Nonlinear Approximation Based Image Recovery Using Adaptive Sparse Reconstructions and Iterated Denoising: Part I - Theory

Onur G. Guleryuz
DoCoMo Communications Laboratories USA, Inc.
181 Metro Drive, Suite 300,
San Jose, CA 95110
guleryuz@docomolabs-usa.com, 408-451-4719, 408-573-1090(fax)

Abstract

We study the robust estimation of missing regions in images and video using adaptive, sparse reconstructions. Our primary application is on missing regions of pixels containing textures, edges, and other image features that are not readily handled by prevalent estimation and recovery algorithms. We assume that we are given a linear transform that is expected to provide sparse decompositions over missing regions such that a portion of the transform coefficients over missing regions are zero or close to zero. We adaptively determine these small magnitude coefficients through thresholding, establish sparsity constraints, and estimate missing regions in images using information surrounding these regions. Unlike prevalent algorithms, our approach does not necessitate any complex preconditioning, segmentation, or edge detection steps, and it can be written as a sequence of denoising operations. We show that the region types we can effectively estimate in a mean squared error sense are those for which the given transform provides a close approximation using sparse nonlinear approximants. We show the nature of the constructed estimators and how these estimators relate to the utilized transform and its sparsity over regions of interest. The developed estimation framework is general, and can readily be applied to nonstationary signals with a suitable choice of linear transforms. Part I discusses fundamental issues, and Part II is devoted to adaptive algorithms with extensive simulation examples that demonstrate the power of the proposed techniques.
EDICS: 2-LFLT Linear Filtering and Enhancement; 2-MODL Modeling; 2-REST Restoration

The author would like to extend his most sincere gratitude to the three anonymous reviewers and to the Associate Editor for their insightful comments which have significantly improved this paper. Part of this work was done when the author was with Epson Palo Alto Laboratory. Part of this work was supported by NSF (CAREER award IIS-0093179).

LIST OF FIGURES

I. INTRODUCTION

Many applications necessitate the estimation/recovery of missing regions of pixels in an image or video frame using the information provided by the remaining pixels. For example, in image and video compression applications over unreliable channels the decoder has to contend with data corrupted by channel errors. With macroblock based coders [25], [28] these errors lead to missing rectangular regions which need to be estimated under a fidelity criterion by appropriate recovery and concealment algorithms. Similarly failures in capture and storage devices or imperfections in other processes in a generic signal/image/video processing pipeline produce errors which necessitate the application of restoration algorithms that predict missing regions. In this sequence of two papers we develop techniques that are geared toward the recovery of missing regions using the minimum mean squared error criterion and an implicit statistical model. While our main applications will be the estimation of missing regions in images and video, it will be clear that the developed techniques can be generalized to handle missing "regions" in other types of signals and they can also be generalized to accommodate other applications. For example, our techniques can be used as part of an encoder that reduces redundancy by predictively encoding signals, images, and video, or deployed in applications that require prediction in more general scenarios. We will therefore abuse terminology and use the terms recovery, prediction, and estimation interchangeably. The reader is referred to [41], [42], [2], [23], [34], and references therein, for some example prior work specifically related to the recovery application presented in this paper.

In the case of images the missing data has to be predicted spatially while for video both temporal and spatial prediction can be attempted. In this work we will primarily be concerned with the recovery of missing data using spatial prediction alone. As such, for video, the presented techniques are directly applicable in cases where temporal prediction is not possible or prudent, for example, in concealment applications involving severely corrupted motion vectors and/or intra marked macroblocks in the popular MPEG algorithms [24], [28]. We will primarily concern ourselves with the recovery of completely missing image blocks of known location though our algorithms can be adapted to situations where partial information is available and/or the missing data corresponds to non-rectangular or irregularly shaped regions. Of particular interest to us is the robust recovery of image blocks that contain textures, edges, and other features that are not readily handled by prevalent algorithms [41], [42]. While visual appearance and uniformity will be important, we will mainly be after significant PSNR improvements in recovered regions. Figure 1 illustrates some examples of recovery over edge and texture regions by the algorithms proposed in this work.

If we view recovery algorithms as providing estimates of missing data based on an assumed statistical model, we can associate available techniques with the statistical models that they implicitly or explicitly utilize. Loosely speaking, for images these models range from the simple "images are composed of smoothly varying pixels" (see for example, [43], [33], [1], [29]), to the more intermediate "images are composed of locally smooth regions separated by edges" (e.g., [2], [22], [37], [40]), and finally to the more general "images are composed of locally uniform regions separated by edges" (e.g., [23], [30], [3]). Of course, the exact definitions of smooth, local, and uniform depend on the particular method, with many variations proposed for specific applications and associated constraints. Typically smoothness is defined using polynomials or in terms of the frequency (or transform coefficient) content

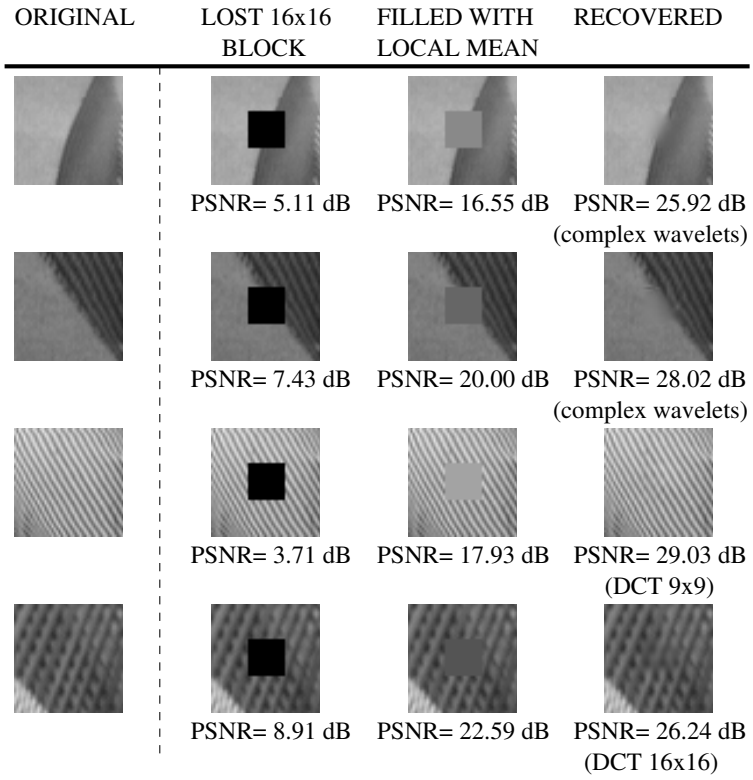


Fig. 1. Some examples of the proposed recovery algorithm over 16×16 missing blocks from various regions of the standard image Barbara.

of the estimates, and uniform regions are allowed to contain either smooth, texture or other structures containing high frequencies. However, as the sophistication of models increases so does the complexity of the associated techniques, and one has to contend with algorithms that are not fully autonomous, techniques that depend on non-robust preconditioning or edge detection steps, and other methods that provide results which are visually acceptable, but which may deviate significantly from the missing data in terms of mean-squared-error. The goal of this work is to address these deficiencies[1].

As we will see, the work presented in this sequence of two papers combines recent advances in the understanding of linear transforms and associated approximation spaces with well-known sparse reconstruction ideas. Classical work in sparse reconstructions considers the case where the signal with the missing information is known apriori to be bandlimited, i.e., a portion of the signal's Fourier transform coefficients are known or assumed to be zero. Then, by employing well-known techniques (see for e.g., [13], [32] and references therein), one can attempt a recovery of the missing information under the mean squared error fidelity metric using the *sparsity constraint* that a portion of the signal's transform coefficients are zero. Designed mainly with stationary Gaussian signals in mind, these approaches assume that the target class of signals have sparse representations in terms of the Fourier transform. As such they encounter serious problems when applied to images and other nonstationary signals since it is well

[1]Our early results were reported in [18], [19]. Some texture generation, resampling, and modeling approaches also utilize linear transforms or formulate their models in transform domain. While not motivated by mean squared error, and typically utilizing very sophisticated nonlinear models (as opposed to the linear transformation and hard-thresholding techniques which we utilize), these techniques may achieve exceptional visual quality in some applications. The reader is referred to [34], [6] and references therein.

known that Fourier transforms are not good at providing sparse linear representations for such signals [10], [4]. Localized basis such as DCTs, wavelets, etc., are much better suited to this task. Another issue with classical work is the need to know *apriori* exactly which subset of the representation coefficients have to be zero. For general images and other nonstationary signals, even with a well-designed localized basis, it is very difficult to know *in advance* which portion of the transform coefficients that yield a sparse representation are zero or close to zero. While "images are composed of locally uniform regions separated by edges" may be a reasonable model, we do not know in advance the distribution of the edges and the types and locations of the locally uniform regions. This second problem can only be avoided by adaptively determining sparsity constraints over the particular image under examination.

For the purposes of this work, a linear orthonormal transformation provides a sparse representation over a class of N dimensional signals if it is such that about Z ($1 << Z < N$) of the transform coefficients $c(i)$ ($i = 1, \ldots, N$) have small magnitudes, i.e., $|c(j)| < T$ for some given T as j ranges in an *index set* that determines the small coefficients[2]. Observe that we are not particular about the index set that determines which of the transform coefficients are small. In fact, one of the key properties of this work is its ability to allow this index set to change for *each* signal in the class in order to derive substantial benefits in adaptivity and robustness through nonlinear approximation principles [11].

Very basically, what this work proposes to do is to adaptively determine the small magnitude transform coefficients of a signal, or equivalently determine the aforementioned *signal specific* index set, and predict the missing region subject to the constraint that these coefficients are small, i.e., subject to adaptively determined sparsity constraints. We will show that all linear estimators have associated sparsity constraints and that sparsity constraints lead to linear estimators. Hence, the adaptive determination of sparsity constraints as presented in this sequence of two papers will lead to adaptive linear estimates of missing regions. We will see that techniques based on apriori sparsity constraints, such as classical work, can at best hope to be optimal under second order ensemble statistics, which may not reflect the underlying nonstationary behavior. In comparison, we will see that our estimates provide robust and effective performance over nonstationary signals by using conditional rather than ensemble statistics. We will further show that when using conditional statistics the distinction between optimal linear estimators and optimal estimators gets blurred, and in fact our techniques have the potential to construct *the* optimal estimates. A natural consequence of our development will be a correspondence among utilized transforms, adaptively determined conditional statistics, and nonlinear approximation classes. As will be shown in this work, well-known good transforms, when coupled with our algorithms, produce simple, robust, and very effective results.

The work is divided into two parts. The first part is composed of Sections I through V, which discuss the basic ideas and theory behind this work. In the remaining section of the Introduction we try to place the contributions of this work within the broad space of estimation methods (Section I-A). In Section II, we provide basic nonlinear approximation ideas that are utilized in this paper, followed by Section III, which includes the main analysis. In order to establish connections with denoising literature and as a prelude to Part II, Section IV is devoted to solutions

[2]One can also accommodate nonorthogonal transforms by insisting that the remaining $K = N - Z$ significant coefficients be adequate in forming a faithful approximation to the original signal under the mean squared error fidelity criterion.

of key design equations using denoising based on hard-thresholding. Section V sketches a simple algorithm for recovery that exposes the nonconvexity inherent in adaptive index set determination. Details of this algorithm must be carefully filled in, and the second part of this work is devoted to its step-by-step construction. We conclude the first part in Section VI of concluding remarks.

