Nonlinear approximation

Function / signal $f \in L^2$

Orthonormal basis $\beta = \{b_i\}_{i=1}^{\infty}$

Let $\{\gamma_k\}$ = "rearrangement" of $\{|\langle f, b_i \rangle|\}$

$\gamma_k = k^{\text{th}}$ largest $|\langle f, b_i \rangle|$

f_N = Best N-term approximation in β

$$= \sum_{i \in I_N} \langle f, b_i \rangle b_i$$

where $I_N = \{\text{indices of } N \text{ largest } |\langle f, b_i \rangle|\}$

ε_N = error of f_N

$$= \|f - f_N\|_2$$

$$= \left(\sum_{K=N+1}^{\infty} \gamma_K^2 \right)^{1/2}$$

not squared like last week

✳ Theorem: $\varepsilon_N \sim N^{-s} \iff \gamma_K \sim K^{-s-1/2}$

Links the decay rate (or sparsity) of the expansion coefficients to the decay rate of the error.

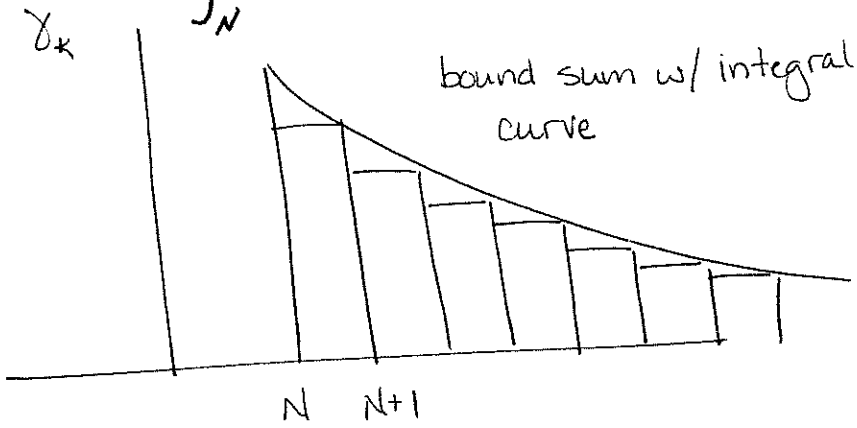
Pf:

$\Leftarrow$

$\gamma_k \sim k^{-s-1/2} \Rightarrow \exists\, C>0$ s.t. $\gamma_k \le C k^{-s-1/2}$

$$\Rightarrow \varepsilon_N^2 \le C \sum_{k=N+1}^{\infty} k^{-2s-1} \le C \int_N^{\infty} x^{-2s-1}\, dx = C' N^{-2s}$$

$$\Rightarrow \varepsilon_N \sim N^{-s}$$



$\Rightarrow$

$\varepsilon_N \sim N^{-s}$,

We have

$$\gamma_{2N}^2 \le \frac{1}{N} \sum_{m=N+1}^{2N} \gamma_m^2 \qquad \text{since } \gamma_k \text{ is monotonically decreasing}$$

$$\le \frac{1}{N} \varepsilon_N^2$$

$$\le C N^{-2s-1}$$

$$\Rightarrow \gamma_k \sim k^{-s-1/2}$$

■

$\varepsilon_N \le C$

$a_n \sim n^{-\alpha}$ means $\exists\, B>0$ s.t. $a_n \le B n^{-\alpha}$ $\forall n$

ℓ^p

We can classify functions by the ℓ^p norms of their expansion coefficients using a basis $\beta = \{b_i\}_{i=1}^{\infty}$

$$\|f\|_{\beta,p} = \left(\sum_{i=1}^{\infty} |\langle f, b_i \rangle|^p \right)^{1/p}$$

for $p < 1$, not called "norm" (triangle inequality doesn't always hold) "quasinorm"

