Nonlinear Approximation and Image Representation using Wavelets

Sudipto Guha and Boulos Harb

Abstract

We address the problem of finding sparse wavelet representations of high-dimensional vectors. We present a lower-bounding technique and use it to develop an algorithm for computing provably-approximate instance-specific representations minimizing general ℓ_p distances under a wide variety of compactly-supported wavelet bases. More specifically, given a vector $f \in \mathbb{R}^n$, a compactly-supported wavelet basis, a sparsity constraint $B \in \mathbb{Z}$, and $p \in [1, \infty]$, our algorithm returns a B-term representation (a linear combination of B vectors from the given basis) whose ℓ_p distance from f is a $O(\log n)$ factor away from that of the optimal such representation of f. Our algorithm applies in the one-pass sublinear-space data streaming model of computation, and it generalize to weighted p-norms and multidimensional signals. Our technique also generalizes to a version of the problem where we are given a bit-budget rather than a term-budget. Furthermore, we use it to construct a *universal representation* that consists of at most $B(\log n)^2$ terms and gives a $O(\log n)$-approximation under all p-norms simultaneously.

1 Introduction

Consider the following problem: Given a vector in high-dimensional space, represent it as *closely as possible* using a linear combination of a *small number* elements of a pre-defined dictionary. These *sparse* representations, so-called because the number of dictionary elements we are constrained to use is much smaller than the dimension of the target vector, have become more pertinent in light of the large amounts of data that we encounter. The benefits we gain from the sparsity, however, are counteracted by a loss in our representation's fidelity and its ability to model the signal accurately. A sparse representation will be, in general, an *approximation* to the given vector, and the quality of this approximation is affected by both the sparsity constraint and the choice of dictionary to utilize for the representation.

As indicated above, we will work in the discrete setting, and accordingly, we assume that we are given a vector f in high-dimensional space $\mathbb{R}^n$. We are also given a dictionary of m elements of $\mathbb{R}^n$, $\{a_1, a_2, \cdots, a_m\}$, and an integer B which we have referred to earlier as the *sparsity constraint*; hence, B is typically much smaller than n. The goal is to represent f as a linear combination of B elements of the dictionary. That is, we wish to find a vector $\hat{f}$ with

$$\hat{f} = \sum_{k \in \mathcal{S} \,:\, |\mathcal{S}|=B} x_k a_k \ , \tag{1}$$

Dagstuhl Seminar Proceedings 07071
Web Information Retrieval and Linear Algebra Algorithms
http://drops.dagstuhl.de/opus/volltexte/2007/1063

that is a good representation of f. We need a way to measure the success of our candidate representation, and for that we will use its ℓ_p distance from the target vector f: $\left\|f - \hat{f}\right\|_p$. Recall that if $y \in \mathbb{R}^n$, then

$$\|y\|_p = \begin{cases} \left(\sum_{i=1}^n |y_i|^p\right)^{1/p} & 1 \le p < \infty \\ \max_{i=1,\ldots,n} |y_i| & p = \infty \end{cases} .$$

As a simple example, suppose as before that we are given a set of observations $f_i = f(t_i)$ of an underlying signal f taken at times $t_i, i = 1, \ldots, n$. Suppose we wish to model the signal f as a linear combination of simpler functions $\varphi_k, k = 1, \ldots, B$ where the performance of our model is measured using the least-squares error. Then we can set $a_k[i] = \varphi_k(t_i)$ for $i = 1, \ldots, n$ and $k = 1, \ldots, B$, and our task becomes that of finding $\hat{f} = \sum_{k=1}^B x_k a_k$ that minimizes $\sum_{i=1}^n |f_i - \hat{f}_i|^2$. A few things to note about this example. First, if we set $A_{ik} = a_k[i]$, then we are simply solving the familiar least-squares regression problem $\|f - Ax\|_2$. Second, we designed our dictionary $\{a_k\}_{k=1}^B$. In general, however, we will assume that the dictionary will be given to us. Finally, we were not asked to *choose* a set of B vectors for our representation. Instead, we were given the B vectors, and we were allowed to use all of them. This is an important distinction between two approaches known as *linear* and *nonlinear approximation*.

A B-term *linear approximation* of f is specified by a linear combination of a *given* B elements from the dictionary. In other words, the set $\mathcal{S}$ of indices of dictionary elements in expression (1) is given to us, and all we need to do is to compute the x_k's in $\hat{f} = \sum_{k \in \mathcal{S}} x_k a_k$ that minimize $\left\|f - \hat{f}\right\|_p$ for the desired p-norm. For simplicity of notation, we may assume that we are given the *first* B elements of the dictionary; i.e., $\hat{f} = \sum_{k=1}^B x_k a_k$. Notice that the vector space from which we are approximating any $f \in \mathbb{R}^n$ is the linear space $\mathcal{F}_{[B]} = \operatorname{span}\{a_k, k \in [B]\}$, which explains why this type of approximation is called *linear approximation*. The error of the best representation of f in this space is given by the ℓ_p distance of f from $\mathcal{F}_{[B]}$: $\mathcal{E}_{[B],p} = \min_{\hat{f} \in \mathcal{F}_{[B]}} \left\|f - \hat{f}\right\|_p$. For example, in the case of the 2-norm, this is a least-squares regression problem whose solution is the projection of f onto $\mathcal{F}_{[B]}$. More generally, the problem can be solved using convex programming methods (see, e.g., [1]).

In *nonlinear approximation*, neither the set of indices $\mathcal{S}$ of dictionary elements nor their corresponding coefficients in the representation $\hat{f} = \sum_{k \in \mathcal{S}} x_k a_k$ are given to us. The choice of the B vectors to use in the representation is *instance specific*, meaning it depends on the target vector f. This type of approximation replaces the linear space $\mathcal{F}_{[B]}$ with the space of all vectors that can be represented as a linear combination of *any* B dictionary elements,

$$\mathcal{F}_B = \left\{\sum\nolimits_{k \in \mathcal{S}} x_k a_k : x_k \in \mathbb{R}, \mathcal{S} \subset [n], |\mathcal{S}| = B\right\} .$$

This is a *nonlinear* space since, in general, the sum of two arbitrary vectors in $\mathcal{F}_B$ uses more than B elements from the dictionary, and thus does not belong to the space. We measure the error of a candidate representation as before using its ℓ_p distance from f, and the error of the best representation of f in $\mathcal{F}_B$ is given by

$$\mathcal{E}_{B,p} = \min_{\hat{f} \in \mathcal{F}_B} \left\|f - \hat{f}\right\|_p .$$

We address this nonlinear approximation problem when the given dictionary is a *wavelet basis*:

Problem 1.1 (B-term Representation)**.** *Given $f \in \mathbb{R}^n$, $p \in [1, \infty]$, a compactly-supported wavelet basis for $\mathbb{R}^n$ $\Psi = [\psi_1, \psi_2, \cdots, \psi_n]$, and an integer B, find a solution vector $x \in \mathbb{R}^n$ with at most B non-zero components x_i such that $\|\Psi x - f\|_p$ is minimized.*

We will refer to this problem as the *unrestricted* B-term representation problem in order to contrast it with a *restricted* version where the non-zero components of x can only take on values from the set $\{\langle f, \psi_i \rangle, i \in [n]\}$. That is, in the restricted version, each x_i can only be set to a coefficient from the wavelet expansion of f using Ψ, or zero.