The second part of the work is devoted to the derivation and analysis of our adaptive algorithm. In Section Part II - II, we use the introduced concepts to derive a powerful algorithm, and examine its properties. Section Part II - III includes our extensive simulation results, where we also discuss the properties of linear transforms and sparse decompositions that can be used to yield successful prediction with our algorithm. Section Part II - IV discusses future work and concludes the sequence.

A. Contributions of this Work

This work demonstrates a systematic way of constructing adaptive estimators for nonstationary signals. One of its key properties is its delegation of the required adaptive statistical modeling to the sparsity properties of linear transforms and filter banks. This allows one to readily take advantage of the existing literature on transform and filter bank design in rapidly obtaining general and powerful predictors for new problems without explicitly building statistical models. (What we mean by explicit building is the statistical inference of pdfs, parameters, etc., in order to fit data to a carefully defined statistical model which is then used to build predictors.) The work requires no explicit covariance/correlation modeling, averaging, or other explicit statistics inference procedure. Hence, we simply bypass all issues related to statistics inference on nonstationary signals, for e.g., we do not need to segment the signal into statistically uniform regions so that we can infer the correct local statistics, we likewise do not need to have labeled data to train/learn, etc. These issues typically force image processing techniques to applications on piecewise smooth signals since edge information, and hence the segmentation, is easier to determine. The proposed work has no such restrictions.

Using the techniques of this work, statistical modeling comes about in an implicit way through the action of choosing a particular transform for a problem. As we will see however, the modeling accomplished by this choice is general, since a given transform allows the capture of a multitude of possible statistical models under one umbrella. Hence, while one does in a sense commit to a class of stochastic processes by choosing a particular transform, thanks to our use of nonlinear approximation principles, the implied class of signals over which successful estimation is possible is very broad. As we will show, even well-known transforms, such as DCTs, provide sophisticated estimation performance over a large variety of image regions with the application of the techniques in this work.

It is important to note that the algorithm proposed in this work knows virtually nothing about the type of data being operated on, since all it effectively does is iterations of simple denoising, based on hard-thresholding. Armed with a good transform that is expected to provide sparse decompositions, it is straightforward to directly apply the algorithm proposed in Part II [16] to estimation applications on other nonstationary signals. As such, this work should not be judged as just another block recovery algorithm but more as a general estimation paradigm (see for e.g., [17], for an interpolation application).

Being based on sparse nonlinear approximants and the nonconvex data modeling they provide (Section II), the

sequence of two papers recognizes that the target nonstationary data "lives in" nonconvex sets, and realistically builds the estimation algorithm to deal with this issue from the ground up (Sections V and Part II - II-B). This results in a *progression* of estimates from coarse to fine as provided by iterated denoising from larger to smaller thresholds.

More generally, the work provides a very useful categorization of adaptive linear estimators. With the established duality between estimators and transforms (Section III), it is straightforward to see that *any* work doing adaptive linear estimation is effectively choosing an orthonormal transform that the work expects will provide sparse decompositions on the target data. As examined here, this choice can be done in three ways, non data-adaptive transform and index set (linear approximation), non data-adaptive transform but data-adaptive index set (nonlinear approximation), and finally data-adaptive transform and data-adaptive index set (adaptive nonlinear approximation). Our results indicate that sophisticated adaptive performance is possible even with estimators of the type "non data-adaptive transform but data-adaptive index set". Through this categorization it is also possible to draw further ties with results in harmonic analysis (and its classification of function spaces), to more precisely determine the class of stochastic processes over which successful estimation is possible by using a particular technique.

Finally, beyond providing systematic estimators, categorization, and ties with harmonic analysis, the results of this work can also be used to obtain better signal representations by providing another benchmark application for transform and filter bank design.

II. Basic Nonlinear Approximation Ideas

Recent activity in signal processing has resulted in an important shift in the way linear signal representations are designed and exploited. Using results from harmonic analysis [11], it is now recognized that in many signal processing applications it is advantageous to switch from linear approximation with a predetermined basis to a nonlinear one (see for example, [31], [11], [10], [4], and references therein). Given a basis and a signal's representation in terms of this basis, i.e., the coefficients or coordinates of the signal in this basis, linear approximation based techniques insist on viewing this representation in terms of a specific order, namely the order determined by the apriori ordering of the basis functions. Nonlinear approximation based techniques on the other hand have no apriori order preferences, and they have the capability to utilize different orderings depending on the *signal* and application.

Let x $(N \times 1)$ be an N dimensional signal and assume we are given a linear, invertible transform. Let h_i $(N \times 1)$, $i = 1, \ldots, N$, denote the reconstruction basis, and let c_i, $i = 1, \ldots, N$, denote the corresponding transform coefficients of x. We have

$$x = \sum_{i=1}^{N} c_i h_i. \tag{1}$$

The distinction between the two approaches manifests itself when we consider approximations $\hat{x}_{linear}$ and $\hat{x}_{nonlinear}$ of x with a limited number, say $K < N$, of transform coefficients. The two types of approximations can be written

as

$$\hat{x}_{linear}(K) = \sum_{i=1}^{K} c_i h_i, \tag{2}$$

$$\hat{x}_{nonlinear}(K) = \sum_{i \in \mathcal{E}(x)} c_i h_i, \tag{3}$$

where the cardinality of the index set $\mathcal{E}(x)$ in Equation (3) is $card(\mathcal{E}(x)) = K$, and the notation indicates the dependence of the index set on the signal. As can be seen, linear approximation becomes one particular form of nonlinear approximation if we set $\mathcal{E}(x) = \{1, \ldots, K\}$, however nonlinear approximation becomes much more advantageous when we allow for the optimal choice of $\mathcal{E}(x)$ that minimizes the mean squared approximation error for *each* x. Observe that such a signal-adaptive choice has the consequence that a linear combination of two signals, which can be represented by K coefficients each, may require more than K coefficients in the given basis, i.e., when viewed as an operator, the approximation process becomes nonlinear, and hence the name nonlinear approximation.

For orthonormal transforms, it is easy to see that the optimal $\mathcal{E}(x)$ can be constructed as the indices of the K largest magnitude transform coefficients of x. Then, for each type of approximation, using analysis on continuous time signals, one can consider how the mean squared approximation error decays as K increases, and construct classes of continuous time signals by grouping together those signals for which the error decays faster than a prescribed rate. Here we limit ourselves to general terminology and refer the reader to [11] for a systematic treatment with orthonormal as well as biorthogonal transforms. Suffice to say that through such constructions it can be shown that nonlinear approximation yields classes that are significantly richer than linear approximation classes. For important results in this direction that are particularly relevant to signal processing, the reader should also consult [10], which compares linear and nonlinear approximation using various transforms. In particular, this work directly shows the vast superiority of nonlinear approximation using localized transforms over linear approximation using Fourier and Karhunen-Loeve transforms on nonstationary signals having edges.

The approximation notions that will be important for us are concerned with the classes of signals for which high fidelity approximations can be achieved with $K << N$, i.e., $\hat{x}_{\cdot}(K) \stackrel{\sim}{=} x$, and the size of these classes. Let us loosely refer to these classes as the sparse classes of the given transform in order to make some intuitive arguments. For both types of approximation the sparse classes are composed of signals that are sparse in the transform domain since only a few transform coefficients are required in a high fidelity reconstruction. In the case of linear approximation (Equation (2)), the class of sparse signals is limited by the predetermined ordering. For example, if the ordering of the transform basis is such that low frequency basis functions have lower indices, then the sparse class is expected to contain mostly low-pass or smooth signals. Similarly, by using different apriori orderings one can have *individual* sparse classes of band-pass or high-pass signals. In comparison, with nonlinear approximation (Equation (3)), we do not have to commit to an apriori fixed order and we can construct a *single* class of sparse signals to contain not only *all* of the aforementioned classes but also significantly richer combinations. In linear approximation the size of the sparse class is essentially determined by assigning different values to the K retained coefficients. On the other hand in nonlinear approximation, there is a further combinatorial factor of $\begin{pmatrix} N \\ K \end{pmatrix}$ which determines which K coefficients are retained. The simple nonlinear approximation extension of allowing the index set to vary for each

signal in the class yields a much bigger sparse class than is possible with linear approximation. As will become clear, for a given transform, the mean squared error effectiveness of our reconstructions will be directly tied to the respective class of sparse signals and it will be very important for us to have as large a class as possible. We will accomplish this by using nonlinear approximation, which can take advantage of sparseness wherever it may exist.

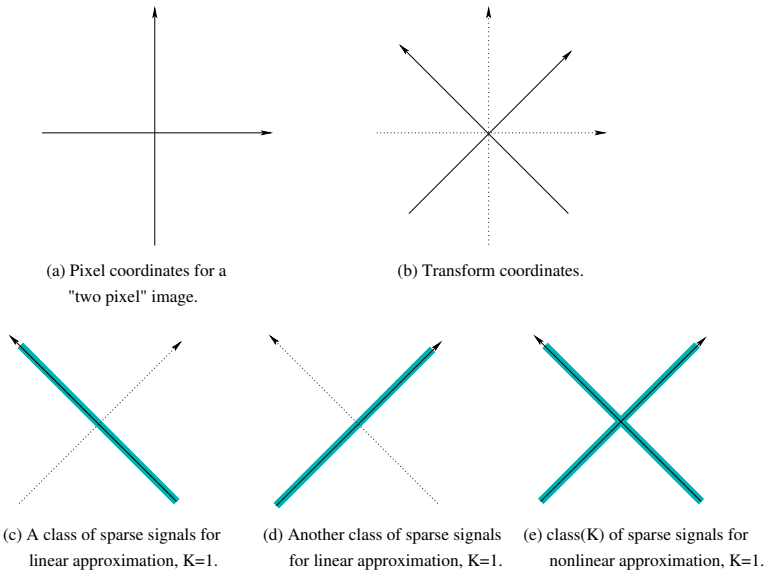


Fig. 2. Sparse classes for linear and nonlinear approximation on a "two pixel" image. Linear approximation classes are convex. Nonlinear approximation classes are more general, star-shaped sets [35], [20].

It is interesting to note that the class of sparse signals under linear approximation form convex sets whereas sparse signals under nonlinear approximation make up non-convex sets (a convex combination of two signals, which can be represented by K coefficients each, may require more than K coefficients in the given basis). Figure 2 (a) and (b) illustrate pixel and transform coordinates for a "two pixel" image. As shown in Figure 2 (c) and (d), we can have two distinct sparse classes using linear approximation with $K = 1$. The single sparse class using nonlinear approximation is shown in Figure 2 (e). Note that the set shown in Figure 2 (e) is non-convex and contains the sets in Figure 2 (c) and (d) as subsets. As can be seen, the sparse classes for nonlinear approximation are more general star-shaped sets. (A set $\mathcal{C} \subset \mathcal{R}^n$ is said to be *star–shaped*, if for any $x \in \mathcal{C}$, the line segment joining the origin to x lies in $\mathcal{C}$.). Star-shaped sets, while substantially different, enjoy some similar basic properties with convex sets (see for e.g., [35], for general properties, and [20], for entropy results for probability distributions defined on star-shaped sets). Such sets and nonlinear approximation form good models for most natural images since the convex combination of two images (say Lena and Barbara) has different properties and can typically be separated into its constituents (Figure 3). While it will force us to deal with non-convex optimization, we believe this non-convex modeling of signals through nonlinear approximation is of fundamental importance.