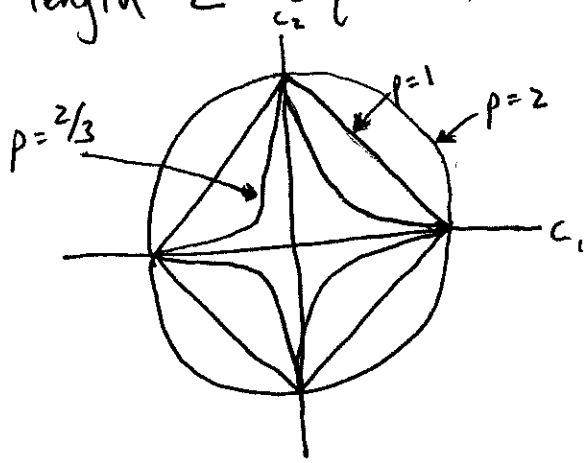
Define the space

$$B_{\beta,p} = \{ f \in L^2 : \|f\|_{\beta,p} < \infty \}$$

For $p = 2$, $\|f\|_{\beta,p} = \|f\|_2$, $B_{\beta,2} = L^2$

For $p < 2$, $\|f\|_{\beta,p}$ gives us another notion of sparsity

ex: For a length 2 sequence, look at $\|f\|_{\beta,p} \leq 1$



$\|\{c_1, c_2\}\|_2 = \sqrt{c_1^2 + c_2^2} = c$

$\|\{c_1, c_2\}\|_1 = |c_1 + c_2|$

as $p \to 0$, curve approaches origin

As p gets small, the sequences cluster around one of the axes.

Functions in $B_{\beta,p}$ with $p < 2$ have expansion coefficients that drop off rapidly.

Theorem:

$$f \in B_{\beta,p} \Rightarrow \gamma_k \sim k^{-1/p} \qquad (*)$$

pf.

$$\|f\|_{\beta,p}^p = \sum_{i=1}^{\infty} \gamma_i^p \geq \sum_{i=1}^{k} \gamma_i^p \geq k \gamma_k^p$$

$$\Rightarrow \gamma_k \leq \|f\|_{\beta,p} \, k^{-1/p}$$

$$|a+b|^\alpha \leq 2^\alpha (|a|^\alpha + |b|^\alpha)$$

$$\Rightarrow \gamma_k \sim k^{-1/p}$$

Side note:

$$f \in B_{\beta,p} \Rightarrow \{\gamma_k\} \in \ell^p$$

In fact, $\{\gamma_k\} \in \ell^p$ is a slightly stronger condition than (*) above. The space of sequences satisfying (*) above is called "weak-ℓ^p"

$$\ell^p \subset \text{weak-}\ell^p$$

$$\ell^q \not\subset \text{weak-}\ell^p \quad \forall\, q > p$$

$$\{\gamma_k\} \in \ell^p \Rightarrow \varepsilon_N \sim N^{-1/p + 1/2}$$

We can relate our previous notion of smoothness (Lipschitz regularity) to ℓ^p spaces of wavelet coefficients.

Let

$\Psi = \left[\{\phi_{0,k}\}, \{\psi_{j,k}\} \right]$ wavelet basis

ψ has q vanishing moments, compactly supported

We call $B_p := B_{\Psi,p}$ a <u>Besov space</u>

<u>Theorem:</u>

If $f \in L^2[0,1]$ is uniformly Lipschitz $\alpha \leq q$, then

$f \in B_p$ for $\frac{1}{p} < \alpha + 1/2$ $\left(p > \frac{1}{\alpha+1/2}\right)$ as $\alpha \uparrow$, $p \downarrow$ (need fewer coeffs for smoother func)

<u>pf:</u>

From before, $\exists$ A s.t. $|\langle f, \psi_{j,k} \rangle| \leq A 2^{-j(\alpha+1/2)}$

$$\|f\|_{B_p}^p = \sum_{j=0}^{\infty} \sum_{k} |\langle f, \psi_{j,k} \rangle|^p$$

$$\leq A^p \sum_{j=0}^{\infty} 2^j 2^{-pj(\alpha+1/2)}$$

$$= A^p \sum_{j=0}^{\infty} 2^{j(1-p(\alpha+1/2))}$$

$$< \infty \quad \text{if} \quad p(\alpha+1/2) > 1.$$

So, for uniformly Lip $\alpha \leq q$ functions,

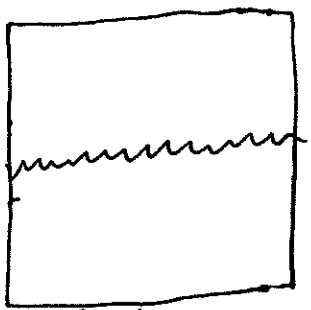
$$\varepsilon_N \sim N^{-\alpha}$$

Using a wavelet basis with q vanishing moments

What happens when we add discontinuities?