Additionally, we wish to be able to compute our representation without having to store the whole target vector f. This is important especially when n is very large, and f, for instance, is a time-series signal (for example, f is a set of n observations $f_i = f(t_i), i = 1, \ldots, n$, taken in succession). It implies, however, that our algorithms can only use sublinear space. More general versions of the problem that we also consider (but do not present here) include having a bound on the number of bits rather than the number of coefficients; and, having multiple bases in the dictionary among which to choose for the representation.

In addressing the problems mentioned above we focus on measuring the error (or goodness) of our representation using general ℓ_p norms. These problems have all been studied under the ℓ_2 error measure; however, it is not clear what techniques are needed for solving them under other ℓ_p norms. Unless a space is equipped with the ℓ_2 norm, it ceases to be a Euclidean space; hence, we lose all the convenient properties of the associated inner-product. For example, even the notion of projection is not entirely intuitive under non-ℓ_2 norms. Under the ℓ_2 norm, the techniques we use coincide with known ℓ_2-specific ones; hence, they extend these existing techniques.

1.1 Background

Let us start by giving a brief idea of wavelet bases. A wavelet basis $\{\psi_k\}_{k=1}^n$ for $\mathbb{R}^n$ is a basis where each vector is constructed by dilating and translating a single function referred to as the *mother wavelet* ψ. For example the Haar mother wavelet, due to Haar [15], is given by:

$$\psi_H(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2 \\ -1 & \text{if } 1/2 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}$$

The Haar basis for $\mathbb{R}^n$ is composed of the vectors $\psi_{j,s}[i] = 2^{-j/2}\psi_H\left(\frac{i - 2^j s}{2^j}\right)$ where $i \in [n]$, $j = 1, \ldots, \log n$, and $s = 0, \ldots, n/2^j - 1$, plus their orthonormal complement $\frac{1}{\sqrt{n}} 1^n$. This last basis vector is closely related to the Haar *multiresolution scaling function* $\phi_H(t) = 1$ if $0 \leq t < 1$ and 0 otherwise. In fact, there is an explicit recipe for constructing the mother wavelet function ψ from ϕ [18, 22] (see also Daubechies [19], and Mallat [6]). Notice that the Haar mother wavelet is compactly supported on the interval $[0, 1)$. This wavelet, which was discovered in 1910, was the only known wavelet of compact support until Daubechies constructed a family of compactly-supported wavelet bases [5] in 1988 (see also [6, Chp. 6]).

Why wavelets? Wavelets are fundamental in the field of *nonlinear approximation theory*. Nonlinear approximation has a rich history starting from the work of Schmidt [25] in 1907, and has been studied under various contexts. More recently, in a substantial review, DeVore [8] explores the advantages of nonlinear approximation over the simpler linear one. The advantages are investigated in terms of the rate of decay of the approximation error relative to the number of terms in the approximate representation. Indeed, the question usually considered in approximation theory is: Given a basis and a parameter $\alpha > 0$, what is the class of functions whose B-term approximation error $\mathcal{E}_{B,p}$ decays like $O(B^{-\alpha})$? The success of *wavelet bases* in this context was first displayed by DeVore, Jawerth, and Popov [9]. They show that if f is sufficiently "smooth" (belongs to a Besov space), then its B-term approximation error using certain wavelet bases decays quickly. In fact, they show that the (properly normalized) wavelet coefficients of functions belonging to these function spaces decay rapidly; hence, retaining the largest B of the coefficients suffices to give the result. (See also the survey by Temlyakov [27]). This leads to wavelet-based algorithms for noise reduction (called *wavelet shrinkage*) developed by Donoho and Johnstone [10] (see also Donoho *et al.* [11]). The idea here is that "large" (i.e., larger than a threshold) wavelet coefficients mostly carry signal information, while "small" wavelet coefficients are thought to be mostly noise and can be set to zero. Under a Gaussian noise model, they find thresholds that minimize the expected ℓ_2 error.

Wavelets have since found extensive use in image representation and signal compression (see, e.g., Mallat [19], and Stollnitz, Derose, and Salesin [26]). The basic paradigm for using wavelets in image compression is shown in Figure 1, and it goes as follows [8]. The problem is viewed as a nonlinear approximation with the image as the target function. The wavelet expansion coefficients of the image are computed and the smallest of these are pruned using a threshold. Quantization of the remaining coefficients then takes place. These remaining quantized coefficients are the representation of the approximate image (which can be reconstructed using the inverse wavelet transform). Finally, the representative coefficients are compressed using a loss-less encoder. A more direct encoding approach, however, would optimize the number of bits stored directly. Cohen *et al.* [3] show decay results for this *bit-budget* version of the problem that are similar to those developed be DeVore *et al.* [9].

Why are the largest wavelet coefficients retained? Retaining the largest coefficients is the optimal strategy for minimizing the ℓ_2 norm in a nonlinear approximation. In the nonlinear compression procedure above, we are computing a representative image $\hat{f}$ that approximates the original image f and minimizes the ℓ_2 error $\mathcal{E}_{B,2} = \min_{\hat{f}\in\mathcal{F}_B} \|f - \hat{f}\|_2$. As in the case for linear approximation, this problem is well-understood under this error measure. The optimal representation $\hat{f} \in \mathcal{F}_B$ simply retains the B wavelet coefficients $\langle f, \psi_k\rangle$ that are largest in absolute value; i.e., if $|\langle f, \psi_{k_1}\rangle| \geq |\langle f, \psi_{k_2}\rangle| \geq \cdots \geq |\langle f, \psi_{k_n}\rangle|$, then $\hat{f} = \sum_{i=1}^{B} \langle f, \psi_{k_i}\rangle \psi_{k_i}$. In fact, this is true for any orthonormal basis when the error is measured under the ℓ_2 norm.

Recently, wavelets have been utilized in databases to facilitate selectivity estimation in query optimization [20], for approximate query processing [2, 28, 29], and for content-based image querying [17, 23]. For example, Chakrabarti *et al.* [2] show how to create synopses of relational tables using wavelet coefficients, then perform database queries, that are fast but approximate, over these compact sets of retained coefficients. Many researchers observe,

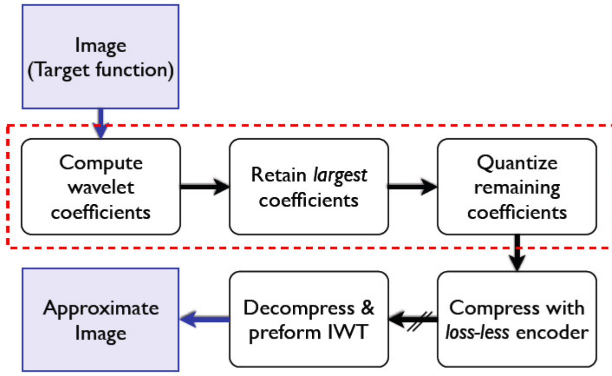


Figure 1: Image compression using wavelets. The dashed box highlights the nonlinear approximation that takes place.

however, that the choice of least-squares error is not natural for the representation of data or distributions. Matias, Vitter, and Wang [20], for example, suggest using the ℓ_1 and the ℓ_∞ error measures. The ℓ_1 norm is a *robust* measure that is less sensitive to outliers than the ℓ_2 norm, and it is well-suited for measuring distributions. Even in image compression, Mallat [19, p. 528] and Daubechies [6, p. 286] point out that while the ℓ_2 measure does not adequately quantify perceptual errors, it is used, nonetheless, since other norms are difficult to optimize.

