \beta_1) \cdot \text{clip}(g_2; \beta_2) \right].$$

Then, $\Theta$ is continuous, $\Theta(0) = 0$, and $\Theta$ is convex on $[0, 1]$. In particular, $\Theta(\rho) \leq \rho \cdot \Theta(1)$ for all $0 \leq \rho \leq 1$.

The proof of Lemma A.3 is a consequence of Price’s Theorem [McM64; Pri58], which is a powerful tool to calculate the derivative of non-linearly distorted, correlated Gaussian random variables. The following version of Price’s Theorem was established by Voigtlaender [Voi17] and even applies to tempered distributions:

Theorem A.4 (Price, [Voi17, Cor. 2]) For $\rho \in (-1, 1)$, let

$$\phi_p: \mathbb{R}^2 \to (0, \infty), (x_1, x_2) \mapsto \frac{1}{2\pi\sqrt{1-\rho^2}} \cdot \exp \left(-\frac{x_1^2 + x_2^2 - 2\rho x_1 x_2}{2(1-\rho^2)} \right)$$

be the density function of a random vector $g \sim N\left(0, \left[\begin{array}{c} 1 \\ \rho \end{array}\right] \right)$. If $f \in \mathcal{S}'(\mathbb{R}^2)$ is a tempered distribution, then the function

$$\mathcal{F}: (-1, 1) \to \mathbb{C}, \rho \mapsto \langle f, \phi_p \rangle_{\mathcal{S}', \mathcal{S}}$$

is infinitely often differentiable with

$$\mathcal{F}^{(k)}(\rho) = \left\langle \frac{d^k}{dx_1^k dx_2^k} f, \phi_p \right\rangle_{\mathcal{S}', \mathcal{S}}$$

for all $\rho \in (-1, 1)$ and $k \geq 0$.

In particular, if $f \in L^1_{\text{loc}}(\mathbb{R}^2)$, we have $\mathcal{F}(\rho) = E[f(g)]$ for $g \sim N\left(0, \left[\begin{array}{c} 1 \\ \rho \end{array}\right] \right)$.

Proof of Lemma A.3. We consider the function

$$f_{\beta_1, \beta_2}: \mathbb{R}^2 \to \mathbb{R}, (x_1, x_2) \mapsto \text{clip}(x_1; \beta_1) \cdot \text{clip}(x_2; \beta_2).$$

It is not hard to see that $f_{\beta_1, \beta_2}$ is 1-Lipschitz and therefore $f_{\beta_1, \beta_2} \in \mathcal{S}'(\mathbb{R}^2)$. The (weak) derivative of $f_{\beta_1, \beta_2}$ is given by the set indicator function of the box $(-\beta_1, \beta_1) \times (-\beta_2, \beta_2) \subset \mathbb{R}^2$:

$$\frac{d^2}{dx_1 dx_2} f_{\beta_1, \beta_2} = \mathcal{X}_{(-\beta_1, \beta_1) \times (-\beta_2, \beta_2)}.$$

For any $\phi \in \mathcal{S}(\mathbb{R}^2)$, Fubini’s theorem and the fundamental theorem of calculus now yield

$$\left\langle \frac{d^k}{dx_1^k dx_2^k} f_{\beta_1, \beta_2}, \phi \right\rangle_{\mathcal{S}', \mathcal{S}} = \left\langle \frac{d^2}{dx_1 dx_2} \mathcal{X}_{(-\beta_1, \beta_1) \times (-\beta_2, \beta_2)}, \frac{d^2}{dx_1 dx_2} \phi \right\rangle_{\mathcal{S}', \mathcal{S}}$$

$$= \int_{-\beta_2}^{\beta_2} \int_{-\beta_1}^{\beta_1} \frac{d^2}{dx_1 dx_2} \phi(x_1, x_2) dx_1 dx_2$$

$$= \phi(\beta_1, \beta_2) - \phi(-\beta_1, -\beta_2) - \phi(-\beta_1, \beta_2) + \phi(-\beta_1, -\beta_2).$$

Hence, by Theorem A.4, the map $\Theta$ is smooth on $(-1, 1)$ and

$$\Theta''(\rho) = \phi_p(\beta_1, \beta_2) - \phi_p(\beta_1, -\beta_2) - \phi_p(-\beta_1, \beta_2) + \phi_p(-\beta_1, -\beta_2)$$

$$= 2(\phi_p(\beta_1, \beta_2) - \phi_p(\beta_1, -\beta_2))(A.2)$$

for $\rho \in (-1, 1)$, where we have also used the symmetry of $\phi_p$. Now, let us show the continuity of $\Theta$ on the interval $[-1, 1]$. For this, let $g := (g, g') \sim N\left(0, I_2\right)$ and set $g_\rho := \rho \cdot g + \sqrt{1-\rho^2} \cdot g'$ for $\rho \in [-1, 1]$.