<u>Theorem:</u> If $f \in L^2[0,1]$ has a finite number K of discontinuities and is uniformly Lip $\alpha < q$ between these discontinuities, Then $f \in B_p$ for $\frac{1}{p} < \alpha + \frac{1}{2}$

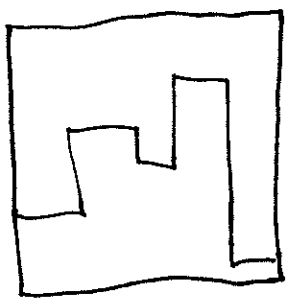
Outline of proof:



Unif. Lip α

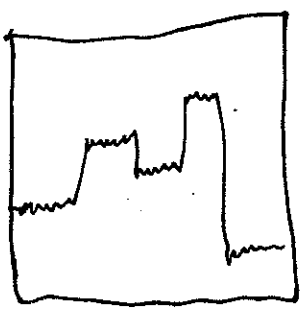
$\longrightarrow$ has wavelet coeffs. in ℓ^p $\frac{1}{p} < \alpha + 1/2$

$+$



$\longrightarrow$ has wavelet coeffs in ℓ^p $\forall p > 0$

$\Rightarrow$



$\longrightarrow$ has wavelet coeffs in ℓ^p, $\frac{1}{p} < \alpha + 1/2$

since ℓ^p is a linear space.

pf:

Divide the wavelet coefficients into two types

I. those with basis functions whose support does not include a discontinuity ("touching")

II. those with basis functions whose support includes a discontinuity

I. $\{\psi_{j,k}\}_{\{j,k\}\in I}$ represent uniformly Lip α regions,

so $\sum_{\{j,k\}\in I} |\langle f, \psi_{j,k}\rangle|^p < \infty \qquad \frac{1}{p} < \alpha + \frac{1}{2}$

II.

The $\psi_{j,k}$ have compact support, so $\exists\ C > 0$ such that there are at most C wavelets at each scale whose support includes a specific abcissa.

$\Rightarrow \leq CK$ wavelets of type II at each scale j

From last time, we know $\exists\ A$ s.t.

$$|\langle f, \psi_{j,k}\rangle| \leq A 2^{-j/2} \qquad \forall \{j,k\} \in II$$

so $$\sum_{\{j,k\}\in II} |\langle f, \psi_{j,k}\rangle|^p \leq \sum_{j=0}^{\infty} CKA^p 2^{-pj/2} < \infty \qquad \forall p > 0$$

$\Rightarrow \ell^p$ norm of the type I & type II wavelet coefficients is finite if $\frac{1}{p} < \alpha + \frac{1}{2}$

* Adding a finite number of discontinuities does not affect the approximation rate, we still have $\varepsilon_N \sim N^{-\alpha}$
(doesn't happen in Fourier domain.)

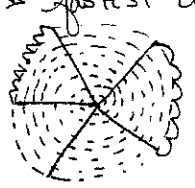
Notes:

- Real image slices have $p \sim .7$
- These results can be extended to higher dimensions BUT, the model for images does not extend

An edge in an image is a discontinuity along a curve, not at a point

It turns out that despite the success wavelets have enjoyed in image processing, they are suboptimal.

Current research tools: woohoo!
wedgelets, ridgelets, curvelets, ...



fastest decay rate as $N \to \infty$ (asymptotic)
doesn't say how low error will be for given signal

provably optimal for 1-D
(Besov spaces, piecewise Lipschitz)
MP3 - cosine transform

5-10 slides talk (10-12 min w/ few mins for q+a, 15 total)

~8

Why use wavelets ?

- The wavelet transform gives a sparse representation of smooth functions.

Implication: A signal f is well approximated using just a few wavelet coefficients

N-term nonlinear approximation

$f \in L^2$ is approximated with N basis functions selected from an orthonormal basis $\beta = \{b_n\}_{n=1}^{\infty}$

- $$f_N = \sum_{n \in I_N} \langle f, b_n \rangle b_n$$

$I_N \subset \mathbb{N}$, $|I_N| = N$

Approximation error

$$\varepsilon[N] = \|f - f_N\|^2 = \sum_{n \notin I_n} |\langle f, b_n \rangle|^2$$

For a given N, $\varepsilon[N]$ is minimized by choosing I_N corresponding to the N-largest magnitude coefficients.