1.2 Contribution

We develop approximation algorithms for finding nonlinear approximate representations that minimize general ℓ_p error measures (including ℓ_∞) under a large class of wavelet bases. We present a greedy algorithm with performance guarantees when the given wavelet is compactly supported, and we develop a dynamic programming algorithm more suited for use in the Haar wavelet case. Given an arbitrary target vector $f \in \mathbb{R}^n$, a sparsity constraint B, a p-norm, and, for example, any Daubechies wavelet basis for $\mathbb{R}^n$, our greedy algorithm constructs a function $\hat{f} \in \mathcal{F}_B$ whose representation error $\|f - \hat{f}\|_p$ is no more than $O(\log n)\mathcal{E}_{B,p}$. The algorithm runs in linear time, performs one pass over the given function, and requires logarithmic space $O(\log n + B)$ to compute this solution $\hat{f}$. Hence, our algorithm lends itself well to large data streams where input arrives rapidly as a time-series signal. The crux of the argument is a nonlinear dual program that we use to obtain a lower bound on the optimal representation error. The dual program is easy to solve and its optimal solution is comprised of only B wavelet coefficients; therefore, its solution is also a solution to the restricted problem, and it establishes a $O(\log n)$ gap between the restricted and unrestricted versions of the problem. This program can be a powerful tool for tackling related problems in nonlinear approximation; for example, it can immediately be applied to the bit-budget version of the problem mentioned earlier in the context of image compression. Furthermore, we use it to show that by storing $B(\log n)^2$ coefficients, we obtain a universal representation that $O(\log n)$-approximates the optimal representations for all p-norms *simultaneously*. That is, given f, if $\hat{f}_u$ is this universal representation, then

$\forall p \in [1, \infty]$, $\left\| f - \hat{f}_u \right\|_p \leq O(\log n)\mathcal{E}_{B,p}$.

In the specific case of the Haar wavelet, we have a dynamic programming algorithm that constructs a solution $\hat{f}$ whose representation error is arbitrarily close to the optimal error $\mathcal{E}_{B,p}$. This algorithm also performs one pass over the target function, and requires sublinear space (i.e., it is streaming) when $p > 1$ (we do not present this algorithm here).

Both of our algorithms extend to multi-dimensional signals (belonging to $\mathbb{R}^{n^c}$, for fixed $c \in \mathbb{N}$) and to certain dictionaries composed of multiple wavelet bases. When the dictionary is not composed of a single wavelet basis, the representation problem becomes a form of *highly nonlinear approximation* (in fact, the arbitrary dictionary case, which as we said is NP-hard even to approximate [7, 12] is considered a form highly nonlinear approximation). A notable example of such *redundant* dictionaries are those of Coiffman and Wickerhauser [4], which are binary tree-structured dictionaries composed of $O(n \log n)$ vectors and an exponential number of bases. Given such a tree-structured dictionary of wavelet bases, our algorithms can be used (and combined) to find an approximately-best basis for representing the target $f \in \mathbb{R}^n$ using B-terms under any given ℓ_p error measure. We believe that our two approximation algorithms and their extensions may be used effectively in those new domains where wavelets are being utilized for summarizing large data.

1.2.1 Future Work

One immediate application to investigate is to utilize these sparse wavelet representation algorithms for image categorization and search. In text classification, documents are usually represented as sparse feature vectors in high-dimensional space. Images, however, do not lend themselves easily to sparse representations since the inter-pixel relationships are very important for correctly *understanding* the image. Our algorithms provide a way to represent images sparsely in a meaningful, provably-approximate manner. These wavelet representations may then be combined with well-known data analysis methods, such as Regularized Least Squares Classification (RLSC) and Support Vector Machine (SVM) classification, in order to classify the images, which is an important step for better image search. Engineering decisions include the choice of wavelet, the sparsity bound, and the ℓ_p norm used for measuring the representation error.

2 An Approximation Algorithm for Representing Data Streams using Wavelets

In this section, we first present our greedy $O(\log n)$-approximation algorithm and demonstrate its use for sparsely representing images (Section 2.1). We then show an analysis of the performance (time and space complexities as well we approximation guarantee) of our algorithm using the dual system of constraints mentioned earlier (Section 2.2). Finally, we show how to compute a *universal representation* that stores $B(\log n)^2$ coefficients and gives a $O(\log n)$ approximation to all p-norms *simultanuously* (Section 2.3).

2.1 Sparse Image Representation under non-ℓ_2 Error Measures

In this section we present our greedy algorithm and give three examples that demonstrate uses for it in compressing images. A non-streaming version of the algorithm for Haar and Daubechies wavelets was implemented in MATLAB using the *Uvi_Wave.300* toolbox[1] [24]. Pseudocode of the implementation is provided below in Figure 2. The algorithm takes four parameters as input: the image X, the number of coefficients to retain B, the p-norm to minimize, and the type of Daubechies wavelet to use. The last parameter, q, determines the number of non-zero coefficients in the wavelet filter. Recall that the Haar wavelet is the Daubechies wavelet with smallest support; i.e., it has $q = 1$.

Algorithm *DaubGreedy*(X, B, p, q)
(∗ X is a grayscale image (intensity matrix) ∗)
Perform a 2D wavelet transform of X using the Daubechies D_q wavelet
Let w be the wavelet coefficients of the transform
$p' \leftarrow p/(p-1)$
$y_i \leftarrow |w_i| / \|\psi_i\|_{p'}$
Let $\mathcal{I}$ be the indices of the B largest y_i's
$w_i \leftarrow 0$ if $i \notin \mathcal{I}$
Perform a 2D inverse wavelet transform on the resulting w
Let X' be the resulting image representation
return X'

Figure 2: Pseudocode of the greedy algorithm's implementation.