\[1\] For a detailed introduction to Schwartz functions and tempered distributions, see [Fol99, Chap. 8+9] for example.
Then \((g, g_\rho) \sim \mathcal{N}\left(0, \left[\frac{1}{\rho} \right]\right)\), so that
\[
|\Theta(\rho) - \Theta(\rho')| \leq E_g[|f_{\beta_1, \beta_2}(g, g_\rho) - f_{\beta_1, \beta_2}(g, g_{\rho'})|] \\
\text{ if } f_{\beta_1, \beta_2} \text{ is 1-Lipschitz} \\
\leq E_g[|g_\rho - g_{\rho'}|] \\
\leq |\rho - \rho'| \cdot E[|g|] + |\sqrt{1 - \rho} - \sqrt{1 - \rho'}| \cdot E[|g'|] \xrightarrow{\rho' \to \rho} 0.
\]
Moreover, we observe that
\[
\Theta(0) = E_g[f_{\beta_1, \beta_2}(g, g')] = E[\text{clip}(g; \beta_1)] \cdot E[\text{clip}(g'; \beta_1)] = 0.
\]
Finally, let us verify the convexity of \(\Theta\) on \([0, 1]\). It is sufficient to show \(\Theta''(\rho) \geq 0\) for \(\rho \in (0, 1)\). Indeed, by (A.2), we have
\[
\Theta''(\rho) = 2(\beta_1^2 \rho_1^2 - \beta_1 \rho_1 - \beta_2) \geq 0 \\
\iff \exp\left(-\frac{\beta_1^2 + \beta_2^2 - 2\rho \beta_1 \beta_2}{2(1-\rho^2)}\right) \geq \exp\left(-\frac{\beta_1^2 + \beta_2^2 + 2\rho \beta_1 \beta_2}{2(1-\rho^2)}\right) \\
\iff \beta_1^2 + \beta_2^2 - 2\rho \beta_1 \beta_2 \leq \beta_1^2 + \beta_2^2 + 2\rho \beta_1 \beta_2 \iff 0 \leq 4\rho \beta_1 \beta_2,
\]
where the latter inequality is obviously fulfilled for every \(\rho \in (0, 1)\).

Now, using Lemma A.3, we can easily derive the statement of Lemma 6.12:

**Proof of Lemma 6.12.** First, we observe that
\[
\langle v_1, g \rangle, \langle v_2, g \rangle \sim \mathcal{N}\left(0, \left[\frac{1}{\langle v_1, v_2 \rangle} \langle v_1, v_2 \rangle \right]\right).
\]
If \(\langle v_1, v_2 \rangle \geq 0\), Lemma A.3 implies
\[
E[\text{clip}(\langle v_1, g \rangle; \beta_1) \cdot \text{clip}(\langle v_2, g \rangle; \beta_2)] = \Theta(\langle v_1, v_2 \rangle) \leq \langle v_1, v_2 \rangle \cdot \Theta(1),
\]
and an elementary calculation shows that
\[
\Theta(1) = E[\text{clip}(g; \beta_1) \cdot \text{clip}(g; \beta_2)] \\
\sim \mathcal{N}(0, 1) \\
= \left(1 - \text{erf}\left(\frac{\rho_{\max}}{\sqrt{2}}\right) - \sqrt{\frac{2}{\pi}} e^{-\langle\rho_{\max}, \rho_{\max}\rangle/2} \right) \cdot \beta_{\min} \cdot \beta_{\max} + \text{erf}\left(\frac{\rho_{\min}}{\sqrt{2}}\right).
\]
On the other hand, if \(\langle v_1, v_2 \rangle \leq 0\), we have
\[
E[\text{clip}(\langle v_1, g \rangle; \beta_1) \cdot \text{clip}(\langle v_2, g \rangle; \beta_2)] = -E[\text{clip}(\langle -v_1, g \rangle; \beta_1) \cdot \text{clip}(\langle v_2, g \rangle; \beta_2)] \\
= -\Theta(\langle -v_1, v_2 \rangle) \\
\langle -v_1, v_2 \rangle \geq 0 \\
\geq -\langle -v_1, v_2 \rangle \cdot \Theta(1) \\
= \langle v_1, v_2 \rangle \cdot \Theta(1),
\]
and by taking the absolute value, the claim follows.
A.3 Proof of Lemma 6.14

Before proving the actual statement of Lemma 6.14, we recall the definition of the function \( h \) defined in (2.3):

\[
h: (0, \infty) \to (0, \infty), \ \tau \mapsto \sqrt{\frac{2}{\pi}} e^{-\tau^2/2} + \text{erf}\left(\frac{\tau}{\sqrt{2}}\right) - 1.
\]

The well-definedness of \( h \), i.e., \( h(\tau) > 0 \) for all \( \tau > 0 \), is a consequence of standard bounds on the error function, e.g., see [FR13, Prop. 7.5]. Since \( \lim_{\tau \to 0} h(\tau) = \infty \), \( \lim_{\tau \to \infty} h(\tau) = 0 \), and

\[
h'(\tau) = -\sqrt{\frac{2}{\pi}} e^{-\tau^2/2} < 0
\]

for all \( \tau > 0 \), we can conclude that \( h^{-1}: (0, \infty) \to (0, \infty) \) is also well-defined.