- We'll show that with β = wavelet basis, $\varepsilon[N] \to 0$, $N \to \infty$, where the rate depends on the "smoothness" of f.

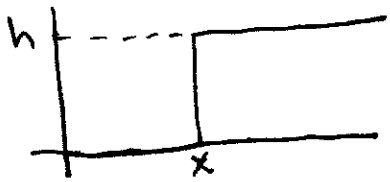
cameraman

1% of wavelet coefs

5% of wavelet coefs

Quick example of "sparsity"

- f = step edge of height h at x



$w_{j,k} = \langle f, \psi_{j,k} \rangle$ = wavelet coefficients of f

① If the support of $\psi_{j,k}$ does not include x, $w_{j,k} = 0$

$$w_{j,k} = \int f \cdot \psi_{j,k}\, dt = c \int \psi_{j,k}\, dt = 0$$

② If the support of $\psi_{j,k}$ includes x,

$$|w_{j,k}| = \left| \int f \cdot \psi_{j,k}\, dt \right| \leq \int |f| \cdot |\psi_{j,k}|\, dt$$

$$\leq |f|_\infty \int |\psi_{j,k}|\, dt = h 2^{j/2} \int |\psi(2^j t)|\, dt$$

(largest value $|f|$ takes.)

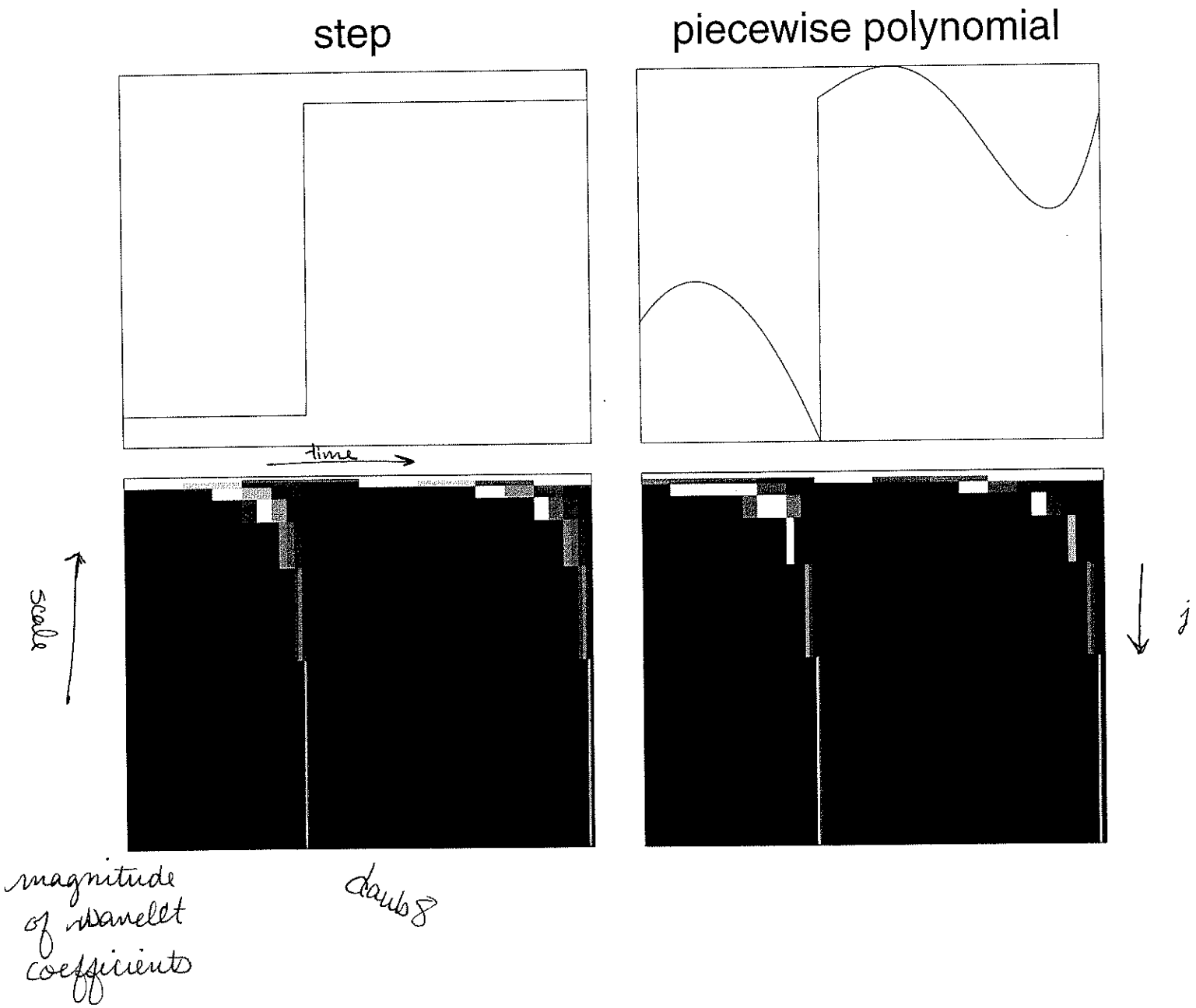
$$= h 2^{-j/2} \int |\psi(x)|\, dx = C_\psi h 2^{-j/2}$$