The first example illustrates a use of the ℓ_∞ measure for sparse representation using wavelets. Minimizing the maximum error at any point in the reconstructed image implies we should retain the wavelet coefficients that correspond to sharp changes in intensity; i.e., the coefficients that correspond to the "details" in the image. The image we used, shown in Figure 3(a), is composed of a gradient background and both Japanese and English texts[2]. The number of non-zero wavelet coefficients in the original image is 65524. We set $B = 3840$ and ran Algorithm *DaubGreedy* with $p = 1, 2$ and ∞ under the Haar wavelet (with $q = 1$). When $p = 2$, the algorithm outputs the optimal B-term representation that minimizes the ℓ_2 error measure. That is, the algorithm simply retains the largest B wavelet coefficients (since $p' = 2$ and $\|\psi_i\|_{p'} = 1$ for all i). When $p = 1$, or $p = \infty$, the algorithm outputs a $O(\log n)$-approximate B-term representation as will be explained in Section 2.2. The results are shown in Figure 3. Notice that the ℓ_∞ representation essentially ignores the gradient in the background, and it retains the wavelet coefficients that correspond to the text in the image. The ℓ_1 representation also does better than the ℓ_2 representation in terms of rendering the Japanese text; however, the English translation in the former is not

[1] For compatibility with our version of MATLAB, slight modifications on the toolbox were performed. The toolbox can be obtained from `http://www.gts.tsc.uvigo.es/~wavelets/`.

[2] The Japanese text is poem number 89 of the *Kokinshu* anthology [21]. The translation is by Helen Craig McCullough.

as clear. The attribution in the ℓ_2 representation, on the other hand, is completely lost. Although the differences between the three representations are not stark, this example shows that under such high compression ratios using the ℓ_∞ norm is more suitable for capturing signal details than other norms.

The second example illustrates a use of the ℓ_1 error measure. Since the ℓ_1 norm is robust in the sense that it is indifferent to outliers, the allocation of wavelet coefficients when minimizing the ℓ_1 norm will be less sensitive to large changes in intensity than the allocation under the ℓ_2 norm. In other words, it implies that under the ℓ_1 norm the wavelet coefficients will be allocated more evenly across the image. The image we used, shown in Figure 4(a), is a framed black and white matte photograph. The number of non-zero wavelet coefficients in the original image is 65536. We set $B = 4096$ and ran Algorithm *DaubGreedy* with $p = 1, 2$ and ∞ under the Daubechies D_2 wavelet. The results are shown in Figure 4. Notice that the face of the subject is rendered in the ℓ_1 representation more "smoothly" than in the ℓ_2 representation. Further, the subject's mouth is not portrayed completely in the ℓ_2 representation. As explained earlier, these differences between the two representations are due to the fact that the ℓ_1 norm is not as affected as the ℓ_2 norm by other conspicuous details in the image; e.g., the frame. The ℓ_∞ representation, on the other hand, focuses on the details of the image displaying parts of the frame and the eyes well, but misses the rest of the subject entirely. This example foregrounds some advantages of the ℓ_1 norm over the customary ℓ_2 norm for compressing images.

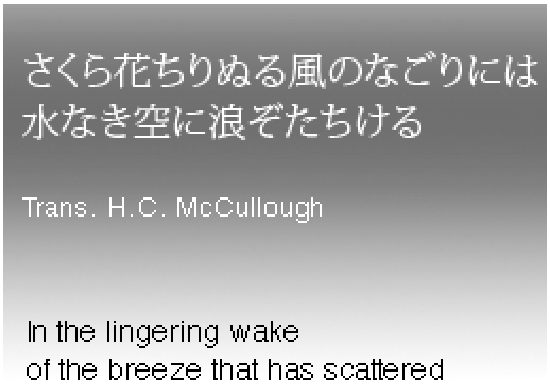
The last example highlights the advantage of representing an image sparsely using a nonlinear wavelet approximation versus using a rank-k approximation of the image. Recall that if X is our image then the best rank-k approximation is given by $U_k \Sigma_k V_k^T$ where $X = U\Sigma V^T$ is the SVD decomposition of X, and U_k is comprised of the k singular vectors corresponding to the largest k singular values of X (see, e.g., [13]). The original image is shown in Figure 5(a)[3] and the number of non-zero coefficients in its Haar wavelet expansion is 65536. Figure 5(c) shows the best rank-12 approximation of the image; i.e., it displays $X_{12} = U_{12}\Sigma_{12}V_{12}^T$. This representation stores 6144 values corresponding to the number of elements in $U_{12}\Sigma_{12}$ plus V_{12}. We set $B = 3072$ and ran Algorithm *DaubGreedy* with $p = 1, 2$ under the Haar wavelet (Figures 5(d) and 5(b)). (The B-term representation problem implicitly requires storing $2B$ numbers: the B *values* of the solution components that we compute, and the B indices of these components.) It is clear that the nonlinear approximations offer perceptually better representations that the approximation offered by the SVD. Also, as in the previous example, the ℓ_1 representation is again "smoother" than the ℓ_2 with less visible artifacts.

2.2 Analysis of the Approximation Algorithm

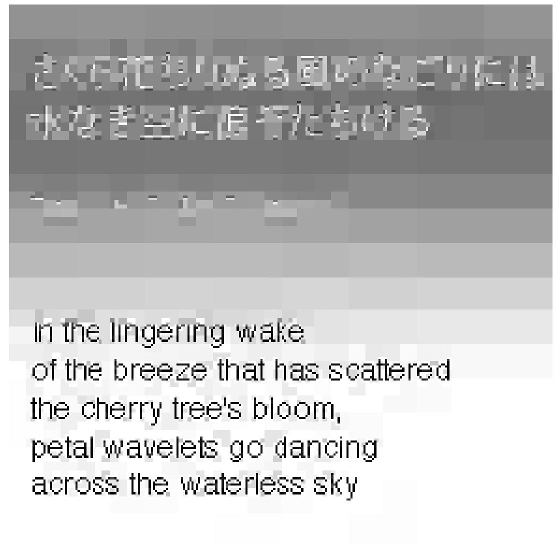
The results in this section appear in [16] (see also [14]).

Recall our optimization problem: Given a compactly-supported wavelet basis $\{\psi_i\}$ and a target vector f, we wish to find $\{z_i\}$ with at most B non-zero numbers to minimize $\|f - \sum_i z_i\psi_i\|_p$.

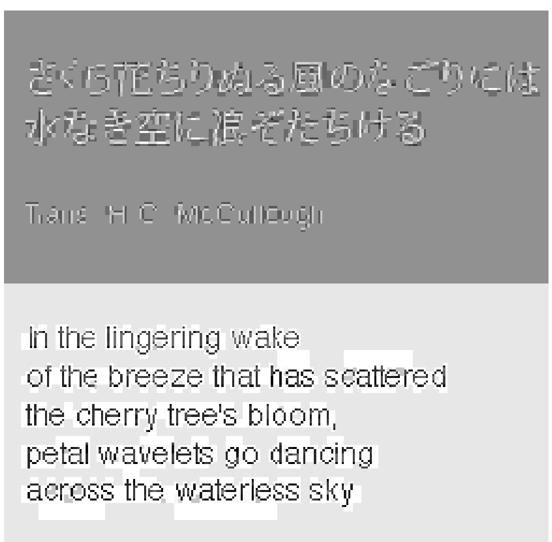
[3] The image is taken from a water painting by Shozo Matsuhashi. It is untitled.