Proof of Lemma 6.14. Adapting the notation from Proposition 2.7, let us set \( S := S_S^e \), \( L := L_{S^e} \), \( L := L_{S^e} \), and for the sake of convenience, let us also restate the definition of \( F \):

\[
F(\tau, \lambda) = n + (S - L \cdot h(\tau)) \cdot \tau^2 \cdot \lambda^2 + (\lambda^2 \cdot L - 2\lambda \cdot \bar{L}) \cdot \text{erf}\left(\frac{\tau}{\sqrt{2}}\right), \quad \tau, \lambda \geq 0.
\]

First, we calculate the partial derivatives of \( F \) for \( \tau, \lambda > 0 \):

\[
\begin{align*}
\frac{\partial F}{\partial \tau}(\tau, \lambda) &= 2\lambda^2 \cdot \tau \cdot (S - L \cdot h(\tau)) + 2\lambda \cdot \sqrt{\frac{2}{\pi}} e^{-\tau^2/2} \cdot (\lambda \cdot L - \bar{L}), \\
\frac{\partial F}{\partial \lambda}(\tau, \lambda) &= 2\lambda^2 \cdot \tau^2 \cdot (S - L \cdot h(\tau)) + 2 \text{erf}\left(\frac{\tau}{\sqrt{2}}\right) \cdot (\lambda \cdot L - \bar{L}).
\end{align*}
\]

Hence, the gradient of \( F \) takes the following form:

\[
\nabla F(\tau, \lambda) = \begin{bmatrix}
2\lambda^2 \cdot \tau & 2\lambda \cdot \sqrt{\frac{2}{\pi}} e^{-\tau^2/2} \\
2\lambda \cdot \tau^2 & 2 \text{erf}\left(\frac{\tau}{\sqrt{2}}\right)
\end{bmatrix} \cdot \begin{bmatrix}
S - L \cdot h(\tau) \\
\lambda \cdot L - \bar{L}
\end{bmatrix}.
\]

Since \( S, L, \bar{L} > 0 \) and

\[
\det(M) = 4\lambda^2 \cdot \tau \cdot \left( \text{erf}\left(\frac{\tau}{\sqrt{2}}\right) - \tau \cdot \sqrt{\frac{2}{\pi}} e^{-\tau^2/2} \right) > 0
\]

for all \( \tau, \lambda > 0 \), the only critical point of \( F \) on \( (0, \infty) \times (0, \infty) \) is attained if

\[
0 = S - L \cdot h(\tau) = \lambda \cdot L - \bar{L} \quad \Leftrightarrow \quad (\tau, \lambda) = (\tau', \lambda') = \left(h^{-1}\left(\frac{S}{\bar{L}}\right), \frac{\bar{L}}{\bar{S}}\right).
\]

Computing the value of \( F \) at this point, we observe that (cf. (6.17))

\[
F(\tau', \lambda') = n - \frac{(L)^2}{\bar{L}} \cdot \text{erf}\left(\frac{\tau}{\sqrt{2}}\right) < n = F(\tau, 0) = F(0, \lambda)
\]

for all \( \lambda, \tau \geq 0 \), implying that \( F \) cannot attain a global minimum on the boundary of \( [0, \infty) \times [0, \infty) \). Consequently, it suffices to seek for a global minimum on \( (0, \infty) \times (0, \infty) \). For this purpose, let us first consider the quadratic function \( \lambda \mapsto F_\tau(\lambda) := F(\tau, \lambda) \) for \( \lambda > 0 \) and a fixed \( \tau > 0 \). It is not hard to see that \( F_\tau \) has its unique minimum at

\[
\lambda_0(\tau) := \frac{\bar{L} \cdot \text{erf}\left(\frac{\tau}{\sqrt{2}}\right)}{L \cdot \text{erf}\left(\frac{\tau}{\sqrt{2}}\right) + \tau^2 \cdot (S - L \cdot h(\tau))}.
\]

Note that \( \lambda_0(\tau) > 0 \), since \( L \cdot (\text{erf}(\tau/\sqrt{2}) - \tau^2 \cdot h(\tau)) > 0 \) for all \( \tau > 0 \). Next, we minimize the
parameterized mapping
\[
\tau \mapsto \tilde{F}(\tau) := F(\tau, \lambda_0(\tau)) = n - \frac{L^2 \cdot \text{erf}(\frac{\tau}{\sqrt{2}})^2}{L \cdot \text{erf}(\frac{\tau}{\sqrt{2}}) + \tau^2 \cdot (S - L \cdot h(\tau))}
\]
on \((0, \infty)\). An elementary calculation shows that
\[
0 = \tilde{F}'(\tau) \iff 0 = (S - L \cdot h(\tau)) \cdot \left( \tau \cdot \sqrt{\frac{2}{\pi}} e^{-\tau^2/2} - \text{erf}(\frac{\tau}{\sqrt{2}}) \right)_{<0}
\]
and \(\tilde{F}''(\tau') > 0\). Thus, \(\tilde{F}\) attains its unique minimum at \(\tau'\).