The sets shown in Figure 2 for two dimensions and extensions in higher dimensions are ideal, since the elements of these sets have some transform coefficients that are precisely equal to zero. The intuitive picture provided in Figure 4, which shows "extensions" of sparse classes as functions of a threshold $T > 0$, will be more useful when we discuss adaptive algorithms and nonconvex optimization (Sections V and Part II - II-B). The elements of these

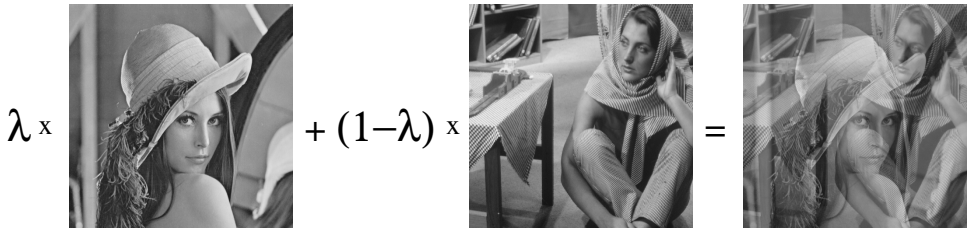
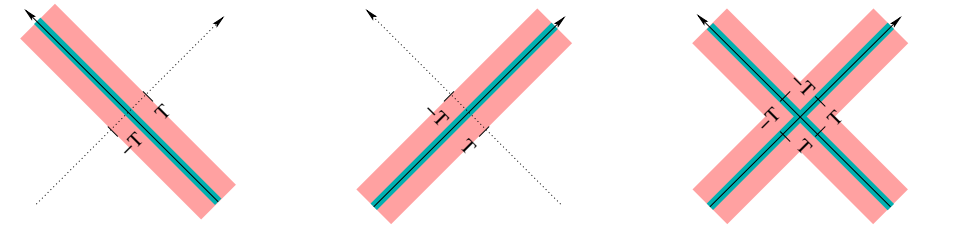


Fig. 3. Natural images do not lie in convex sets. A convex combination of two images has distinctly different properties and can typically be separated into its constituents (see for e.g., [27] and references therein).

sets have some transform coefficients with magnitudes less than T, and we will say that the elements of these sets have some small transform coefficients. For future reference we make the following definition.

Definition 1 (Sparse Class Extension): Given $T > 0$, the $class(K,T)$ of signals with respect to a basis h_i, $i = 1, \ldots, N$, is the set of all signals x $(N \times 1)$ such that,

$$class(K,T) = \{x | x = \sum_{i \in \mathcal{E}(x)} c_i h_i + \sum_{j \notin \mathcal{E}(x)} c_j h_j,$$
$$card(\mathcal{E}(x)) = K, \ |c_j| < T \text{ if } j \notin \mathcal{E}(x)\}. \tag{4}$$



(a) A class of sparse signals for linear approximation, K=1.

(b) Another class of sparse signals for linear approximation, K=1.

(c) class(K,T) of sparse signals for nonlinear approximation, K=1.

Fig. 4. Extensions of sparse classes for linear and nonlinear approximation on a "two pixel" image using a threshold $T > 0$. Linear approximation classes are convex. Nonlinear approximation classes are more general, star-shaped sets [35], [20].

We will see later that an intuitive picture for the algorithms presented in this work will be the estimation of x via the chain

$$\textit{Given } T > 0 \textit{ and the observed signal} \rightarrow class(K,T) \rightarrow$$
$$\rightarrow \hat{\mathcal{E}}(x) \rightarrow \textit{estimated signal} \cong \hat{x}_{nonlinear}(K), \tag{5}$$

where $\hat{\mathcal{E}}(x)$ denotes an estimate of the index set $\mathcal{E}(x)$. One of the main tools we will develop in this work will be this adaptive determination of $\mathcal{E}(x)$, which will replace the unnecessary presumption that linear approximation makes with conditional statistics. Figure 5 shows a sequence of results that illustrate the use of a 16×16 DCT in obtaining successful estimates over different region types. The algorithms we will present accomplish these results by effectively searching for estimates over nonlinear approximation classes.

In this work we will use the index set of small or *insignificant* coefficients given by

$$V(x) = \{1, \ldots, N\} - \mathcal{E}(x), \tag{6}$$

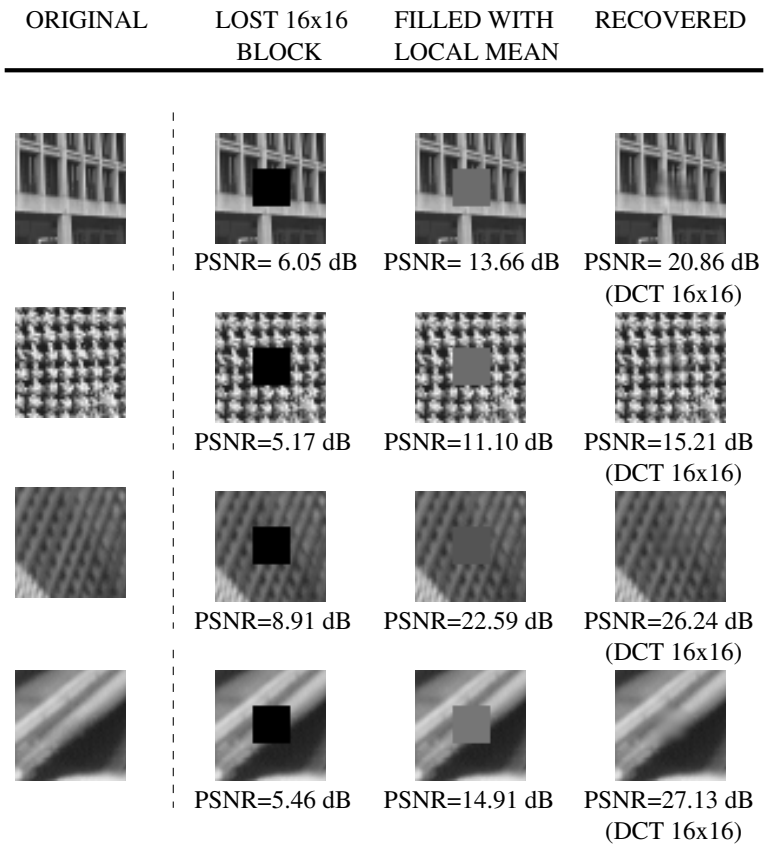


Fig. 5. Some recovery results using 16×16 DCTs. The algorithms of this work can recover different types of regions by using a fixed transform, but by adaptively changing the index set of insignificant coefficients.

in place of the index set of significant coefficients $\mathcal{E}(x)$. We will not be specifically concerned with K or $Z = N-K$, but instead we will use thresholds to determine the insignificant coefficients. Hence it will be important to note that K or Z are not design parameters or constraints in our setting. The set of small or insignificant coefficients $V(x)$ will really become dependent on the utilized threshold T, i.e., $V(x) \rightarrow V(x,T)$, as we will be utilizing different thresholds in our estimation framework. In this sense there is a very direct connection between our work and thresholding based denoising techniques such as [8], [5], which also adaptively determine the insignificant set $V(x,T)$ and estimate the denoised signal accordingly. Similar to denoising, the determination of $V(x,T)$ will not be exact as we too will be trying to determine this set from noisy data. However, by using layering algorithms and localized transforms progressively in Part II, we will limit possible discrepancies under basic assumptions.

III. Sparse Reconstructions Using a Linear Transform $\mathbf{G}$

In this section we formulate the main estimation constructs that are utilized in this work. We do so using a single orthonormal transform $\mathbf{G}$ as it simplifies notation and enables one to see our very basic construction in terms of familiar language and ideas. Much of the properties and performance of this work is due to the way the simple ideas introduced in this section are generalized to develop a full-fledged algorithm in Part II of this sequence. While Part II builds up from the base we establish here, the reader is cautioned that it incorporates significantly more sophisticated ideas (iterated denoising, nonconvex optimization, progressive estimates, overcomplete transforms,

etc.) on to the underlying theory. In Section III-A we show the equivalence between sparsity constraints and linear estimators. We derive two simple results, namely that sparsity constraints result in linear estimators of missing data (Proposition 3.1), and conversely, linear estimators of missing data determine sparsity constraints (Proposition 3.2). Since sparsity constraints are established through the utilized transform there exists a correspondence between transforms (together with the index set of insignificant transform coefficients) and estimators. We illustrate this correspondence in Section III-B and derive optimality results for ensemble and conditional statistics. We show that the optimal transforms introducing the sparsity constraints are tied to optimal estimators and vice versa, where optimality is in the mean squared error sense and means are calculated over ensemble or conditional statistics (Propositions 3.3 and 3.4). In Section III-B.3 we look at the estimates constructed in this work in light of the results in Section III-B. We investigate the class of estimators for a given linear transform and tie these to nonlinear approximation classes.

A. Sparsity Constraints and Linear Estimators

Suppose that the *original* image is arranged into a vector x $(N \times 1)$, such that

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}, \tag{7}$$

where x_0 $(n_0 \times 1)$ constitutes the available pixels and x_1 $(n_1 \times 1)$ denotes the pixels in the missing region. We have $n_0 + n_1 = N$. Given the image containing the missing region, we would like to form an estimate of the original image x by

$$y = \begin{bmatrix} x_0 \\ \hat{x}_1 \end{bmatrix}, \tag{8}$$

where $\hat{x}_1$ is our estimate of the missing region x_1. Assume zero mean quantities.

Without loss of generality let $\mathbf{G}$ $(N \times N)$ denote an *orthonormal* transformation acting on y to yield transform coefficients c $(N \times 1)$ via

$$c = \mathbf{G}y. \tag{9}$$

For now we leave issues regarding the determination of sparsity constraints to Part II and assume that we are given the indices of the significant and insignificant coefficients. Arrange and partition the rows of $\mathbf{G}$ into $\mathbf{G}_I$ $(Z \times N)$ and $\mathbf{G}_S$ $((N - Z) \times N)$ to indicate the portions of the transform that are known to produce insignificant and significant transform coefficients respectively, i.e., let

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_I \\ \mathbf{G}_S \end{bmatrix}. \tag{10}$$

We start by recovering x_1 subject to the sparsity constraint

$$\mathbf{G}_I y = 0, \tag{11}$$

i.e., the insignificant transform coefficients are zero. Partition the columns of $\mathbf{G}_I$ into $\mathbf{G}_{I,0}$ $(Z \times n_0)$ and $\mathbf{G}_{I,1}$ $(Z \times n_1)$ to indicate portions that overlap x_0 and $\hat{x}_1$ such that

$$\mathbf{G}_I = \left[\begin{array}{c|c} \mathbf{G}_{I,0} & \mathbf{G}_{I,1} \end{array} \right], \tag{12}$$

and our constraint becomes

$$\mathbf{G}_{I,0} x_0 + \mathbf{G}_{I,1} \hat{x}_1 = 0. \tag{13}$$

In order to avoid issues related to equation ranks let us reformulate this constraint by considering the *equivalent* problem where we obtain $\hat{x}_1$ that minimizes $||\mathbf{G}_I y||^2$. This results in

$$\mathbf{G}_{I,1}^T \mathbf{G}_{I,0} x_0 + \mathbf{G}_{I,1}^T \mathbf{G}_{I,1} \hat{x}_1 = 0, \tag{14}$$

where $(\ldots)^T$ denotes transpose. Depending on the rank of $\mathbf{G}_{I,1}$ it is clear that Equation (14) can be solved either exactly to recover $\hat{x}_1$ or it can be solved within the positive eigenspace of $\mathbf{G}_{I,1}^T \mathbf{G}_{I,1}$ to recover the portion of $\hat{x}_1$ lying in this subspace. In the latter case we assume that the component of $\hat{x}_1$ orthogonal to the alluded to subspace is set to zero. (The reader is cautioned that in our formulation with progressive thresholds to be discussed later in Part II, we will not set this component to zero and we will actually allow for a non-zero mean value to propagate as determined by prior solutions.). We thus have the following result.