$\Rightarrow$ wavelet coeffs. decay exponentially across scale!

- Compare to Fourier domain

$$|f_k| \sim 1/k$$

- Same for piecewise polynomials, thanks to vanishing moments


step
piecewise polynomial
time
scale
j
magnitude of wavelet coefficients
daub 8

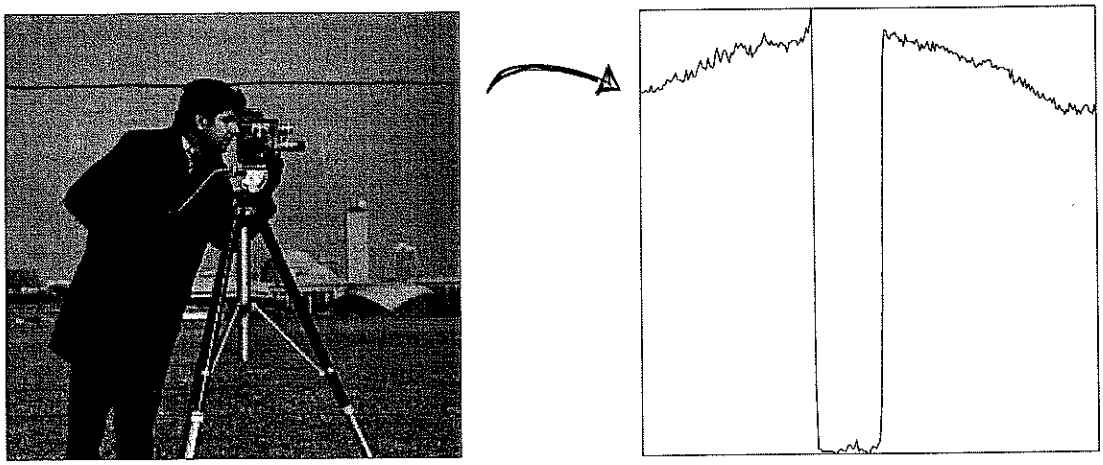
3 issues

1. Introduce a broad class of "piecewise-smooth" functions
   - Lipschitz continuity
   - Besov spaces
2. Describe how these functions behave in the wavelet domain
   - decay rates of w_{jk} across scale ← l_p spaces
3. Using 2, calculate how well these functions can be approximated using the "N largest" wavelet coefficients
   - decay rate of approx. error as $N \to \infty$