Finally, let \((\tau, \lambda) \in (0, \infty) \times (0, \infty)\) be arbitrary. By the above observations,
\[
F(\tau, \lambda) \geq F(\tau, \lambda_0(\tau)) \geq F(\tau', \lambda_0(\tau')) = F(\tau', \lambda')
\]
where equality holds if, and only if, \(\tau = \tau'\) and \(\lambda = \lambda_0(\tau') = \lambda'\). Therefore, \(F\) has indeed its unique global minimum at \((\tau', \lambda')\).

\section{Details on Implementation}

\subsection{Greedy Approach to Constructing Sparse Representations}

In this section, we present a greedy strategy for finding a good approximation space \(U \subset \mathbb{R}^n\) according to the argumentation of Subsection 2.4. Recall that our basic geometric idea is to select \(U \subset \mathbb{R}^n\) in such a way that the orthogonal projection of \(P_U x^*\) enjoys a small value of \(\mathcal{M}(\Psi, P_U x^*)\) while not being too distant from \(x^*\) at the same time. This is precisely the purpose of Algorithm B.1 below, where we fix a target analysis sparsity \(S\) in advance and then identify a subspace \(U\) with \(\|\Psi P_U x^*\|_0 \leq S\) and small approximation error \(\|x^* - P_U x^*\|_2\).

\begin{algorithm}[H]
\caption{(Greedy subspace selection)}
\begin{algorithmic}
\State **Input:** Analysis operator \(\Psi \in \mathbb{R}^{N \times n}\), signal vector \(x^* \in \mathbb{R}^n\), target level of sparsity \(S \in [N]\), (machine) precision parameter \(\epsilon \geq 0\).
\State **Output:** Linear approximation space \(U \subset \mathbb{R}^n\).
\State **Initialize:** \(S_0 := N, S_0 := [N], U := \{0\}\).
\State **While** \(S_0 > S:\)
\State \hspace{1em} (1) Select an index \(k \in S_0\) such that \(\|P_{U_k} x^*\|_2\) becomes minimal, where \(U_k := \text{span}\{\psi_k\}\).
\State \hspace{1em} (2) Determine all indices \(k' \in S_0\) with \(|\langle \psi_{k'}, P_{U_k} x^* \rangle| \leq \epsilon\) and exclude them from \(S_0\).
\State \hspace{1em} (3) Set \(U := \text{span}(\{\psi_k \mid k \in S_0^c\})^\perp\).
\State \hspace{1em} (4) Update \(S_0 := \|\Psi P_U x^*\|_0\).
\State **Return** \(U\).
\end{algorithmic}
\end{algorithm}
Since \( \|x^* - P_{U_k} x^*\|_2 = \|P_{U_k} x^*\|_2 \), the actual intention of Step (1) in Algorithm B.1 is to identify an atom \( \psi_k \in \mathbb{R}^n \) of \( \Psi \) that yields a best approximation of \( x^* \) while enforcing \( \langle \psi_k, P_{U_k} x^* \rangle = 0 \). The coherence structure of \( \Psi \) is then incorporated by Step (2), where all (almost) vanishing analysis coefficients of \( P_{U_k} x^* \) are excluded from the index set \( S_0 \). Let us emphasize that this procedure could be suboptimal in the sense of best S-analysis-sparse approximations (cf. (2.8)), but it still may allow for accurate predictions by means of Theorem 2.8, as demonstrated in Subsection 3.1.4. Finally, let us mention that the statement of Theorem 2.8 actually considers a scaled version of \( P_{U} x^* \) in (2.9), which is merely an artifact of its proof and therefore ignored here.

### B.2 Numerical Estimation of the Sample Complexity

To validate the quality of our sampling-rate predictions via \( M(\Psi, x^*) \), we have also reported the true sample complexity \( \delta(\Psi, x^*) \) of (BP\(_{\Psi=0}\)) in Subsection 3.1 for several choices of \( \Psi \) and \( x^* \). Technically, this actually relies on computing the so-called statistical dimension, which is known to capture the sample complexity of many convex programs [ALMT14].

Let us sketch how this quantity can be numerically calculated. Recalling that \( \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \) denotes the descent set of the \( \ell^1 \)-analysis objective \( \|\Psi(\cdot)\|_1 \) at \( x^* \) (see Definition 6.3), the statistical dimension of the associated descent cone \( \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \) is given by

\[
\delta(\Psi, x^*) := \mathbb{E}\left[ \left( \sup_{h \in \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \cap B_2^n} \langle g, h \rangle \right)^2 \right],
\]

where \( g \sim \mathcal{N}(0, I_n) \); see [ALMT14, Prop. 3.1(5)]. Since \( \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \) is a polytope containing the origin, it is not hard to see that there exists \( t_0 > 0 \), such that

\[
t^{-1} \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \cap B_2^n = \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \cap B_2^n
\]

for all \( 0 < t \leq t_0 \), and in particular,

\[
\delta(\Psi, x^*) = \mathbb{E}\left[ \left( \sup_{h \in t^{-1} \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \cap B_2^n} \langle g, h \rangle \right)^2 \right]. \tag{B.1}
\]