Proposition 3.1: The constraint given by Equation (14) results in a linear estimate of x_1 in terms of x_0, i.e.,

$$\hat{x}_1 = \mathbf{A} x_0, \tag{15}$$

where $\mathbf{A}$ $(n_1 \times n_0)$ is a matrix determined by $\mathbf{G}_I$ using Equation (14).

Proof: Using Equation (14), if $\mathbf{G}_{I,1}^T \mathbf{G}_{I,1}$ is invertible,

$$\mathbf{A} = -(\mathbf{G}_{I,1}^T \mathbf{G}_{I,1})^{-1} \mathbf{G}_{I,1}^T \mathbf{G}_{I,0}. \tag{16}$$

Otherwise since $\mathbf{G}_{I,1}^T \mathbf{G}_{I,1}$ is symmetric and positive semidefinite, let

$$\mathbf{G}_{I,1}^T \mathbf{G}_{I,1} = \sum_{\{j|\lambda_j > 0\}} \lambda_j v_j v_j^T, \tag{17}$$

be its eigen decomposition, where $\lambda_j > 0$ are the non-zero eigenvalues and v_j are the corresponding eigenvectors. Then

$$\mathbf{A} = -(\sum_{\{j|\lambda_j > 0\}} \frac{1}{\lambda_j} v_j v_j^T) \mathbf{G}_{I,1}^T \mathbf{G}_{I,0}. \tag{18}$$

□

On the other hand, suppose we form a linear estimate of x_1 using a matrix $\mathbf{A}$ $(n_1 \times n_0)$ via $\hat{x}_1 = \mathbf{A} x_0$. We have

$$y = \left[\begin{array}{c} x_0 \\ \hat{x}_1 \end{array} \right] = \left[\begin{array}{c} \mathbf{1} \\ \mathbf{A} \end{array} \right] x_0, \tag{19}$$

where $\mathbf{1}$ is the n_0 dimensional identity. Thus y is constrained to lie in a n_0 dimensional subspace determined by the columns of $\left[\begin{array}{c} \mathbf{1} \\ \mathbf{A} \end{array} \right]$ and we have the following result.

Proposition 3.2: Any linear estimate of x_1 given by Equation (19) results in estimates constrained to an n_0 dimensional linear subspace which yields a sparsity constraint of the form

$$\mathbf{G}_I y = 0, \tag{20}$$

where $\mathbf{G}_I$ $(n_1 \times N)$ is a matrix of orthonormal rows determined by $\mathbf{A}$, up to an n_1 dimensional rotation. Each row of $\mathbf{G}_I$ is orthogonal to the constraining subspace, i.e.,

$$\mathbf{G}_I \begin{bmatrix} \mathbf{1} \\ \mathbf{A} \end{bmatrix} = \mathbf{0}.$$

$\mathbf{G}_I$ can be augmented to a complete orthonormal transformation $\mathbf{G}$ $(N \times N)$ via the selection of a $\mathbf{G}_S$ of orthonormal rows (up to an n_0 dimensional rotation).

Proof: A variety of techniques, including Gramm-Schmidt orthogonalization procedures, can be utilized to put Proposition 3.2 to effect. We refer the reader to a compendium such as [14] for details. □

Having established the simple connection among transforms, sparsity constraints, and estimators, we turn our attention to optimality under the mean squared error metric.

B. Optimal Sparsity Constraints and Optimal Estimators

In any estimation procedure on stochastic data one would like to form the optimal estimates under a given fidelity criterion. In this work we are interested in the best estimates in the mean squared error sense. With the aid of Propositions 3.1 and 3.2 we can immediately see that the optimal sparsity constraint ($\mathbf{G}_I y = 0$) is established (up to rotations) by the optimal linear estimator ($\mathbf{A}$) and vice versa, where optimality is in the mean squared error sense. We next look at the form of these optimal estimates by invoking well known estimation theory results [39].

It is important to note that given a linear estimation matrix $\mathbf{A}$ we can obtain $\mathbf{G}_I$ and the compound transform $\mathbf{G}$ by algebraic operations without requiring any further statistical information (Proposition 3.2). Similarly, given $\mathbf{G}$ and the insignificant set, we can obtain the corresponding linear estimation matrix without requiring further statistical information (Proposition 3.1). Hence the sparsity constraint used to generate the estimate summarizes the required statistics and the manner in which this sparsity constraint is obtained determines the type of statistics utilized. Let $E[\ldots]$ denote expectation.

1) Optimal Apriori Sparsity Constraints and Ensemble Statistics: When the utilized sparsity constraint is determined apriori for a class of signals, i.e., when the sparsity constraint is not allowed to adapt to each signal, the optimal estimate minimizes

$$E[||x - y||^2], \tag{21}$$

where y is given through Equation (19), with $\mathbf{A}$ *fixed* for the entire class of signals to yield

$$E[||x_1 - \mathbf{A}x_0||^2], \tag{22}$$

as the mean squared error. In this case, assuming that the covariance of x_0 is full rank, it is well known that the optimal $\mathbf{A}$ is given by ([39])

$$\mathbf{A}_e^* = E[x_1 x_0^T](E[x_0 x_0^T])^{-1}, \tag{23}$$

i.e., the optimal sparsity constraint should be chosen such that one forms the optimal linear estimate using second order ensemble statistics.

Proposition 3.3: Assume the covariance matrix $E[x_0 x_0^T]$ is full rank. Then, the optimal transform $\mathbf{G}$ (and associated insignificant set) establishing the minimum mean squared error linear estimate can be obtained through Proposition 3.2 (up to rotations) using $\mathbf{A}_e^*$ given by Equation (23).

Remark: Observe that the optimal estimate is restricted to construct the minimum mean squared error *linear* estimate, due to the linear nature of the estimation formulation. Observe also that the cases where $E[x_0 x_0^T]$ is of reduced rank can be handled in a straightforward fashion by restricting the quantities to nonzero eigenspaces.

Note that $\mathbf{A}_e^*$ and hence $\mathbf{G}_I$ are fixed for the entire class, i.e., by observing a particular signal we do not make any changes to $\mathbf{A}_e^*$ or $\mathbf{G}_I$. This is a technique motivated by linear approximation where sparsity constraints are determined in a signal invariant fashion by the use of the apriori ordering of the basis functions in Equation (2). Let us now consider adaptive sparsity constraints to see connections to nonlinear approximation.

2) Optimal Adaptive Sparsity Constraints and Conditional Statistics: If the utilized sparsity constraint is allowed to adapt to each signal in the class, then the optimal estimate is one that minimizes the *conditional* mean squared error given that we have observed x_0

$$E[||x - y||^2 | x_0], \tag{24}$$

where y is again given through Equation (19), $E[\dots|x_0]$ indicates that the expectation is conditioned on x_0, and this time $\mathbf{A}$ in Equation (19) is chosen to vary for each realization. The optimal $\mathbf{A}$ is found by minimizing

$$E[||x_1 - \mathbf{A}x_0||^2 | x_0]. \tag{25}$$

It can be seen that for $x_0 \neq 0$, the optimal $\mathbf{A}$ varies with x_0 and it satisfies ([39])

$$\mathbf{A}_c^*(x_0)x_0 = E[x_1|x_0]. \tag{26}$$

The optimal sparsity constraint is thus chosen to vary for each realization to result in such a value for $\mathbf{A}_c^*(x_0)$.

Proposition 3.4: Assume $x_0 \neq 0$. Then the optimal transform $\mathbf{G}$ (and associated insignificant set) establishing the minimum mean squared error estimate can be obtained through Proposition 3.2 (up to rotations) using any $\mathbf{A}_c^*(x_0)$ satisfying Equation (26).

Remark: Observe that unlike the ensemble case, which is constrained to construct the minimum mean squared error *linear* estimator, when $x_0 \neq 0$, the conditional case can reconstruct *the* optimal minimum mean squared error estimator. (We will see later that our adaptive estimation process will yield $\hat{x}_1 = 0$, whenever $x_0 = 0$).

Corollary 3.5: Assume $x_0 \neq 0$. Then,

$$E[||x_1 - \mathbf{A}_c^*(x_0)x_0||^2 | x_0] \leq E[||x_1 - \hat{x}_1||^2 | x_0], \tag{27}$$

and

$$E[E[||x_1 - \mathbf{A}_c^*(x_0)x_0||^2 | x_0]] \leq E[||x_1 - \hat{x}_1||^2], \tag{28}$$

for any estimate $\hat{x}_1$ of x_1, and in particular, the mean squared errors resulting from conditional estimates due to optimal adaptive sparsity constraints ($\mathbf{A}_c^*(x_0)x_0$) and ensemble estimates due to optimal apriori sparsity constraints

$(\mathbf{A}_e^* x_0)$ satisfy

$$E[||x_1 - \mathbf{A}_c^*(x_0)x_0||^2|x_0] \leq E[||x_1 - \mathbf{A}_e^* x_0||^2|x_0], \tag{29}$$

and

$$E[E[||x_1 - \mathbf{A}_c^*(x_0)x_0||^2|x_0]] \leq E[||x_1 - \mathbf{A}_e^* x_0||^2]. \tag{30}$$

Proof: The inequalities follow since $\mathbf{A}_c^*(x_0)x_0$ is the optimal conditional estimator given x_0 [39]. □

Remark: In the sense of Equation (27), it is a misnomer to refer to conditional estimates constructed through adaptive sparsity constraints as linear, since they have the capacity to construct *the* minimum mean squared error estimate.

The significance of the corollary can be seen in the following sense for linear/nonlinear approximation based estimators. For nonstationary signals like images, edges and other localized singularities play important roles. It is well-known that ensemble statistics "hide" the influence of edges, and estimation or approximation based on ensemble statistics has poor performance on images and similar nonstationary signals [10], [4]. Indeed, the performance difference between estimators based on ensemble statistics and those based on conditional statistics is in general overwhelmingly in favor of estimators based on conditional statistics on images. In other words, the performance difference in inequalities (29) and (30) is likely to be substantial on images and similar nonstationary signals. Note that it is not possible to tap into this performance difference with linear approximation based techniques since such techniques, with their apriori choices, are lower bounded by the right side of the inequalities. We next consider where the estimates constructed in this work fit in.

3) Estimates Constructed in This Work: In this work we will be constructing sparsity constraints conditioned on x_0 with the ultimate aim of constructing estimators of x_1 that minimize Equation (25). Intuitively, if we assume that the class of signals we are working on allows for "good" conditional estimates, i.e., if we are working on a set of signals $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$ such that

$$||x_1 - \mathbf{A}_c^*(x_0)x_0|| < T, \tag{31}$$

for some $T > 0$ independent of x, then using Proposition 3.2, we may expect x to be close to a nonlinear approximation class in some basis. This intuition can then be statistically formalized by generalizing Equation (31) to hold almost surely or in mean squared error, etc., in order to define such classes of signals. However, it is of course possible to construct arbitrarily sophisticated processes for which Equation (31) and its generalizations hold but there is no *single* basis (or associated nonlinear approximation class) that can be used to approximate all of the resulting set of signals.