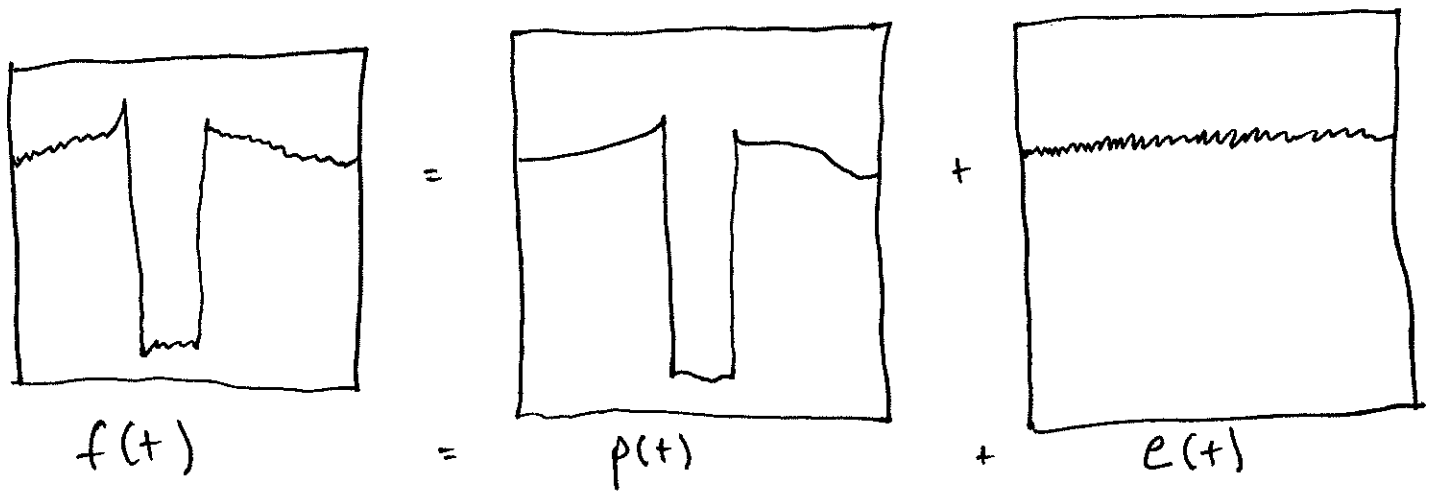
piecewise-smooth $\Rightarrow$ fast error decay of wavelets

Real data is not piecewise polynomial

Image Slice



Basic structure is piecewise polynomial, but with a small "texture" on top

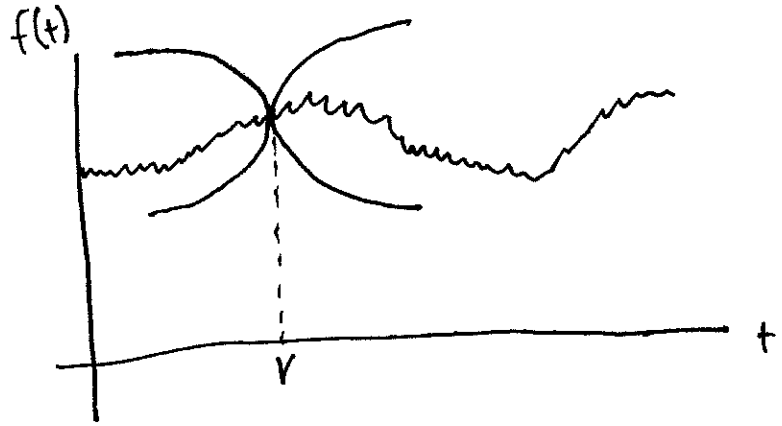


e(t) is continuous everywhere, but differentiable nowhere

Lipschitz Continuity

- **f. (for $0 \le \alpha < 1$):** A function f is **Lipschitz α**, $0 \le \alpha < 1$ at a point r if $\exists\ K_r$ such that

$$|f(t) - f(r)| \le K_r |t-r|^\alpha \qquad \forall\, t \in \mathbb{R}$$



- Catch phrase: "more than continuous, less than differentiable"

Def. (for all $\alpha \ge 0$): A function f is Lipschitz α at a point r if $\exists\ K_r$ and a polynomial $p_r(t)$ of degree $m = \lfloor \alpha \rfloor$ such that

$$|f(t) - p_r(t)| \le K_r |t-r|^\alpha \qquad \forall\, t \in \mathbb{R}$$

Basically, subtract a polynomial approximation to f at r (Taylor expansion) and look at the residual

- A function f is Lip α, $\alpha > 1$, if $f^{(m)}$ is lip $\alpha - m$, $m = \lfloor \alpha \rfloor$

Lipschitz cont.

In general, α can vary from point to point, so to model homogeneous regions of the signal, we define.

Def: A function f is <u>uniformly Lip α</u> over interval $[a,b]$ if $\exists\, K$ s.t.

$$|f(t) - p_r(t)| \le K \cdot |t-r|^\alpha \qquad \forall t \in \mathbb{R},\ \forall r \in [a,b]$$

- K is the same at every point in the interval

Facts:

+ f is uniformly Lip α on $[a,b]$
  $\Rightarrow f$ is $m = \lfloor \alpha \rfloor$ times continuously differentiable on $[a,b]$

+ f is uniformly Lip $\alpha \iff f^{(m)}$ is uniformly Lip $\alpha - m$

Lip α functions are $\lfloor \alpha \rfloor$ differentiable, but not necessarily $\lceil \alpha \rceil$

* The Lipschitz regularity at a point r classifies the <u>singularity type</u> of f at r.