The inner expression of this expected value can be rewritten as a simple convex program:

\[
\sup_{h \in \mathbb{R}^n} \langle g, h \rangle \quad \text{subject to} \quad \|\Psi(x^* + th)\|_1 \leq \|\Psi x^*\|_1 \text{ and } \|h\|_2 \leq 1, \tag{B.2}
\]

where we have used that \( h \in t^{-1} \mathcal{D}(\|\Psi(\cdot)\|_1, x^*) \) is equivalent to \( \|\Psi(x^* + th)\|_1 \leq \|\Psi x^*\|_1 \). Thus, to approximate the expectation in (B.1), we first select a sufficiently small value of \( t \) (typically \( t = 0.01 \) works well) and then solve program (B.2) for independent samples \( g_1, \ldots, g_k \sim \mathcal{N}(0, I_n) \), using the Mat\( \text{lab} \) software package cvx. Due to the concentration behavior of empirical Gaussian processes, the arithmetic mean of \( k = 200 \) samples already yields a good estimate of \( \delta(\Psi, x^*) \).

Note that the notion of statistical dimension is essentially equivalent to the conic Gaussian mean width considered in Section 6. More precisely, we have (see [ALMT14, Prop. 10.2])

\[
u_2^w(\mathcal{D}(\|\Psi(\cdot)\|_1, x^*)) \leq \delta(\Psi, x^*) \leq \nu_2^w(\mathcal{D}(\|\Psi(\cdot)\|_1, x^*)) + 1.
\]

For numerical evaluations, the statistical dimension however seems to be more appealing due to the convexity of (B.2).
C An Excursion to Analysis Operator Learning

According to Theorem 2.5, the value of \( M(\Psi, x^*) \) determines the number of measurements required to ensure recovery. The number of measurements, in turn, can be regarded as the “costs” of successfully applying (BP)\( _{\eta=0} \). In that sense, the mapping

\[
M: \mathbb{R}^{N \times n} \times \mathbb{R}^n \to \mathbb{R}, \ (\Psi, x^*) \mapsto M(\Psi, x^*)
\]

actually plays the role of a loss function (or cost function). This perspective enables us to formalize the procedure of Subsection 5.1 by means of a (machine) learning problem:

**Problem C.1** (Operator Learning – Expected Loss Minimization) Let \( \mathcal{X} \subset \mathbb{R}^n \) be a collection of feasible signal vectors and assume that \( \mu \) is a probability measure on \( \mathcal{X} \). Then, solve the minimization problem

\[
\min_{\Psi \in \mathcal{H}} \int_{\mathcal{X}} M(\Psi, x^*) d\mu(x^*), \tag{C.1}
\]

where \( \mathcal{H} \subset \mathbb{R}^{N \times n} \) is called hypothesis set (or parameter set), containing all potential choices of analysis operators.

Intuitively, the purpose of (C.1) is to select an operator \( \Psi \in \mathcal{H} \) which minimizes the average costs of reconstructing signals from \( \mathcal{X} \). We would like to emphasize that the significance of this approach clearly depends on both the considered signal class \( \mathcal{X} \) and the hypothesis set \( \mathcal{H} \). For instance, if \( \mathcal{X} \) is just a singleton, i.e., \( \mathcal{X} = \{ x^* \} \), solving (C.1) can lead to vacuous minimizers, being strongly adapted to \( x^* \). Of more interest are larger classes \( \mathcal{X} \) whose elements share certain (geometric) characteristics. In fact, we have already studied a prototypical example in Section 3, namely discrete piecewise constant functions; more formally, \( \mathcal{X} \) might consist of all those vectors \( x^* \in \mathbb{R}^n \) with a fixed number of discontinuities (that is, \( \| \Psi_{TV,1} x^* \|_0 \leq S_{TV} \) for some \( S_{TV} > 0 \)).

The choice of \( \mathcal{H} \) is typically driven by the belief in what properties of \( \Psi \) are appropriate to capture the structural features of \( \mathcal{X} \). Once again, Subsection 3.1.2 provides an interesting example: In the scenario of Haar wavelet systems, \( \mathcal{H} \) could contain all those \( \Psi \in \mathbb{R}^{N \times n} \) arising from a scale-wise weighting of \( \Psi_{Hdwt} \). Then, a minimizer of (C.1) would yield an optimally weighted version of \( \Psi_{Hdwt} \) (with respect to the class of piecewise constant signals). It is also worth mentioning that the dimension of the parameter space \( \mathbb{R}^{N \times n} \) is considerably reduced in that way, since the degrees of freedom of \( \mathcal{H} \) would just equal the number of scales.

The optimization task of (C.1) is however often unfeasible in practice because the underlying probability measure \( \mu \) is usually unknown, e.g., when the \( x^* \)s are supposed to be natural images. Instead, one has merely access to a collection of independent samples drawn from a random vector in \( \mathbb{R}^n \) distributed according to \( \mu \). This leads to an empirical version of Problem C.1:

**Problem C.2** (Operator Learning – Empirical Loss Minimization) Let \( \mathcal{X} \subset \mathbb{R}^n \) be the set of feasible signals. Suppose that \( x_1^*, \ldots, x_L^* \in \mathcal{X} \) are independent samples drawn according to a probability measure \( \mu \) on \( \mathcal{X} \). Then, solve the minimization problem

\[
\min_{\Psi \in \mathcal{H}} \frac{1}{L} \sum_{l=1}^{L} M(\Psi, x_l^*), \tag{C.2}
\]

where \( \mathcal{H} \subset \mathbb{R}^{N \times n} \) is again a hypothesis set.