In this work, our primary assumption will be that the target class of signals allow nonlinear approximation with a well-designed localized transform (and its overcomplete extension to be discussed later) to provide good estimates, i.e., nonlinear approximation in some basis yields close approximations to x using Equation (3), with $K = K(T)$. In our framework we will be choosing $\mathbf{G}$ apriori but we will allow the insignificant set to vary for each signal. As a

result, in practice we will only be able to construct a limited number of estimation matrices $\mathbf{A}_c^*(x_0)$[3]. This number and the resulting $\mathbf{A}_c^*(x_0)$'s will determine the class of signals our method will be successful on in a statistical sense, i.e., the class of signals for which our techniques will perform well using Equation (25). The reader should note that it is possible to use the results of this work to design signal specific transforms or to adaptively choose them from a dictionary of basis in order to further the performance of our estimates and to expand on the class of signals over which successful estimation is possible. We will outline one such procedure consistent with the developed adaptive algorithms in Part II.

Using the given $\mathbf{G}$, the class of signals we can perform successful estimation under the mean squared error metric can readily be established as the class for which

$$E[||x_1 - \hat{x}_1(\hat{V}(x_0))||^2] << E[||x_1||^2], \tag{32}$$

where $\hat{V}(x_0)$ is an adaptive estimate of $V(x)$ and reflects the fact that we will be determining the insignificant set from incomplete data in Part II. Since $\hat{V}(x_0)$ is expected to vary for each estimate, $Z = card(\hat{V}(x_0))$ also varies for each estimate. However, we can always adjust our technique to perform estimation only when Z exceeds a predetermined value Z_0. The resulting class of signals can then be tied to nonlinear approximation classes for which one can perform successful approximation using the given transform basis by keeping no more than $K = N - Z_0$ coefficients [11] (see also Sections II and Section Part II - II-B).

Note that the relation to nonlinear approximation is not exact in two ways. First, our adaptive determination of sparsity constraints through $\hat{V}(x_0)$ may deviate from the "Z_0 smallest in magnitude transform coefficients" as would be demanded by formal nonlinear approximation, i.e., our adaptive determination of the insignificant set $\hat{V}(x_0)$ may deviate from the ideal $V(x)$. As stated earlier, by using layering algorithms and localized transforms progressively in Part II, we will strive to limit possible discrepancies under basic assumptions. Second, since nonlinear approximation classes are generally defined using continuous time analysis, obtaining very precise connections that apply to finite dimensional vectors is difficult. Regardless, by using continuous time analysis and arguments based on harmonic analysis results, researchers have advocated various discrete time transforms that are expected to have good nonlinear approximation properties on different classes of discrete time signals (see for e.g., [4], [38], [7], and references therein). The arguments of this section are intended in similar vein to allow connections with that line of transform design research. In particular, we would like to argue that we generally expect these well-known "good" transforms to be successful in our framework. We further expect the results from this work to serve as another performance benchmark, similar to denoising literature, for the ability of transforms in representing various image regions under nonlinear approximation.

IV. Denoising Reconstructions

In this section we establish two computational results that allow us to implement sparse reconstructions via denoising iterations. Our main purpose is to make way for progressive estimates and to establish connections with

[3]Letting Z denote the cardinality of the insignificant set, we can obtain a straightforward bound to the number of estimation matrices we can construct in this simple setting as $\sum_{Z=0}^{N} \begin{pmatrix} N \\ Z \end{pmatrix} = 2^N$.

threshold based denoising techniques. However, denoising iterations will also help accommodate transforms that have basis functions of large spatial support (such as wavelet basis functions corresponding to coarse resolutions), which may otherwise dictate large matrix dimensions when solving linear systems like Equation (14). We first formulate a procedure that solves Equation (14) using iterations. We start with a single orthonormal transform (and sparsity constraint) but generalize to an overcomplete set of orthonormal transforms (yielding a set of sparsity constraints) in Section IV-A.

The role of the procedures we will derive in this section, in relation to the algorithm derived in Part II, is shown in Figure 6. Inside each progression, the procedures will iteratively carry out steps that denoise and enforce known data to yield an estimate. This estimate will then be used as the initializer for the algorithm to rederive insignificant sets and so on. As we will show, the procedures, if carried out a sufficient number of times, will converge to the solution of relevant equations. The progressions on the other hand, will correspond to a search over wider and wider approximation classes as detailed in Part II. Much of the performance of the algorithm will be due to the

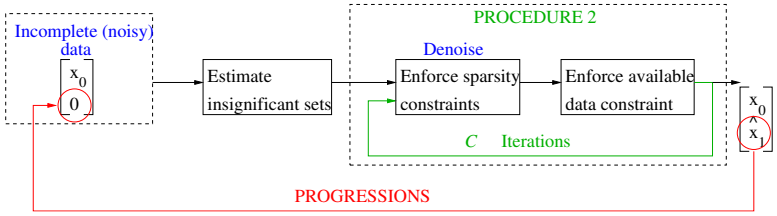


Fig. 6. Simplified outline of the algorithm constructed in Part II. Initial data is used to derive insignificant sets, which are in turn used to construct a denoising operator. Application of this operator in conjunction with available data constraints yields an estimate for each progression. The estimate from each progression is used to reinitialize data and feed the next progression.

utilization of progressive estimates, which tackle the nonconvex optimization issues we alluded to earlier. Starting with an initial "denoising operator", Part II will detail progressive estimates that can be obtained by applying Procedure 2, reinitializing, obtaining a new denoising operator, and then reapplying the procedure, and so on. In this fashion, the final estimate obtained will be of the form $\mathbf{A}_c^*(x_0)x_0$ with the equivalent $\mathbf{A}_c^*(x_0)$ constructed through coarse to fine progressions (or through coarse to fine denoising operators). The reader should keep in mind that that unlike earlier work on denoising [8], or on applying thresholding techniques to inverse problems [9], our method establishes adaptive linear constraints *subject to available information* and produces substantially different estimates by applying denoising iteratively rather than a single application as is done in earlier work.

As before, let $\mathbf{G}$ $(N \times N)$ be a linear, orthonormal transform with portions $\mathbf{G_S}$ $((N-Z)\times N)$ and $\mathbf{G_I}$ $(Z \times N)$, known to produce significant and insignificant coefficients as in Equation (10), i.e., let

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_I \\ \mathbf{G}_S \end{bmatrix}.$$

We postpone the adaptive determination of $\mathbf{G}_I$ to Part II and proceed with the following definitions, which will be useful in the derivations.

Definition 2 (Selection Matrix): Let $\mathbf{S}$ $(N \times N)$ be the diagonal matrix with diagonal entries of 0 and 1 such

that

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{G}_S \end{bmatrix} = \mathbf{SG}. \tag{33}$$

In what follows we will refer to $\mathbf{S}$ as the selection matrix that selects the significant portion of $\mathbf{G}$ or effectively the significant transform coefficients (c_S) of a vector y via $\begin{bmatrix} 0 \\ c_S \end{bmatrix} = \mathbf{SG}y$.

Definition 3 (Recovery Projection Matrix): Assume that the pixels in the original image x are arranged as in Equation (7). We define the diagonal matrix $\mathbf{P}_1$ $(N \times N)$ having diagonal entries 0 and 1 such that

$$\begin{bmatrix} 0 \\ x_1 \end{bmatrix} = \mathbf{P}_1 x. \tag{34}$$

In particular, for an estimate $y = \begin{bmatrix} x_0 \\ \hat{x}_1 \end{bmatrix}$ of x, $\mathbf{P}_1 y = \begin{bmatrix} 0 \\ \hat{x}_1 \end{bmatrix}$.

Orthonormal transform denoising based on hard thresholding of a vector y will obtain the coefficients $\mathbf{G}y$, threshold these coefficients to determine significant ones, i.e., construct $\mathbf{SG}y$ and inverse transform to form $\mathbf{G}^{-1}\mathbf{SG}y$. The following definition formalizes this process.

Definition 4 (Denoising Matrix): Let $\mathbf{D}$ $(N \times N)$ denote the matrix that when applied to a vector y yields a new vector with only the significant components of y via,

$$\begin{aligned} \mathbf{D}y &= \mathbf{G}^{-1}\mathbf{SG}y, \\ \mathbf{D} &= \mathbf{G}^T\mathbf{SG}. \end{aligned} \tag{35}$$

It is important to observe that the hard-thresholding operation is hidden inside $\mathbf{S}$.

Definition 5 (Contraction): We will say that a matrix $\mathbf{B}$ is a contraction if for all vectors y,

$$||\mathbf{B}y|| \leq ||y||. \tag{36}$$

We immediately have the following proposition.

Proposition 4.1: The denoising matrix $\mathbf{D}$ in Definition 4 is a contraction.

Proof: Let $c = \mathbf{G}y$. Using Equation (35),

$$\begin{aligned} y^T\mathbf{D}^T\mathbf{D}y &= y^T\mathbf{D}y \\ &= c^T\mathbf{S}c \\ &\leq c^Tc = ||y||^2, \end{aligned}$$

since $\mathbf{S}$ selects a portion of the coefficients and $\mathbf{G}$ is orthonormal. □

We are now ready to discuss the following simple procedure that solves Equation (14) via iterations.

Procedure 1 (Basic Iterations): Let u $(n_1 \times 1)$ be an arbitrary vector and let $y^0 = \begin{bmatrix} x_0 \\ u \end{bmatrix}$. Let C denote the maximum iteration count. For $k = 0, 1, \ldots, C$, and for a given $\mathbf{D}$, define the iterations

$$y^{k+1} = \mathbf{P}_1\mathbf{D}y^k + (\mathbf{1} - \mathbf{P}_1)y^k, \tag{37}$$

where $\mathbf{1}$ is the $N \times N$ identity.

Remark: Note that $(\mathbf{1}-\mathbf{P}_1)y^k = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$ for all k, and y^{k+1} is obtained by "denoising" y^k (via the term $\mathbf{D}y^k$), taking those pixels in the missing regions ($\mathbf{P}_1\mathbf{D}y^k$), and adding the available information x_0 via the term $(\mathbf{1} - \mathbf{P}_1)y^k$. Observe also that the denoising matrix $\mathbf{D}$ is fixed throughout the iterations, i.e., the coefficient thresholding or selection that is hidden inside $\mathbf{S}$ in Equation (35) is determined in the beginning, and then *kept fixed* throughout the iterations.

Proposition 4.2: The basic iterations of Procedure 1 converge to a vector $y^* = \begin{bmatrix} x_0 \\ \hat{x}_1 \end{bmatrix}$ where $\hat{x}_1$ satisfies Equation (14).

Proof: See Appendix .

Remark: The procedure converges to *a* solution of Equation (14) regardless of u. As mentioned following Equation (14), if necessary, we can ensure a unique solution by zeroing out the contribution of certain subspaces. We will however allow the procedure to stay as is in preparation for the progressive threshold version.

So far we have assumed that there is a single transform that determines the sparsity constraint with which we form reconstructions. We next turn our attention to sparsity constraints formed by an overcomplete set of transforms that allow us to retain our basic formulation while yielding significantly improved descriptions of sparsity.

A. Overcomplete Transforms

In this section we incorporate translation invariant sparsity constraints to further the performance of our estimates. Our motivation is provided by transform based denoising applications, where it is well-known that using an overcomplete bank of transforms provides translation invariant operation and significantly improves performance [5]. The rationale for using translation invariant transforms can be illustrated using the intuitive example in Figure 7 that utilizes translations of a DCT. Observe that each "shift" of the DCT constitutes a signal wide orthonormal transform. The figure illustrates a piecewise smooth image that has two smooth regions separated by an edge[4]. If we adopt the simplified viewpoint where we assume DCT blocks over smooth portions are sparse, and blocks over singularities (i.e., the edge in the figure) are not, then it is easy to see that each of the four DCTs provide a sparse decomposition over slightly different portions of this image. By utilizing an overcomplete set of transforms and combining the insignificant portions of each transform via Equation (39) below, we will establish a much better overall constraint than would be the case by using any of the transforms alone. Note that while overcomplete transforms are typically used to establish translation invariance, we do not concern ourselves with transform design issues and use general notation. This allows us to use overcomplete transforms that are, say, approximately translation invariant (such as complex wavelets [26]), or use a combination of transforms that satisfy other desirable properties.