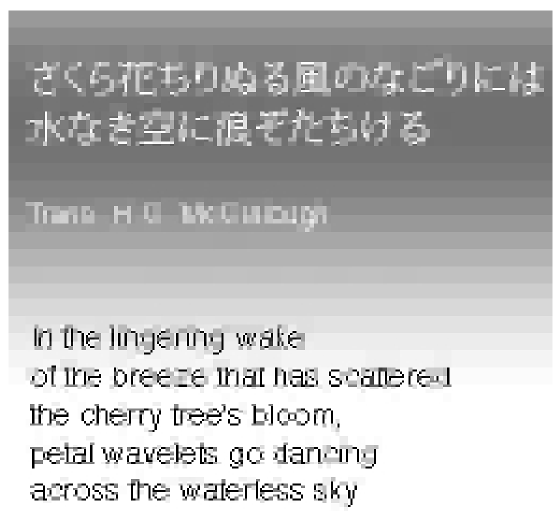
(a) The original image



(b) Output of the optimal ℓ_2 algorithm (which retains the largest B wavelet coefficients)



(c) Output of our greedy algorithm under ℓ_∞

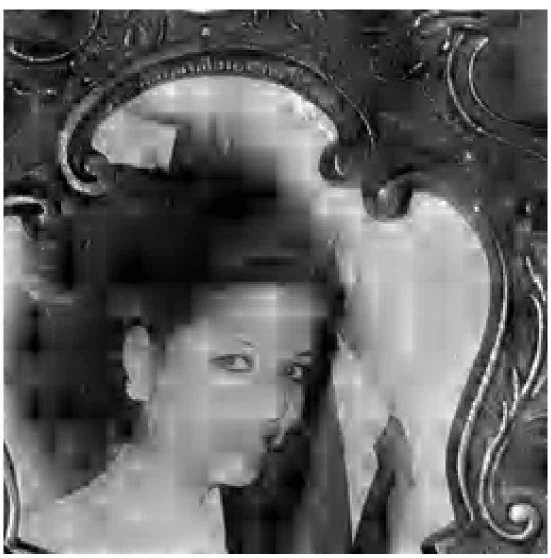


(d) Output of our greedy algorithm under ℓ_1

Figure 3: Representing an image with embedded text using the optimal strategy that minimizes the ℓ_2 error, and our greedy approximation algorithm under the ℓ_∞ and ℓ_1 error measures. The Haar wavelet is used in all three representations, and the number of retained coefficients is $B = 3840$.

(a) The original image

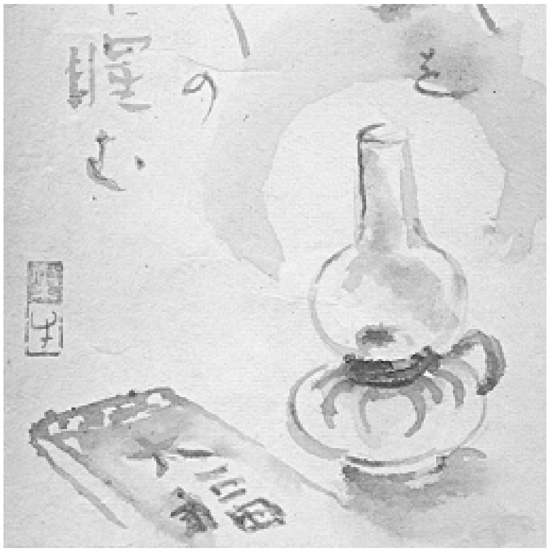
(b) Output of the optimal ℓ_2 algorithm (which retains the largest B wavelet coefficients)



(c) Output of our greedy algorithm under ℓ_∞

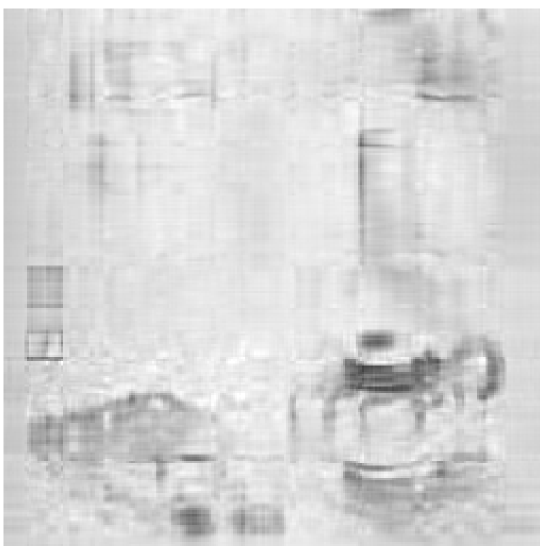
(d) Output of our greedy algorithm under ℓ_1

Figure 4: Representing an image using the optimal strategy that minimizes the ℓ_2 error, and our greedy approximation algorithm under the ℓ_∞ and ℓ_1 error measures. The Daubechies D_2 wavelet is used in all three representations, and the number of retained coefficients is $B = 4096$.

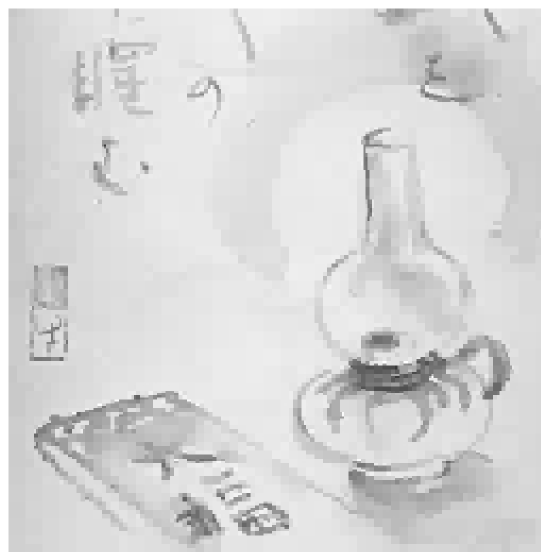


(a) The original image

(b) Output of the optimal ℓ_2 algorithm (which retains the largest B wavelet coefficients)



(c) Output of the best rank-12 approximation



(d) Output of our greedy algorithm under ℓ_1

Figure 5: Representing an image using the optimal strategy that minimizes the ℓ_2 error and using our greedy approximation algorithm under the ℓ_1 error measure versus its best rank-k approximation. Here $k = 12$, and the number of values stored in all three representations is 6144. The Haar wavelet is used in the two nonlinear representations (the number of retained wavelet coefficients is $B = 3072$).

We present two analyses below corresponding to ℓ_∞ and ℓ_p errors when $p \in [1, \infty)$. In each case we begin by analyzing the sufficient conditions that guarantee the error. A (modified) greedy coefficient retention algorithm will naturally fall out of both analyses. The proof shows that several of the algorithms that are used in practice have bounded approximation guarantee. Note that the optimum solution can choose any values in the representation $\hat{f}$.

In what follows the pair (p, p') are the usual conjugates; i.e., $\frac{1}{p} + \frac{1}{p'} = 1$ when $1 < p < \infty$, and when $p = 1$ we simply set $p' = \infty$. For simplicity, we start with the $p = \infty$ case.