The idea behind (C.2) is quite simple: By computing the empirical mean instead of an integral, one may end up with a tractable problem, whose solutions are still very close to minimizers of (C.1). In fact, solving and analyzing programs of the type (C.2) is one of the key objectives in statistical learning theory [MRT12; Vap00]. This field of research offers a rich toolbox that allows us to investigate many related issues, such as regularization, non-convex optimization, iterative algorithms, or sample complexity. We hope that this methodology, combined with our refined sampling-rate bounds, could be
a starting point of novel design rules and guidelines in analysis-based modeling. On the other hand, due to the non-convexity and discontinuity of $\mathcal{M}(\Psi, x^*)$, the specific program of (C.2) is expected to be (algorithmically) challenging in general. A detailed study would go beyond the scope of this paper and is therefore deferred to future research. Let us finally point out that similar learning strategies were recently proposed in the literature as well, but using very different kinds of loss functions, e.g., see [HKD13; RB13; SWGK16; YNGD12].

Acknowledgments

The authors thank Claire Boyer, Peter Jung, Jackie Ma, and Pierre Weiss for fruitful discussions. Our special thanks goes to Ali Hashemi, who referred us to Price’s theorem, and Felix Voigtlaender, who assisted us by working out a fairly general version of it [Voi17], which is also used in this work. M.G. is supported by the European Commission Project DEDALE (contract no. 665044) within the H2020 Framework Program as well as the Einstein Center for Mathematics Berlin. G.K acknowledges partial support by the Einstein Foundation Berlin, the DFG Collaborative Research Center TRR 109 “Discretization in Geometry and Dynamics,” and the European Commission Project DEDALE (contract no. 665044) within the H2020 Framework Program, DFG Grant KU 1446/18 as well as DFG-SPP 1798 Grants KU 1446/21 and KU 1446/23. M.M. is supported by the DFG Priority Programme DFG-SPP 1798.

References

[ALMT14] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. “Living on the edge: phase transitions in convex programs with random data”. Inf. Inference 3.3 (2014), 224–294.

[B+17] R. Baraniuk, H. Choi, R. Neelamani, et al. Rice Wavelet Toolbox, Version 3. URL: https://github.com/ricedsp/rwt. Aug. 2017.

[BD09] T. Blumensath and M. E. Davies. “Sampling Theorems for Signals From the Union of Finite-Dimensional Linear Subspaces”. IEEE Trans. Inf. Theory 55.4 (2009), 1872–1882.

[BF13] E. van den Berg and M. P. Friedlander. Spot – A Linear-Operator Toolbox. URL: http://www.cs.ubc.ca/labs/sci/sot/index.html. Aug. 2013.

[BPCPE11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. Found. Trends Mach. Learn. 3.1 (2011), 1–122.

[CD95] R. R. Coifman and D. L. Donoho. “Translation-Invariant De-Noising”. Wavelets and Statistics. Ed. by A. Antoniadis and G. Oppenheim. Vol. 103. Lecture Notes in Statistics. Springer New York, 1995, 125–150.

[CDS98] S. S. Chen, D. L. Donoho, and M. A. Saunders. “Atomic Decomposition by Basis Pursuit”. SIAM J. Sci. Comput. 20.1 (1998), 33–61.

[CENR11] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. “Compressed sensing with coherent and redundant dictionaries”. Appl. Comput. Harmon. Anal. 31.1 (2011), 59–73.

[Cha04] A. Chambolle. “An Algorithm for Total Variation Minimization and Applications”. J. Math. Imaging Vision 20.1 (2004), 89–97.

[CK13] P. G. Casazza and G. Kutyniok, eds. Finite Frames: Theory and Applications. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[COS10] J.-F. Cai, S. Osher, and Z. Shen. “Split Bregman Methods and Frame Based Image Restoration”. Multiscale Model. Simul. 8.2 (2010), 337–359.

[CRT06] E. J. Candès, J. K. Romberg, and T. Tao. “ Stable signal recovery from incomplete and inaccurate measurements”. Found. Comput. Math. 6.5 (2006), 561–586.

[CRPW12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. “The convex geometry of linear inverse problems”. Found. Comput. Math. 12.6 (2012), 805–849.

[CRT05] E. J. Candès and T. Tao. “ Decoding by linear programming”. IEEE Trans. Inf. Theory 51.4 (2005), 4203–4215.

[CX15] J.-F. Cai and W. Xu. “Guarantees of total variation minimization for signal recovery”. Inf. Inference 4.4 (2015), 328–353.

[DJ94] D. L. Donoho and J. M. Johnstone. “Ideal spatial adaptation by wavelet shrinkage”. Biometrika 81.3 (1994), 425–455.