Let $\mathbf{G}^1, \mathbf{G}^2, \ldots, \mathbf{G}^J$ denote an overcomplete set of orthonormal transforms with each transform arranged so that,

[4]The use of a piecewise smooth image and DCTs is exemplary. As we will see in Section Part II - III, a variety of transforms will provide sparse decompositions over regions that are significantly richer than just the smooth portions of an image.

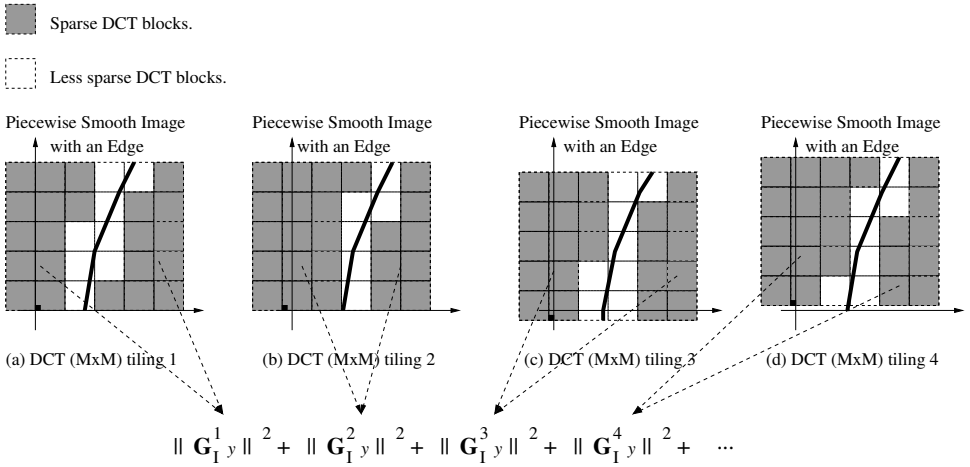


Fig. 7. An overcomplete set of DCTs tiling a piecewise smooth image with an edge. The figures in (a) through (d) show the tiling of the image due to different orthonormal transforms ($\mathbf{G}^1$ through $\mathbf{G}^4$) that are translations of a DCT. If we adopt the simplified viewpoint that DCT blocks over smooth portions of the image lead to a sparse set of coefficients, it becomes easy to visualize blocks from each one of the DCTs contributing to Equation (39). When these contributions are put together as in Equation (39), we obtain a much better description of the sparse portions of the image.

using the notation of Section III-A,

$$\mathbf{G}^l = \begin{bmatrix} \mathbf{G}_I^l \\ \mathbf{G}_S^l \end{bmatrix}, \quad l = 1, \ldots, J,$$

where $\mathbf{G}_I^l$ and $\mathbf{G}_S^l$ are the insignificant and significant portions of the transform $\mathbf{G}^l$. As before we assume that we are given sparsity constraints using each transform, i.e., we are given

$$\mathbf{G}_I^l y = 0, \quad l = 1, \ldots, J, \tag{38}$$

where y is as in Equation (8). Similar to the development immediately before Equation (14), we convert these constraints into a minimization problem where we choose an estimate $\hat{x}_1$ of x_1 that minimizes

$$\sum_{l=1}^{J} ||\mathbf{G}_I^l y||^2 = \sum_{l=1}^{J} ||\mathbf{G}_{I,0}^l x_0 + \mathbf{G}_{I,1}^l \hat{x}_1||^2 \tag{39}$$

$$= \sum_{l=1}^{J} [x_0^T \mathbf{G}_{I,0}^{l\,T} \mathbf{G}_{I,0}^l x_0 + 2 x_0^T \mathbf{G}_{I,0}^{l\,T} \mathbf{G}_{I,1}^l \hat{x}_1 + \hat{x}_1^T \mathbf{G}_{I,1}^{l\,T} \mathbf{G}_{I,1}^l \hat{x}_1].$$

This results in the overcomplete analog of Equation (14) given by

$$(\sum_{l=1}^{J} \mathbf{G}_{I,1}^{l\,T} \mathbf{G}_{I,0}^l) x_0 + (\sum_{l=1}^{J} \mathbf{G}_{I,1}^{l\,T} \mathbf{G}_{I,1}^l) \hat{x}_1 = 0, \tag{40}$$

from which $\hat{x}_1$ can be solved either exactly or within the positive eigenspace of $(\sum_{l=1}^{J} \mathbf{G}_{I,1}^{l\,T} \mathbf{G}_{I,1}^l)$.

In order to provide updated algorithms with minimal notational disruption we provide the following definitions as analogs of those in Section IV. Let $\tilde{\mathbf{G}}$ $(JN \times N)$ denote the "overcomplete transform", i.e.,

$$\tilde{\mathbf{G}} = \begin{bmatrix} \mathbf{G}^{1T} & \mathbf{G}^{2T} & \ldots & \mathbf{G}^{JT} \end{bmatrix}^T. \tag{41}$$

Then the JN coefficients of the overcomplete transform are given by $\tilde{c} = \tilde{\mathbf{G}} y$, and we have

Definition 6 (Overcomplete Selection Matrix): Let $\tilde{\mathbf{S}}$ $(JN \times JN)$ be the diagonal matrix with diagonal entries of 0 and 1 such that

$$\left[\begin{array}{cccccc} \mathbf{0} & \mathbf{G}_S^{1T} & \mathbf{0} & \mathbf{G}_S^{2T} & \ldots & \mathbf{0} \; \mathbf{G}_S^{JT} \end{array} \right]^T = \tilde{\mathbf{S}}\tilde{\mathbf{G}}. \tag{42}$$

Similar to Definition 2, $\tilde{\mathbf{S}}$ is the selection matrix that selects the significant portion of $\tilde{\mathbf{G}}$ or effectively the significant overcomplete transform coefficients $(\tilde{c}_S)$ of a vector y via $\tilde{\mathbf{S}}\tilde{\mathbf{G}}y$.

Observe that the "inverse overcomplete transform" is given by

$$\tilde{\mathbf{G}}^{-1} = \frac{1}{J}\tilde{\mathbf{G}}^T, \tag{43}$$

since

$$\frac{1}{J}\tilde{\mathbf{G}}^T\tilde{\mathbf{G}} = \frac{1}{J}\sum_{l=1}^{J} \mathbf{G}^{l\ T}\mathbf{G}^l = \mathbf{1},$$

where we have used the fact that $\mathbf{G}^{l\ T}\mathbf{G}^l = \mathbf{1}$ due to orthonormal transforms.

Definition 7 (Overcomplete Denoising Matrix): Let $\tilde{\mathbf{D}}$ $(JN \times JN)$ denote the matrix that when applied to a vector y yields a new vector obtained with only the significant overcomplete transform coefficients of y via,

$$\begin{aligned} \tilde{\mathbf{D}}y &= \tilde{\mathbf{G}}^{-1}\tilde{\mathbf{S}}\tilde{\mathbf{G}}y, \\ \tilde{\mathbf{D}} &= \frac{1}{J}\tilde{\mathbf{G}}^T\tilde{\mathbf{S}}\tilde{\mathbf{G}}. \end{aligned} \tag{44}$$

Note that Equation (44) implies $\tilde{\mathbf{D}}y = \frac{1}{J}\sum_{l=1}^{J} \mathbf{G}_S^{lT}\mathbf{G}_S^l y = \frac{1}{J}\sum_{l=1}^{J}(\mathbf{1} - \mathbf{G}_I^{lT}\mathbf{G}_I^l)y$ and hence multiplying y with $\tilde{\mathbf{D}}$ amounts to "overcomplete denoising" of y with the given orthonormal transforms and hard-thresholding [5]. Similar to the earlier case in Definition 4, observe that the hard thresholding operation is hidden inside $\tilde{\mathbf{S}}$. We immediately have the following equivalent of Proposition 4.1.

Proposition 4.3: The overcomplete denoising matrix $\tilde{\mathbf{D}}$ in Definition 7 is a contraction.

Proof: Since $\tilde{\mathbf{D}}$ is an average of denoising matrices, applying the triangle inequality followed by Proposition 4.1 establishes the result. □

The following procedure is the analog of Procedure 1, and it solves Equation (40) via iterations.

Procedure 2 (Overcomplete Iterations): Let u $(n_1 \times 1)$ be an arbitrary vector and let $y^0 = \left[\begin{array}{c} x_0 \\ u \end{array} \right]$. Let C denote the maximum iteration count. For $k = 0, 1, \ldots, C$, consider the iterations

$$y^{k+1} = \mathbf{P}_1\tilde{\mathbf{D}}y^k + (\mathbf{1} - \mathbf{P}_1)y^k, \tag{45}$$

where $\mathbf{1}$ is the $N \times N$ identity.

Remark: Again, note that $(\mathbf{1} - \mathbf{P}_1)y^k = \left[\begin{array}{c} x_0 \\ 0 \end{array} \right]$ for all k, and y^{k+1} is obtained by overcomplete denoising y^k (via the term $\tilde{\mathbf{D}}y^k$), taking those pixels in the missing regions ($\mathbf{P}_1\tilde{\mathbf{D}}y^k$), and adding the available information x_0 via the term $(\mathbf{1} - \mathbf{P}_1)y^k$. As in Procedure 1, observe that the denoising matrix $\tilde{\mathbf{D}}$ is fixed throughout the iterations, i.e., the coefficient thresholding or selection that is hidden inside $\tilde{\mathbf{S}}$ in Equation (44) is determined in the beginning, and then *kept fixed* throughout the iterations.

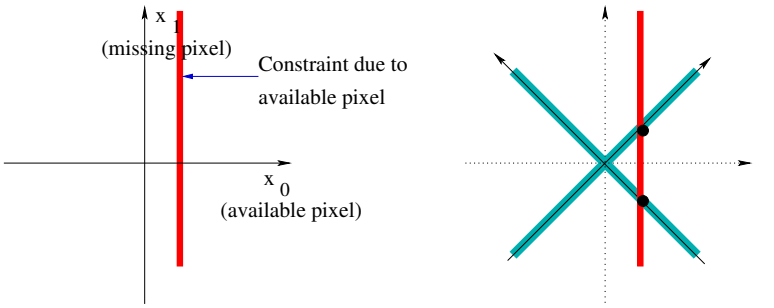
Similar to Proposition 4.2 we have,

Proposition 4.4: The basic iterations of Procedure 2 converge to a vector $y^* = \begin{bmatrix} x_0 \\ \hat{x}_1 \end{bmatrix}$ where $\hat{x}_1$ satisfies Equation (40).

Proof: Similar to the proof of Proposition 4.2 since $\tilde{\mathbf{D}}$ is a positive semidefinite contraction. □

V. Interlude

The main computational result we have established so far is that sparsity constraints can be converted into estimates for the missing data using Procedure 1 or Procedure 2, depending on whether we are utilizing a single transform or an overcomplete set of transforms. In order to establish a complete recovery algorithm that enjoys the favorable properties of nonlinear approximation, it is clear that we also need a means for the adaptive determination of the required sparsity constraints from incomplete data. Since the sparse classes of nonlinear approximation are nonconvex (Figure 2), the determination of sparsity constraints amounts to finding the intersections of non-convex, star-shaped sets with a set that is determined by the available data constraint, $\mathbf{P}_0 x = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$. However, as exemplified in Figure 8, such intersections in general do not result in unique sparsity constraints, and hence, do not lead to unique estimates for the missing data.