2.2.1 An ℓ_∞ Algorithm and Analysis

The main lemma, which gives us a lower bound on the optimal error, is:

Lemma 2.1. *Let $\mathcal{E}$ be the minimum error under the ℓ_∞ norm and $\{z_i^*\}$ be the optimal solution, then*

$$-\|\psi_i\|_1 |\mathcal{E}| \leq \langle f, \psi_i \rangle - z_i^* \leq \|\psi_i\|_1 |\mathcal{E}| \ .$$

Proof. For all j we have $-|\mathcal{E}| \leq f(j) - \sum_i z_i^* \psi_i(j) \leq |\mathcal{E}|$. Since the equation is symmetric multiplying it by $\psi_k(j)$ we get,

$$-|\mathcal{E}||\psi_k(j)| \ \leq f(j)\psi_k(j) - \psi_k(j) \sum_i z_i^* \psi_i(j) \ \leq |\mathcal{E}||\psi_k(j)|$$

If we add the above equation for all j, since $-|\mathcal{E}| \sum_j |\psi_k(j)| = -|\mathcal{E}| \|\psi_k\|_1$ we obtain (consider only the left side)

$$\begin{aligned} -|\mathcal{E}| \|\psi_k\|_1 &\leq \sum_j f(j)\psi_k(j) - \sum_j \psi_k(j) \sum_i z_i^* \psi_i(j) \\ &= \langle f, \psi_k \rangle - \sum_i z_i^* \sum_j \psi_k(j)\psi_i(j) \\ &= \langle f, \psi_k \rangle - \sum_i z_i^* \delta_{ik} = \langle f, \psi_k \rangle - z_k^* \ . \end{aligned}$$

The upper bound follows analogously. □

A Relaxation. Consider the following program:

$$\begin{array}{rcl} & \text{minimize } \tau & \\ -\tau\|\psi_1\|_1 & \leq \langle f, \psi_1 \rangle - z_1 & \leq \tau \|\psi_1\|_1 \\ \vdots & \vdots & \vdots \\ -\tau\|\psi_n\|_1 & \leq \langle f, \psi_n \rangle - z_n & \leq \tau \|\psi_n\|_1 \\ & \text{At most } B \text{ of the } z_i\text{'s are non-zero} & \end{array} \tag{2}$$

Observe that $\mathcal{E}$ is a feasible solution for the above program and $\mathcal{E} \geq \tau^*$ where τ^* is the optimum value of the program. Also, Lemma 2.1 is not specific to wavelet bases, and indeed we have $\mathcal{E} = \tau^*$ when $\{\psi_i\}$ is the standard basis, i.e. ψ_i is the vector with 1 in the i^{th} coordinate and 0 elsewhere. The next lemma is straightforward.

Lemma 2.2. *The minimum τ of program (2) is the $(B+1)^{th}$ largest value $\frac{|\langle f, \psi_i \rangle|}{\|\psi_i\|_1}$.*

The Algorithm. We choose the largest B coefficients based on $|\langle f, \psi_i\rangle|/\|\psi_i\|_1$. This can be done over a one pass stream, and in $O(B + \log n)$ space for any compact wavelet basis. Note that we need not choose $z_i = \langle f, \psi_i\rangle$ but any z_i such that $|z_i - \langle f, \psi_i\rangle|/\|\psi_i\|_1 \le \tau^*$. But in particular, we may choose to retain coefficients and set $z_i = \langle f, \psi_i\rangle$. The alternate choices may (and often will) be better. Also note that the above is only a necessary condition; we *still* need to analyze the guarantee provided by the algorithm.

Lemma 2.3. *For all basis vectors ψ_i of a compact system there exists a constant C s.t. $\|\psi_i\|_p \|\psi_i\|_{p'} \le \sqrt{q}C$.*

Proof. Suppose first that $p < 2$. Consider a basis vector $\psi_i[] = \psi_{j,s}[]$ of sufficiently large scale j that has converged to within a constant r (point-wise) of its continuous analog $\psi_{j,s}()$ [19, pp. 264–5]. That is, $|\psi_{j,s}[k] - \psi_{j,s}(k)| \le r$ for all k such that $\psi_{j,s}[k] \ne 0$. The continuous function $\psi_{j,s}()$ is given by $\psi_{j,s}(t) = 2^{-j/2}\psi(2^{-j}t - s)$, which implies $\psi_{j,s}[k] = O\left(2^{-j/2}\psi(2^{-j}k - s)\right) = O(2^{-j/2})$. Note that we are assuming $\|\psi\|_\infty$ itself is some constant since it is independent of n and B. Combining the above with the fact that $\psi_{j,s}[]$ has at most $(2q)2^j$ non-zero coefficients, we have $\|\psi_{j,s}\|_{p'} = O(2^{-j/2}((2q)2^j)^{1/p'}) = O(2^{j(\frac{1}{p'}-\frac{1}{2})}(2q)^{\frac{1}{p'}})$.

Now by Hölder's inequality, $\|\psi_{j,s}\|_p \le ((2q)2^j)^{\frac{1}{p}-\frac{1}{2}} \|\psi_{j,s}\|_2 = 2^{j(\frac{1}{p}-\frac{1}{2})}(2q)^{\frac{1}{p}-\frac{1}{2}}$. Therefore, for sufficiently large scales j, $\|\psi_{j,s}\|_p \|\psi_{j,s}\|_{p'} = O(2^{j(\frac{1}{p}+\frac{1}{p'}-1)}(2q)^{\frac{1}{p}+\frac{1}{p'}-\frac{1}{2}}) = O(\sqrt{q})$, and the lemma holds. For basis vectors at smaller (constant) scales, since the number of non-zero entries is constant, the ℓ_p norm and the $\ell_{p'}$ norm are both constant.

Finally, for $p > 2$, the argument holds by symmetry. □

For the proof of our theorem, we will also use the following proposition which is a consequence of the dyadic structure of compactly-supported wavelet bases.

Proposition 2.4. *A compactly-supported wavelet whose filter has $2q$ non-zero coefficients generates a basis for $\mathbb{R}^n$ that has $O(q \log n)$ basis vectors with a non-zero value at any point $i \in [n]$.*

Theorem 2.5. *The ℓ_∞ error of the final approximation is at most $O(q^{3/2} \log n)$ times $\mathcal{E}$ for any compactly supported wavelet.*

Proof. Let $\{z_i\}$ be the solution of the system (2), and let the set of the inner products chosen be $\mathcal{S}$. Let τ^* is the minimum solution of the system (2). The ℓ_∞ error seen at a point j is $|\sum_{i \notin \mathcal{S}} \langle f, \psi_i\rangle \psi_i(j)| \le \sum_{i \notin \mathcal{S}} |\langle f, \psi_i\rangle||\psi_i(j)|$. By Lemma 2.2, this sum is at most $\sum_{i \notin \mathcal{S}} \tau^* \|\psi_i\|_1 |\psi_i(j)|$, which is at most $\tau^* \max_{i \notin \mathcal{S}} \|\psi_i\|_1 \|\psi_i\|_\infty$ times the number of vectors that are non-zero at j. By Proposition 2.4 the number of non-zero vectors at j is $O(q \log n)$. By Lemma 2.3, $\|\psi_i\|_1 \|\psi_i\|_\infty \le \sqrt{q}C$ for all i, and since $\tau^* \le \mathcal{E}$ we have that the ℓ_∞ error is bounded by $O(q^{3/2} \log n)\mathcal{E}$. □

2.2.2 An ℓ_p Algorithm and Analysis for $p \in [1, \infty)$

Under the ℓ_p norm, a slight modification to the algorithm above also gives an $O(q^{3/2} \log n)$ approximation guarantee.