[Don06] D. L. Donoho. “Compressed sensing”. IEEE Trans. Inf. Theory 52.4 (2006), 1289–1306.

[DPSS14] A. Denițiu, S. Petra, C. Schnörr, and C. Schnörr. “Phase Transitions and Cosparse Tomographic Recovery of Compound Solid Bodies from Few Projections”. Fund. Inform. 135.1-2 (2014), 73–102.
[DV03] P. L. Dragotti and M. Vetterli. “Wavelet footprints: theory, algorithms, and applications”. IEEE Trans. Signal Process. 51.5 (2003), 1306–1323.

[EMR07] M. Elad, P. Milanfar, and R. Rubinstein. “Analysis versus synthesis in signal priors”. Inverse Probl. 23.3 (2007), 947–968.

[FG05] M. Fornasier and K. Gröchenig. “Intrinsic Localization of Frames”. Constr. Approx. 22.3 (2005), 395–415.

[FMSY12] M. P. Friedlander, H. Mansour, R. Saab, and O. Yilmaz. “Recovering compressively sampled signals using partial support information”. IEEE Trans. Inf. Theory 58.2 (2012), 1122–1134.

[Fol99] G. B. Folland. Real Analysis: Modern Techniques and Their Applications. 2nd Edition. Wiley, 1999.

[Fou14] S. Foucart. “Stability and robustness of $\ell_1$-minimizations with Weibull matrices and redundant dictionaries”. Linear Algebra Appl. 441 (2014), 4–21.

[Fow05] J. E. Fowler. “The Redundant Discrete Wavelet Transform and Additive Noise”. IEEE Signal Process. Lett. 12.9 (2005), 629–632.

[FR13] S. Foucart and H. Rauhut. A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[GB08] M. Grant and S. Boyd. “Graph implementations for nonsmooth convex programs”. Recent Advances in Learning and Control. Ed. by V. Blondel, S. Boyd, and H. Kimura. Vol. 371. Lecture Notes in Control and Information Sciences. Springer London, 2008, 95–110.

[GB14] M. Grant and S. Boyd. CVX: Matlab Software for Disciplined Convex Programming, version 2.1. URL: http://cvxr.com/cvx. Mar. 2014.

[Gen17] M. Genzel. “High-Dimensional Estimation of Structured Signals From Non-Linear Observations With General Convex Loss Functions”. IEEE Trans. Inf. Theory 63.3 (2017), 1601–1619.

[GLPU12] M. Guerquin-Kern, J. Lejeune, K. P. Pruessmann, and M. Unser. “Realistic analytical phantoms for parallel Magnetic Resonance Imaging”. IEEE Trans. Med. Imaging 31.3 (2012), 626–636.

[GNEG14] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies. “Greedy-like algorithms for the cosparse analysis model”. Linear Algebra Appl. 441 (2014), 22–60.

[Gor88] Y. Gordon. “On Milman’s inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$”. Geometric aspects of functional analysis. Ed. by J. Lindenstrauss and V. D. Milman. Vol. 1317. Lecture Notes in Mathematics. Springer, 1988, 84–106.

[GPV15] R. Giryes, Y. Plan, and R. Vershynin. “On the Effective Measure of Dimension in the Analysis Cosparse Model”. IEEE Trans. Inf. Theory 61.10 (2015), 5745–5753.

[Grö04] K. Gröchenig. “Localization of Frames, Banach Frames, and the Invertibility of the Frame Operator”. J. Fourier Anal. Appl. 10.2 (2004), 105–132.

[HKD13] S. Hawe, M. Kleinsteuber, and K. Diepold. “Analysis Operator Learning and its Application to Image Reconstruction”. IEEE Trans. Image Process. 22.6 (2013), 2138–2150.

[KABG16] S. Kitić, L. Albera, N. Bertin, and R. Gribonval. “Physics-Driven Inverse Problems Made Tractable With Cosparse Regularization”. IEEE Trans. Signal Process. 64.2 (2016), 335–348.

[KKS17] F. Krahmer, C. Kusche, and M. Sandbichler. “Total Variation Minimization in Compressed Sensing”. arXiv preprint: 1608.08852. 2017.

[KLR16] G. Kutyniok, W.-Q. Lim, and R. Reisenhofer. “Shearlab 3D: Faithful digital shearlet transforms based on compactly supported shearlets”. ACM Trans. Math. Softw. 42.1 (2016), Art. 5.

[KNW15] F. Krahmer, D. Needell, and R. Ward. “Compressive Sensing with Redundant Dictionaries and Structured Measurements”. SIAM J. Math. Anal. 47.6 (2015), 4606–4629.

[KR15] M. Kabanava and H. Rauhut. “Analysis $\ell_1$-recovery with Frames and Gaussian Measurements”. Acta Appl. Math. 140.1 (2015), 173–195.

[KRZ15] M. Kabanava, H. Rauhut, and H. Zhang. “Robust analysis $\ell_1$-recovery from Gaussian measurements and total variation minimization”. Eur. J. Appl. Math. 26.6 (2015), 917–929.

[LDSP08] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly. “Compressed Sensing MRI”. IEEE Signal Process. Mag. 25.2 (2008), 72–82.