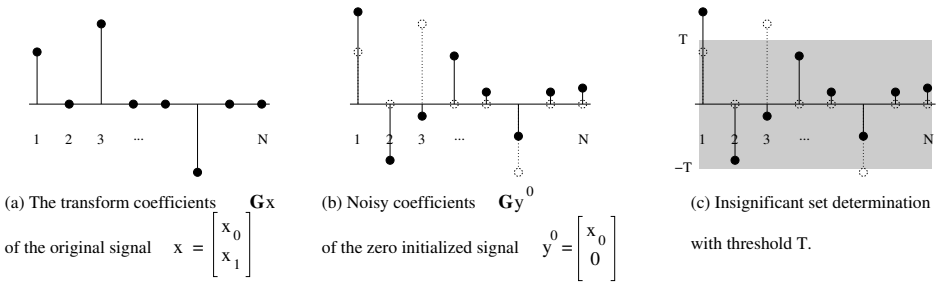
(a) Pixel coordinates for a "two pixel" image.

(b) Non−unique sparsity constraints for class(K=1).

Fig. 8. Insignificant set determination using a two-pixel image. In general, star-shaped classes do not allow for the determination of a unique sparsity constraint from incomplete data. This results in multiple estimates, one for each of the shown intersections.

In Figure 9, we consider insignificant set determination using initial values for the missing data and hard-thresholding, assuming recovery with a single transform. As illustrated in Figure 9 (b), in transform domain one can view the initialized signal as a noisy version of the original, and apply hard-thresholding to determine an initial estimate for the insignificant set. As long as the initialization and the threshold are selected in statistically meaningful ways, we can use the resulting insignificant set to obtain an estimate for the missing data. Sophisticated algorithms that accomplish this using layering (in order to ensure noise in Figure 9 (b) is not overwhelming), statistically meaningful threshold selection (in order to ensure robust insignificant set determination under noise), and selective thresholding (in order to ensure trusted information is given more weight in the estimation) will be introduced in Part II.

Once the sparsity constraints are determined we are reduced to iterations of Procedure 1 for recovery. As shown in

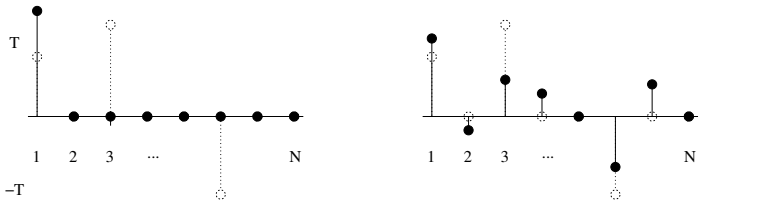


(a) The transform coefficients $\mathbf{G}x$ of the original signal $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$

(b) Noisy coefficients $\mathbf{G}y^0$ of the zero initialized signal $y^0 = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$

(c) Insignificant set determination with threshold T.

Fig. 9. Insignificant set determination in transform domain, using initial values for the missing data. We start with a signal where the missing information is initialized to the mean value of zero in (b), and obtain the set of insignificant coefficients in (c), using thresholding (all coefficients other than the first are deemed insignificant).



(a) Coefficients after thresholding, resulting in the signal $\tilde{y}^0 = \begin{bmatrix} * \\ \hat{} \\ x_1 \end{bmatrix}$

(b) Coefficients after available data constraint, resulting in the signal $y^1 = \begin{bmatrix} x_0 \\ \hat{} \\ x_1 \end{bmatrix}$

Fig. 10. One iteration of Procedure 1 applied to the "noisy" signal in Figure 9 (b), using the insignificant set determined in Figure 9 (c). After the iteration one can carry out more iterations using the same insignificant set or determine a new insignificant set and reapply the procedure.

Figure 10, the application of a single iteration of Procedure 1 amounts to enforcing the sparsity constraints (Figure 10 (a)) followed by enforcing the available data constraint (Figure 10 (b)). Suppose we carry out C iterations of the procedure with C possibly large to ensure convergence to Equation (14) via Proposition 4.2 if desired. Do we have the best estimate for the missing data? Not necessarily. Our determination of the sparsity constraints was carried out in less than ideal conditions, the amount of noise in the data and the threshold selected in relation to the noise may have resulted in a somewhat incorrect insignificant set. Hence, assuming the recovered data is closer to the original compared to the initialization, we can revisit the insignificant set determination with a slightly smaller threshold (to account for the optimistic assumption that we removed some of the noise) and reiterate. Part II will expand further on this threshold reduction strategy, motivate it as a search for the missing data in progressively larger nonlinear approximation classes, and combine it with adaptive insignificant set determination to arrive at our fully autonomous algorithm that accomplishes iterated denoising for image recovery.

VI. Conclusion

In this paper we provide a systematic way of constructing adaptive linear estimators for nonstationary signals through the combination of orthonormal transforms (and filter banks) and insignificant sets. We show that implicitly, any work doing adaptive linear estimation is likewise choosing a linear transform and an insignificant set. The orthonormal transform and index set formulation allows us to establish connections with harmonic analysis results that help determine the classes of stochastic processes over which successful estimation is possible. Through such

connections we recognize that nonconvex data models are fundamental to adaptive linear estimators, i.e., whether the final reconstruction/design equations are convex or not, the underlying problem is inherently nonconvex. With the proposed progressions and adaptive index set determination we deal with the resulting issues directly. A carefully constructed, adaptive recovery algorithm that expands on our results is delegated to Part II of this manuscript.

As we have seen in Procedures 1 and 2, the solution of sparsity constraints subject to available data can be written in terms of denoising iterations through

$$y^{k+1} \quad = \quad \mathbf{P}_1 \mathbf{D} y^k + (\mathbf{1} - \mathbf{P}_1) y^k. \tag{46}$$

Hence adaptive linear estimation techniques can be said to differ through the effective denoising matrix that they implicitly or explicitly choose. Our robust determination of a progressive sequence of denoising matrices (as established through a progressive sequence of thresholds), the resulting iterations, and the simulation results that demonstrate the power of our adaptive algorithms and nonlinear approximation can be found in Part II [16].

The reader versed in linear regression and atomic decomposition literature may be curious about the differences of this work from some established statistical techniques (see for e.g., [21]), especially in light of very recent results that can find maximally sparse solutions under limited scenarios [12]. We conclude with a short summary of the differences leaving the details to other articles [36], [15]. Let $\mathbf{H}$ $(N \times W)$ be a possibly overcomplete matrix of basis vectors with $W \geq N$, and let $y^0 = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$ be an initial estimate of x. In the general regression setting one can pose $y = \mathbf{H}c$ as an estimate of x by minimizing,

$$deviation\_measure(y^0 - y) + \lambda \; regularization\_measure(c). \tag{47}$$

Of course, the problem is not only how one solves (47), but also how one chooses the two measures so that solutions result in *minimum* mse estimates. Using Section III, it is clear that all measures will yield some form of sparsity, which in turn will determine the performance of the resulting estimators. Interesting recent results indicate that if y^0 is very close to x, i.e., if the "noise" due to initial conditions is small, if $\mathbf{H}$ satisfies various stringent properties, if the desired $regularization\_measure$ is the ℓ^0 norm, and if x admits very sparse decompositions using $\mathbf{H}$, then one can use convex techniques (replacing ℓ^0 regularization with ℓ^1 regularization) to find estimates that are very close to the optimal [12]. We note however that such restrictive conditions, including the conditions that allow favorable performance of related techniques, are rarely satisfied in image recovery.

1. *Noise Issues and Nonconvexity:* Let σ^2 denote the variance of a pixel or a suitable bound for it. Observe that in the recovery application, the energy of "noise" due to missing data is $||x - y^0||^2 \cong n_1 \sigma^2$. Hence, unless the missing region is very small, y_0 is not close to x and one can show that there is sufficient noise to deviate recent work from discovering the underlying "true" sparsity, even if measured in the ℓ^0 sense [15]. Our work uses progressions, layering, selective thresholding, and it is designed from the ground up to deal with this issue.

2. *Degree of Sparsity:* While trying to maximize sparsity by assuming x is very sparse may be acceptable in certain applications such as source separation, in missing data prediction, this type of modeling is often erroneous and leads to estimates that are not competitive in mse. For example, even simplified forms of the algorithm of Part II (no layering, no selective hard thresholding), decisively outperform ℓ^1 regularized results using the same

overcomplete basis [36][5]. Clearly we would like to take advantage of any sparsity in the data even if the data is not *very* sparse. Maximizing sparsity is not the goal; correctly determining it's degree, and taking advantage of the "existing" sparsity is. The adaptive algorithms of Part II are specifically designed to do just that.

3. *Robustness to Basis Selection:* Atomic decompositions are overly sensitive to the choice of $\mathbf{H}$ as they insist on restricting the estimate y into $k < N$ dimensional subspaces formed by using only k columns of $\mathbf{H}$, with typically $k << N$. This severely restricts the structure of estimation matrices they can construct, and hence limits their ability in forming versatile estimates. Our reconstructions are significantly different, since our work takes advantage of the overcomplete basis in a very different fashion. The final estimates we obtain simply cannot be written using $k << N$ basis functions of the utilized overcomplete basis (see Figures 1 and 5, as well as Part II).

4. *Lack of Progressions:* Suppose one manages to find an acceptable solution to Equation (47). The solution is acceptable in the sense that it is a better approximation to x, when compared to y^0. Is this the best possible estimate? Replacing y^0 with the found solution and reiterating after adjusting parameters may allow one to construct a better solution. The progressive estimates of this sequence allow such constructions, and enable us to successfully recover signals that are not sparse in the ℓ^0 sense by progressively recovering details under very non-ideal conditions. (Conceptually, after all progressions, an equivalent denoising matrix $\mathcal{D}$ and $regularization\_measure = y^T(\mathcal{D}-\mathbf{1})y$ are generated. The structure of this final $\mathcal{D}$ is complicated though, see Part II, Section Part II - IV).

In conclusion, the formulation of this paper and the algorithms of Part II provide the adaptivity and robustness to handle the issues of missing data recovery. The work of this sequence can recover missing regions very successfully (some with perfect reconstruction) even when the utilized basis fails the requirements of [12], even when such regions cannot be obtained as a combination of $k < N$ basis functions of the utilized basis (let alone $k << N$ basis functions), even when substantial chunks of data are missing and y_0 is substantially away from x, and even when x is not very sparse. Under certain conditions, the estimation matrices $\mathbf{A}_c(x_0)$ that established work constructs will be in the range of the proposed techniques, whereas the matrices that our techniques will prefer to construct will not be in the range of established work.