Lemma 2.6. *Let $\mathcal{E}$ be the minimum error under the ℓ_p norm and $\{z_i^*\}$ be the optimal solution, then for some constant c_0,*

$$\left(\sum_k \frac{1}{\|\psi_k\|_{p'}^p}|\langle f,\psi_k\rangle - z_k^*|^p\right)^{\frac{1}{p}} \leq (c_0 q\log n)^{\frac{1}{p}}\,\mathcal{E} \quad .$$

Proof. An argument similar to that of Lemma 2.1 gives

$$\begin{aligned}
\sum_i \left| f_i\psi_k(i) - \sum\nolimits_j z_j^*\psi_j(i)\psi_k(i)\right| &= \sum_i \xi_i|\psi_k(i)| \leq \left(\sum_{i\in \text{ support of } \psi_k} \xi_i^p\right)^{1/p} \|\psi_k\|_{p'} \\
\Rightarrow \qquad \frac{1}{\|\psi_k\|_{p'}^p}|\langle f,\psi_k\rangle - z_k^*|^p &\leq \sum_{i\in \text{ support of } \psi_k} \xi_i^p \\
\Rightarrow \qquad \sum_k \frac{1}{\|\psi_k\|_{p'}^p}|\langle f,\psi_k\rangle - z_k^*|^p &\leq c_0 q\log n \sum_i \xi_i^p \quad ,
\end{aligned}$$

where the last inequality follows from Proposition 2.4, that each i belongs to $O(q\log n)$ basis vectors (c_0 is the constant hidden by the this O-term). □

A Relaxation. Consider the following system of equations,

$$\begin{aligned}
&\text{minimize } \tau \\
&\left(\sum_{i=1}^n \frac{|\langle f,\psi_i\rangle - z_i|^p}{\|\psi_i\|_{p'}^p}\right)^{\frac{1}{p}} \leq (c_0 q\log n)^{\frac{1}{p}}\tau \\
&\text{At most } B \text{ of the } z_i\text{'s are non-zero}
\end{aligned} \tag{3}$$

The Algorithm. We choose the largest B coefficients based on $|\langle f,\psi_k\rangle|/\|\psi_k\|_{p'}$, which minimizes the system (3). This computation can be done over a one pass stream, and in $O(B+\log n)$ space.

Theorem 2.7. *Choosing the B coefficients $\langle f,\psi_k\rangle$ that are largest based on the ordering $|\langle f,\psi_k\rangle|/\|\psi_k\|_{p'}$ is a streaming $O(q^{3/2}\log n)$ approximation algorithm for the unrestricted optimization problem under the ℓ_p norm.*

Note this matches the ℓ_∞ bounds, but stores a (possibly) different set of coefficients.

Proof. Let the value of the minimum solution to the above system of equations (3) be τ^*. Since $\{z_i^*\}$ is feasible for system (3), $\tau^* \leq \mathcal{E}$. Assume $\mathcal{S}$ is the set of coefficients chosen, the resulting error $\mathcal{E}_{\mathcal{S}}$ is,

$$\begin{aligned}
\mathcal{E}_{\mathcal{S}}^p &= \sum_i \left| \sum_{k\notin\mathcal{S}} \langle f, \psi_k\rangle \psi_k(i) \right|^p \leq \sum_i (c_0 q \log n)^{p-1} \sum_{k\notin\mathcal{S}} |\langle f, \psi_k\rangle|^p |\psi_k(i)|^p \\
&= (c_0 q \log n)^{p-1} \sum_{k\notin\mathcal{S}} |\langle f, \psi_k\rangle|^p \, \|\psi_k\|_p^p \\
&\leq (c_0 q \log n)^{p-1} \sum_{k\notin\mathcal{S}} \frac{C^p q^{\frac{p}{2}}}{\|\psi_k\|_{p'}^p} |\langle f, \psi_k\rangle|^p \\
&= C^p q^{\frac{p}{2}} (\tau^* c_0 q \log n)^p \ .
\end{aligned}$$

Here, the first inequality is Hölder's inequality combined with Proposition 2.4 and the fact that $p/p' = p-1$; the second inequality follows from Lemma 2.3; and the final equality follows from the optimality of our choice of coefficients for the system (3). Now since $\tau^* \leq \mathcal{E}$, we have that $\mathcal{E}_{\mathcal{S}} \leq c_0 C q^{\frac{3}{2}} \mathcal{E} \log n$. □

2.2.3 Summary and Examples

In the two preceeding subsections we showed the following:

Theorem 2.8. *Let $\frac{1}{p} + \frac{1}{p'} = 1$. Choosing the largest B coefficients based on the ordering $|\langle f, \psi_i\rangle|/\|\psi_i\|_{p'}$, which is possible by a streaming $O(B + \log n)$ algorithm, gives a $O(q^{\frac{3}{2}} \log n)$ approximation algorithm for the unrestricted optimization problem (Problem 1.1) under the given ℓ_p norm. The argument naturally extends to multiple dimensions.*

As is well-known, this choice of coefficients is optimal when $p = 2$ (since $p' = 2$ and $\|\psi_i\|_2 = 1$).

Note that the above theorem bounds the gap between the restricted (where we can only choose wavelet coefficients of the input in the representation) and unrestricted optimizations.

A tight example for the ℓ_∞ measure. Suppose we are given the Haar basis $\{\psi_i\}$ and the vector f with the top coefficient $\langle f, \psi_1\rangle = 0$ and with $\langle f, \psi_i\rangle / \|\psi_i\|_1 = 1 - \epsilon$ for $i \leq n/2$, and $\langle f, \psi_i\rangle / \|\psi_i\|_1 = 1$ for $i > n/2$ (where ψ_i, $i > n/2$, are the basis with smallest support). Let $B = n/c - 1$ where $c \geq 2$ is a constant that is a power of 2. The optimal solution can choose the B coefficients which are in the top $\log n - \log c$ levels resulting in an error bounded by $\log c$. The ℓ_∞ error of the greedy strategy on the other hand will be at least $\log n - 1$ because it will store coefficients only at the bottom of the tree. Hence it's error is at least $\log n / \log c - o(1)$ of the optimal.