(or smoothness)

$$f \in L^2 \Rightarrow f^{(2)} \in L^2 \Rightarrow \int |f^{(2)}|^2 dt < \infty \Rightarrow \int \omega^4 |f(\omega)|^2 d\omega < \infty$$

$$\text{"}f^{(\alpha)}\text{"} \in L^2 \Rightarrow \int \omega^{2\alpha} |f(\omega)|^2 < \infty$$

Lipschitz function fourier coefficients decay quickly BUT add one discontinuity and it all falls apart. $-1/n$ decay again

Lipschitz functions in the wavelet domain

The Lipschitz regularity of a function dictates certain behavior (i.e. the rate of exponential decay across scale) from the wavelet coefficients.

↳ different scales

Let wavelet ψ have q vanishing moments and let $\psi \in C^q$ with derivatives that have "fast decay"

then...

Thm: If $f \in L^2$ is Lip $\alpha \le q$ at v, then $\exists$ A such that

$$|W_{j,k}| = |\langle f, \psi_{j,k} \rangle| \le A 2^{-j(\alpha+1/2)} (1 + |k - 2^j v|^\alpha)$$

pf:

$$|\langle f, \psi_{j,k} \rangle| = \left| \int f(t) \psi_{j,k}(t)\, dt \right| = \left| \int (f(t) - p_r(t)) \psi_{j,k}(t)\, dt \right|$$

since $\int p_r \psi_{j,k}\, dt = 0$ (vanishing moments)

$$\le \int |f(t) - p_r(t)| \cdot |\psi_{j,k}(t)|\, dt$$

$$\le \int K_r |t - v|^\alpha \cdot |\psi_{j,k}(t)|\, dt \qquad \text{let } s = 2^j t - k$$

$$= 2^{-j/2} K_r \int |2^{-j}(s+k) - v|^\alpha |\psi(s)|\, ds$$

$|a+b|^\alpha \le 2^\alpha (|a|^\alpha + |b|^\alpha)$

$$\le 2^\alpha K_r 2^{-j/2} \left(\int |2^{-j} s|^\alpha \cdot |\psi(s)|\, ds + |2^{-j} k - v|^\alpha \int |\psi(s)|\, ds \right)$$

$$= A_1 2^{-j(\alpha+1/2)} \left(\underbrace{\int |s|^\alpha |\psi(s)|\, ds}_{\text{finite}} + |k - 2^j v|^\alpha \underbrace{\int |\psi(s)|\, ds}_{\text{finite}} \right)$$

$$\le A 2^{-j(\alpha+1/2)} (1 + |k - 2^j v|^\alpha)$$

Lip functions in wavelet domain cont.

There is also a converse to this theorem, but it is much more difficult to prove.

The wavelet domain behavior is even simpler if f is uniformly Lipschitz....

Thm: If $f \in L^2$ is uniformly Lip α, then

$\exists\ A > 0$ such that

$$|W_{j,k}| = |\langle f, \psi_{j,k} \rangle| \leq A\, 2^{-j(\alpha + 1/2)} \qquad \forall\, j,k$$

Pf: Take $\nu = 2^{-j} k$ above. ∎

Key idea: Wavelet coefficients of Lipschitz functions decay exponentially across scale (with rate depending on the regularity)

Next time

- We'll use what we know about Lip α functions in the wavelet domain to quantify how well n-term approximation works on them.

Then, we'll show the semi-amazing fact that adding a finite number of discontinuities to a Lip α function does <u>not</u> affect how well n-term approximation using wavelets works.

(compare to Fourier)

Wavelet Denoising

In the last lecture, we proved that for piecewise constant signals in noise the ideal estimator, which knows where the breakpoints are located, results in a "best case" error

$$\frac{E\left[\|x-\hat{x}\|^2\right]}{N} \sim O(N^{-1}),$$

whereas the best linear estimator, which doesn't know the breakpoint locations, produces an error

$$\frac{E\left[\|x-\hat{x}\|^2\right]}{N} \sim O(N^{-\frac{1}{2}}).$$

The same conclusions hold if we consider piecewise polynomial signals.