[LML12] Y. Liu, T. Mi, and S. Li. “Compressed Sensing With General Frames via Optimal-Dual-Based $\ell_1$-Analysis”. IEEE Trans. Inf. Theory 58.7 (2012), 4201–4214.

[LMPV17] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin. “A Simple Tool for Bounding the Deviation of Random Matrices on Geometric Sets”. Geometric Aspects of Functional Analysis. Ed. by B. Klartag and E. Milman. Vol. 2169. Lecture Notes in Mathematics. Springer, 2017, 277–299.

[Mal09] S. Mallat. Cover image A Wavelet Tour of Signal Processing: The Sparse Way. 3rd Edition. Elsevier, 2009.

[McM64] E. McMahon. “An extension of Price’s theorem (corresp.)”. IEEE Trans. Inf. Theory 10.2 (1964), 168–168.
S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. “Reconstruction and subgaussian operators in asymptotic geometric analysis”. Geom. Funct. Anal. 17.4 (2007), 1248–1282.

J. L. Mead and R. A. Renaut. “Least squares problems with inequality constraints as quadratic constraints”. Linear Algebra Appl. 432.8 (2010), 1936–1949.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. Foundations of Machine Learning. MIT press, 2012.

S. Nam, M. E. Davies, M. Elad, and R. Gribonval. “The cosparse analysis model and algorithms”. Appl. Comput. Harmon. Anal. 34.1 (2013), 30–56.

S. Nam and R. Gribonval. “Physics-driven structured cosparse modeling for source localization”. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2012, 5397–5400.

D. Needell and R. Ward. “Stable Image Reconstruction Using Total Variation Minimization”. SIAM J. Imaging. Sci. 6.2 (2013), 1035–1058.

S. Oymak, M. A. Khajehnejad, and B. Hassibi. “Recovery threshold for optimal weight ℓ1 minimization”. 2012 IEEE International Symposium on Information Theory Proceedings (ISIT). 2012, 2032–2036.

G. Plonka and J. Ma. “Curvelet-Wavelet Regularized Split Bregman Iteration for Compressed Sensing”. Int. J. Wavelets Multiresolution Inf. Process. 09.01 (2011), 79–110.

C. Poon. “Structure dependent sampling in compressed sensing: Theoretical guarantees for tight frames”. Appl. Comput. Harmon. Anal. 42.3 (2017), 402–451.

R. Price. “A useful theorem for nonlinear devices having Gaussian inputs”. IRE Trans. Inf. Theory 4.2 (1958), 69–72.

S. Ravishankar and Y. Bresler. “MR Image Reconstruction From Highly Undersampled k-Space Data by Dictionary Learning”. IEEE Trans. Med. Imaging 30.5 (2011), 1028–1041.

S. Ravishankar and Y. Bresler. “Learning Sparsifying Transforms”. IEEE Trans. Signal Process. 61.5 (2013), 1072–1089.

R. T. Rockafellar. Convex Analysis. Princeton University Press, 1970.

L. Rudin, S. Osher, and E. Fatemi. “Nonlinear total variation based noise removal algorithms”. Physica D 60.1–4 (1992), 259–268.

H. Rauhut, K. Schnass, and P. Vanderheynst. “Compressed Sensing and Redundant Dictionaries”. IEEE Trans. Inf. Theory 54.5 (2008), 2210–2219.

I. W. Selesnick and M. A. T. Figueiredo. “Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors”. Proceedings of SPIE, Wavelets XIII. Ed. by V. K. Goyal, M. Papadakis, and D. V. D. Ville. Vol. 7446. 2009.

M. Seibert, J. Wörmann, R. Gribonval, and M. Kleinsteuber. “Learning Co-Sparse Analysis Operators With Separable Structures”. IEEE Signal Process. 64.1 (2016), 120–130.

A. M. Tillmann, R. Gribonval, and M. E. Pfetsch. “Projection onto the cosparse set is NP-hard”. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014, 7148–7152.

J. A. Tropp. “Convex Recovery of a Structured Signal from Independent Random Linear Measurements”. Sampling Theory, a Renaissance. Ed. by G. E. Pfander. Applied and Numerical Harmonic Analysis. Birkhäuser, 2015, 67–101.

V. Vapnik. The Nature of Statistical Learning Theory. Springer New York, 2000.

R. Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. Compressed Sensing Theory and Applications. Ed. by Y. C. Eldar and G. Kutyniok. Cambridge University Press, 2012, 210–268.

F. Voigtlaender. “A general version of Price’s theorem”. arXiv preprint: 1710.03576. 2017.

S. Vaiter, G. Peyre, C. Dossal, and J. Fadili. “Robust Sparse Analysis Regularization”. IEEE Trans. Inf. Theory 59.4 (2013), 2001–2016.

M. Yaghoobi, S. Nam, R. Gribonval, and M. E. Davies. “Noise aware analysis operator learning for approximately cosparse signals”. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2012, 5409–5412.

B. Zhang, W. Xu, J.-F. Cai, and L. Lai. “Precise phase transition of total variation minimization”. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2016, 4518–4522.