2.3 A Universal Representation

In this section, we present a strategy that stores $B(\log n)^2$ coefficients and simultaneously approximates the optimal representations for all p-norms[4]. Notice that in Problem 1.1 we

[4]The results presented in this section have not appeared elsewhere.

know the p-norm we are trying to approximate. Here, we do *not* know p and we wish to come up with a representation such that for all $p \in [1, \infty]$, its error measured with $\|f - \hat{f}_u\|_p$ is $O(\log n)$ times the optimal error $\min_x \|f - \Psi x\|_p$ where x has at most B non-zero components. Notice that we allow our universal representation to store a factor $(\log n)^2$ more components than any one optimal representation; however, it has to approximate all of them concurrently.

We run our algorithm as before computing the wavelet coefficients of the target vector f; however, we need to determine which coefficients to store for our universal representation. To this end, define the set:

$$\mathcal{N} = \{p_t : p_t = 1 + \frac{t}{\log n},\ t = 0, \ldots, \log n(\log n - 1)\} \ . \tag{4}$$

For every $p_t \in \mathcal{N}$, we will store the B coefficients that are largest based on the ordering $|\langle f, \psi_k \rangle| / \|\psi_k\|_{p'_t}$ where p'_t is the dual norm to p_t. Hence, the number of coefficients we store is no more than $B(\log n)^2$ since $|\mathcal{N}| = (\log n)^2$. Note that our dual programs show that for a given p, storing more than B coefficients does not increase the error of the representation. Now let $\hat{f}_u$ be our resultant representation; i.e., if $\mathcal{S}$ contains the coefficients we chose, then $\hat{f}_u = \sum_{i \in \mathcal{S}} \langle f, \psi_i \rangle \psi_i$; and let $f^*_{(p)}$ be the optimal representation under the norm ℓ_p. Consider first the case when $p \in (p_t,\ p_{t+1})$ where $p_t, p_{t+1} \in \mathcal{N}$.

$$\begin{aligned}
\|f - \hat{f}_u\|_p &\le \|f - \hat{f}_u\|_{p_t} && \text{since } p > p_t \\
&\le c q^{\frac{3}{2}} (\log n) \|f - f^*_{(p_t)}\|_{p_t} && \text{by Thereom 2.8} \\
&\le c q^{\frac{3}{2}} (\log n) \|f - f^*_{(p)}\|_{p_t} && \text{by the optimality of } f^*_{p_t} \text{ for } \ell_{p_t} \\
&\le c q^{\frac{3}{2}} (\log n) n^{\frac{1}{p_t} - \frac{1}{p}} \|f - f^*_{(p)}\|_p && \text{by Hölder's inequality}
\end{aligned} \tag{5}$$

However $1/p_t - 1/p \le 1/p_t - 1/p_{t+1}$ since $p < p_{t+1}$; and by their definition,

$$\frac{1}{p_t} - \frac{1}{p_{t+1}} = \frac{\log n}{(\log n + t)(\log n + t + 1)} \le \frac{1}{\log n} \ .$$

Hence, $n^{\frac{1}{p_t} - \frac{1}{p}} \le n^{1/(\log n)} = 2$; and from expression (5) we have that $\|f - \hat{f}_u\|_p = O(q^{\frac{3}{2}} \log n) \|f - f^*_{(p)}\|_p$ as required. When $p > p_t$ for $t = \log n(\log n - 1)$, we immediately have $n^{\frac{1}{p_t} - \frac{1}{p}} \le n^{1/(\log n)}$ and the result follows.

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] Kaushik Chakrabarti, Minos Garofalakis, Rajeev Rastogi, and Kyuseok Shim. Approximate query processing using wavelets. *The VLDB Journal*, 10(2-3):199–223, 2001.

[3] A. Cohen, I. Daubechies, O. Guleryuz, and M. Orchard. On the importance of combining wavelet-based non-linear approximation in coding strategies. *IEEE Transactions on Information Theory*, 48(7):1895–1921, 2002.

[4] Ronald R. Coifman and M. Victor Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

[5] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure. Appl. Math.*, 41:909–996, 1988.

[6] Ingrid Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[7] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximation. *Journal of Constructive Approximation*, 13:57–98, 1997.

[8] R. DeVore. Nonlinear approximation. *Acta Numerica*, pages 1–99, 1998.

[9] R. DeVore, B. Jawerth, and V. A. Popov. Compression of wavelet decompositions. *Amer. J. Math.*, 114:737–785, 1992.

[10] D. Donoho and I. Johnstone. Ideal space adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[11] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *J. Royal Statistical Soc., Ser. B*, 57:301–369, 1996.

[12] Anna C. Gilbert, S. Muthukrishnan, and Martin Strauss. Approximation of functions over redundant dictionaries using coherence. *Proc. of SODA*, pages 243–252, 2003.

[13] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1989.

[14] Sudipto Guha and Boulos Harb. Approximation algorithms for wavelet transform coding of data streams. In *SODA '06: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 698–707, New York, NY, USA, 2006. ACM Press.

[15] A. Haar. Zur Theorie der orthogonalen Funktionensysteme. *Math. Ann.*, 69:331–371, 1910.

[16] Boulos Harb. *Algorithms for linear and nonlinear approximation of large data*. PhD thesis, University of Pennsylvania, 2007.

[17] C. E. Jacobs, A. Finkelstein, and D. H. Salesin. Fast multiresolution image querying. *Computer Graphics*, 29(Annual Conference Series):277–286, 1995.

[18] S. Mallat. Multiresolution approximations and wavelet orthonormal bases of $l^2(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 315:69–87, September 1989.

[19] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1999.

[20] Y. Matias, J. Scott Vitter, and M. Wang. Wavelet-Based Histograms for Selectivity Estimation. *Proc. of ACM SIGMOD*, 1998.

[21] H.C. McCullough. *Kokin Wakashu: The First Imperial Anthology of Japanese Poetry*. Stanford University Press, Palo Alto, 1984. Translated from Japanese.

[22] Y. Meyer. *Wavelets and operators*. Advanced mathematics. Cambridge University Press, 1992.

[23] Apostol Natsev, Rajeev Rastogi, and Kyuseok Shim. Walrus: a similarity retrieval algorithm for image databases. In *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pages 395–406, New York, NY, USA, 1999. ACM Press.

[24] Santiago Gonzalez Sanchez, Nuria Gonzalez Prelcic, and Sergio J. Garca Galan. Uvi_Wave version 3.0—Wavelet toolbox for use with MATLAB.

[25] E. Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. *Mathematische Annalen*, 63(4):433–476, 1907.

[26] Eric J. Stollnitz, Tony D. Derose, and David H. Salesin. *Wavelets for computer graphics: theory and applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996.

[27] V. N. Temlyakov. Nonlinear methods of approximation. *Foundations of Computational Mathematics*, 3:33–107, 2003.

[28] Jeffrey Scott Vitter and Min Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. In *SIGMOD '99: Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, pages 193–204, New York, NY, USA, 1999. ACM Press.

[29] Jeffrey Scott Vitter, Min Wang, and Bala Iyer. Data cube approximation and histograms via wavelets. In *CIKM '98: Proceedings of the seventh international conference on Information and knowledge management*, pages 96–104, New York, NY, USA, 1998. ACM Press.