APPENDIX

Convergence is established if there exists a y^* that satisfies

$$y^* = \mathbf{P}_1\mathbf{D}y^* + (\mathbf{1} - \mathbf{P}_1)y^*, \tag{48}$$

and the sequence $||y^k - y^*||$ converges to 0 regardless of the starting point, i.e., regardless of the value of the vector u. We first show that Equation (48) leads to Equation (14). Starting with Equation (48) we obtain

$$\begin{aligned} 0 &= \mathbf{P}_1(\mathbf{D} - \mathbf{1})y^* = \mathbf{P}_1(\mathbf{G}^T\mathbf{S}\mathbf{G} - \mathbf{1})y^* \\ &= \mathbf{P}_1(\mathbf{G}_S^T\mathbf{G}_S - \mathbf{1})y^* \\ &= \mathbf{P}_1(\mathbf{G}_I^T\mathbf{G}_I)y^* \end{aligned}$$

[5]Comparisons to ℓ^1 regularized results are also important since statistics literature seems to prefer this form of regularization to other forms of regression, including to variants based on boosting and SVMs [21].

$$= \mathbf{P}_1 \begin{bmatrix} \mathbf{G}_{I,0}^T \mathbf{G}_{I,0} & \mathbf{G}_{I,0}^T \mathbf{G}_{I,1} \\ \mathbf{G}_{I,1}^T \mathbf{G}_{I,0} & \mathbf{G}_{I,1}^T \mathbf{G}_{I,1} \end{bmatrix} y^*, \tag{49}$$

which results in

$$0 = \mathbf{G}_{I,1}^T \mathbf{G}_{I,0} x_0 + \mathbf{G}_{I,1}^T \mathbf{G}_{I,1} \hat{x}_1 \tag{50}$$

where we have used Definition 4, Definition 2, properties of orthonormal transforms, and Equation (12) to arrive at Equation (50), which is the same as Equation (14).

It is clear that Equation (50) has at least one solution. Let us refer to the set of all y^* that satisfy Equation (48) as the solution set. To see that $||y^k - y^*|| \to 0$ as $C, k \to \infty$, let $y^{k-1} = y^* + w$ for some vector w. Observe that, by construction we have

$$(\mathbf{1} - \mathbf{P}_1) y^* = (\mathbf{1} - \mathbf{P}_1) y^k = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}, \tag{51}$$

for any k. We thus have $(\mathbf{1} - \mathbf{P}_1) w = 0$ and

$$\begin{aligned} y^k &= (\mathbf{P}_1 \mathbf{D} + (\mathbf{1} - \mathbf{P}_1)) y^{k-1} \\ &= (\mathbf{P}_1 \mathbf{D} + (\mathbf{1} - \mathbf{P}_1))(y^* + w) \\ &= y^* + (\mathbf{P}_1 \mathbf{D} + (\mathbf{1} - \mathbf{P}_1)) w \\ &= y^* + \mathbf{P}_1 \mathbf{D} w \end{aligned} \tag{52}$$

which leads to $||y^k - y^*|| = ||\mathbf{P}_1 \mathbf{D} w||$. Since $\mathbf{D}$ is a contraction so is $\mathbf{P}_1 \mathbf{D}$ and we have

$$||y^k - y^*|| = ||\mathbf{P}_1 \mathbf{D} w|| \leq ||w|| = ||y^{k-1} - y^*||, \tag{53}$$

with equality if and only if $||\mathbf{P}_1 \mathbf{D} w|| = ||w||$. Since $\mathbf{P}_1$ and $\mathbf{D}$ are both contractions, in the equality case it must be that $||\mathbf{P}_1 \mathbf{D} w|| = ||\mathbf{D} w|| = ||w||$. Noting that $\mathbf{D}$ is also symmetric and positive semidefinite, we conclude that $\mathbf{D} w = w$. Hence equality happens if and only if $w = \mathbf{P}_1 \mathbf{D} w$ or if y^{k-1} is also in the solution set. Hence, regardless of u, the procedure converges to *a* solution of Equation (14).

REFERENCES

[1] Z. Alkachouh and M. G. Bellanger, "Fast DCT-Based Spatial Domain Interpolation of Blocks in Images", *IEEE Trans. Image Proc.,* vol. 9, no. 4, pp. 729-732, April 2000.

[2] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-In by Joint Interpolation of Vector Fields and Gray Levels," *IEEE Trans. Image Proc.,* vol. 10, no. 8, pp. 1200-1210, Aug. 2001.

[3] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, "Simultaneous structure and texture image inpainting," *IEEE Trans. Image Proc.*, vol. 12 , no. 8, pp. 882-889, Aug. 2003.

[4] A. Cohen, I. Daubechies, O. G. Guleryuz, and M. T. Orchard, "On the importance of combining wavelet-based nonlinear approximation with coding strategies," *IEEE Trans. Info. Theory*, vol. 48, no. 7, pp. 1895-1921, July 2002.

[5] R. R. Coifman and D. L. Donoho, "Translation invariant denoising," in Wavelets and Statistics, Springer Lecture Notes in Statistics 103, pp. 125-150, New York:Springer-Verlag.

[6] J. S. De Bonet, "Multiresolution sampling procedure for analysis and synthesis of texture images," *Proceedings of ACM SIGGRAPH,* July 1997.

[7] M. N. Do, P. L. Dragotti, R. Shukla, and M. Vetterli, "On the compression of two-dimensional piecewise smooth functions," *IEEE Int. Conf. on Image Proc.*, ICIP 01, Thessaloniki, Greece, Oct. 2001.

[8] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81., pp. 425-455, 1994.

[9] D. L. Donoho, "Nonlinear solution of linear inverse problems by Wavelet-Vaguelette Decomposition," *App. Comp. Harmonic Analysis,* vol. 2, pp. 101-126, 1995.

[10] J.P. D'Ales and A. Cohen, *Non-linear Approximation of Random functions*, Siam J. of Appl. Math 57-2, 518-540, 1997.

[11] R. DeVore, *Nonlinear approximation*, Acta Numerica (7), 1998.

[12] D. Donoho, M. Elad, and V. Temlyakov, "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise", Technical Report, 2004. http://www-stat.stanford.edu/∼donoho/reports.html

[13] P. J. S. G. Ferreira and A. J. Pinho, "Errorless Restoration Algorithms for Band-Limited Images", in *Proc. IEEE Conf. on Image Proc.*, vol. III, pp. 157-161, 1994.

[14] G. H. Golub and C. F. Loan, "Matrix Computations," 3rd edition, Johns Hopkins, 1996.

[15] Onur G. Guleryuz, "On Missing Data Prediction Using Sparse Signal Models: A comparison of Atomic Decompositions with Iterated Denoising," Proc. SPIE Conf. on Wavelets XI, in Mathematical Methods, San Diego, Aug. 2005.

[16] Onur G. Guleryuz, "Nonlinear Approximation Based Image Recovery Using Adaptive Sparse Reconstructions and Iterated Denoising: Part II - Adaptive Algorithms," *IEEE Trans. on Image Processing*.

[17] Onur G. Guleryuz, "Predicting Wavelet Coefficients Over Edges Using Estimates Based on Nonlinear Approximants," *Proc. Data Compression Conference*, IEEE DCC-04, April 2004.

[18] Onur G. Guleryuz, "Iterated Denoising for Image Recovery", Proc. Data Compression Conference, IEEE DCC-02, pp. 3-12, April 2002.

[19] Onur G. Guleryuz, "Nonlinear Approximation Based Image Recovery Using Adaptive Sparse Reconstructions," *Proc. IEEE Int'l Conf. on Image Proc.* (ICIP2003), Barcelona, Spain, Sept. 2003.

[20] Onur G. Guleryuz, E. Lutwak, D. Yang, and G. Zhang, "Information-Theoretic Inequalities for Contoured Probability Distributions," IEEE Transactions on Information Theory, vol. 48, no. 8, pp. 2377-2383, August 2002.

[21] T. Hastie, R. Tibshrirani, and J. Friedman, "The Elements of Statistical Learning," Springer, New York, 2001.

[22] S. S. Hemami and T. H.-Y. Meng, "Transform Coded Image Reconstruction Exploiting Interblock Correlation", *IEEE Trans. Image Proc.,* vol. 4, no. 7, pp. 1023-1027, July 1995.

[23] A. Hirani and T. Totsuka, "Combining Frequency and Spatial Domain Information for Fast Interactive Image Noise Removal", *Proc. SIGGRAPH'96 Conf.,* pp. 269-276, 1996.

[24] ISO/IEC International standard 13818-2, "Generic Coding of Moving Pictures and Associated audio information: Part2 - Video", 1995.

[25] JPEG: ITU-T Rec. T.81-ISO/IEC No. 10918-1, "Information Technology- Digital Compression and Coding of Continuous-Tone Still Images", 1993.

[26] N. G. Kingsbury, "A Dual-Tree Complex Wavelet Transform with improved orthogonality and symmetry properties", in *Proc. IEEE Conf. on Image Processing*, Vancouver, September 11-13, 2000, paper 1429.

[27] P. Kisilev, M. Zibulevshy, and Y. Y. Zeevi, "Blind separation of mixed images using multiscale transforms," *Proc. IEEE Conf. on Image Proc.*, ICIP2003, Barcelona, Spain, 2003.

[28] R. Koenen, "Overview of the MPEG-4 standard," ISO/IEC JTC1/SC29/WG11 N1730, July 1997.

[29] A. C. Kokaram, R. D. Morris, W. J. Fitzgerald, and P. J. W. Rayner, "Interpolation of Missing Data in Image Sequences", *IEEE Trans. Image Proc.,* vol. 4, no. 11, pp. 1509-1519, Nov 1995.

[30] X. Lee, Y-Q. Zhang, and A. Leon-Garcia, "Information Loss recovery for Block-Based Image Coding Techniques - A Fuzzy Logic Approach", *IEEE Trans. Image Proc.,* vol. 4, no. 3, pp. 259-273, March 1995.

[31] S. Mallat, *A Wavelet Tour of Signal Processing.* New York: Academic Press, 1998

[32] A. Papoulis, "A new algorithm in spectral analysis and band-limited extrapolation," *IEEE Trans. Circuits and Syst.*, vol. 22, pp. 735-742, 1975.

[33] J. W. Park, J. W. Kim, and S. U. Lee, "DCT Coefficients Recovery-Based Error Concealment Technique and Its Application to the MPEG-2 Bit Stream Error", *IEEE Trans. CSVT,* vol. 7, no. 6, pp. 845-854, Dec. 1997.

[34] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients", *Int'l Journal of Comp. Vision,* vol. 40, pp. 49-71, Oct. 2000.

[35] R. Schneider, "Convex Bodies : The Brunn-Minkowski Theory," Cambridge University Press, March 2003.

[36] Ivan W. Selesnick, Richard Van Slyke, and Onur G. Guleryuz, "Pixel Recovery via ℓ_1 Minimization in the Wavelet Domain," Proc. IEEE Int'l Conf. on Image Proc. (ICIP2004), Singapore, Oct. 2004. Please see related presentation slides at http://eeweb.poly.edu/∼onur/online_pub.html for results.

[37] S. Shirani, F. Kossentini, and R. Ward, "Reconstruction of Baseline JPEG Coded Images in Error Prone Environments", *IEEE Trans. Image Proc.,* vol. 9, no. 7, pp. 1292-1299, July 2000.

[38] J. L. Starck, E. J. Candes and D. L. Donoho, "The Curvelet Transform for Image Denoising," *IEEE Transactions on Image Processing*, vol 11, pp. 670-684,

[39] H. Stark and J. W. Woods, "Probability, Random Processes, and Estimation Theory for Engineers," Prentice Hall, Englewood Cliffs, NJ, 1986.

[40] H. Sun and W. Kwok, "Concealment of Damaged Block Transform Coded Images Using Projections onto Convex Sets", *IEEE Trans. Image Proc.,* vol. 4, no. 4, pp. 470-477, April 1995.

[41] Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, "Error Resilient Video Coding Techniques," *IEEE Signal Proc. Magazine*, July, 2000, pp. 61-82.

[42] Y. Wang, and Q-F. Zhu "Error Control and Concealment for Video Communication: A Review," *Proceedings of the IEEE*, vol. 86, no. 5, May, 1998, pp. 974-997.

[43] Y. Wang, Q-F. Zhu, and L. Shaw, "Maximally smooth image recovery in transform coding", *IEEE Trans. Comm.*, vol. 41, pp. 1544-1551, Oct. 1993.