In this lecture, we will show that simple wavelet denoising rules (i.e., "killing" small wavelet coefficients and reconstructing) ← thresholding produce estimates of piecewise polynomial signals that <u>almost</u> achieve the best possible rate. Wavelet thresholding results in an error

$$\frac{E\left[\|x-\hat{x}\|^2\right]}{N} \sim O\left(\frac{\log^2 N}{N}\right)$$

Within a logarithmic factor of the "best case" error if we knew the breakpoint locations!!

clairvoyant:	$O(N^{-1})$
linear:	$O(N^{-1/2})$
nonlinear wavelet thresholding:	$O\left(\frac{\log^2 N}{N}\right)$

Coefficient-Wise (Diagonal) Estimators

Let us again decompose the noisy signal

$$Y(n) = X(n) + W(n)$$

in an orthogonal basis $\{g_m\}_{m=0}^{N-1}$:

$$\langle Y, g_m \rangle = \langle X, g_m \rangle + \langle W, g_m \rangle .$$

Since W is Gaussian white noise, the inner products

$$\langle W, g_m \rangle = \sum_{n=0}^{N-1} W(n)\, g_m(n)$$

are independent Gaussian random variables of zero mean and variance σ^2.

Check :

$$E\left[\langle W, g_m \rangle \langle W, g_k \rangle\right] = E\left[\sum_{n,n'=0}^{N-1} W(n) g_m(n)\, W(n') g_k(n'\right.$$

$$= \sum_{n,n'} g_m(n)\, g_k(n')\, \underbrace{E\left[W(n)\, W(n')\right]}_{\sigma^2 \delta(n-n')}$$

$$= \sigma^2 \langle g_m, g_k \rangle = \sigma^2 \delta(m-k).$$

In the following discussion we will estimate the signal x by estimating each coefficient $\theta_m = \langle x, g_m \rangle$ individually and computing the reconstruction:

$$\hat{x} = \sum_{m=0}^{N-1} \hat{\theta}_m g_m$$

where $\hat{\theta}_m$ = estimate of θ_m.

Also, for the time being, we will view the signal $x(n)$ as an unknown deterministic quantity, rather than a random process realization.

Ideal Coefficient Attenuation

Noise must be additive and uncorrelated

Consider an estimator of the form

$$\hat{x} = \sum_{m=0}^{N-1} \alpha_m \langle y, g_m \rangle g_m, \quad \alpha_m \in \mathbb{R},$$

i.e., $\hat{\theta}_m = \alpha_m \langle y, g_m \rangle$. The MSE is (where $\theta_m = \langle x, g_m \rangle$)

$$E\left[\|x - \hat{x}\|^2\right] = E\left[\|\theta - \hat{\theta}\|^2\right] \quad \text{(Parseval)}$$

$$= \sum_{m=0}^{N-1} E\left[(\theta_m - \hat{\theta}_m)^2\right]$$

Minimizing the MSE with respect to $\{\alpha_m\}_{m=0}^{N-1}$ yields the optimal weights

$$\alpha_m = \frac{|\langle x, g_m \rangle|^2}{|\langle x, g_m \rangle|^2 + \sigma^2}$$

Problem: x unknown $\Rightarrow$ $\langle x, g_m \rangle$ unknown $\Rightarrow$ ideal coefficient attenuation is not feasible.

The resulting MSE is

$$\epsilon_a = \sum_{m=0}^{N-1} \frac{|\langle x, g_m \rangle|^2 \sigma^2}{|\langle x, g_m \rangle|^2 + \sigma^2}$$

Although the ideal coefficient attenuation estimator resembles the Wiener filter, its action is very different.

The ideal attenuation factor

$$\alpha_m = \frac{|\langle x, g_m \rangle|^2}{|\langle x, g_m \rangle|^2 + \sigma^2}$$

is based on the coefficients of the actual signal that is observed. On the other hand, the Wiener filter attenuates according to

$$\frac{E\left[|\langle x, g_m \rangle|^2\right]}{E\left[|\langle x, g_m \rangle|^2 + \sigma^2\right]}$$

where x is viewed as a realization of a Gaussian process. Here the attenuation is based on the <u>average</u> or expected squared coeff. value, rather than the actual value of a particular realization.

Practical Coefficient Attenuation

Although we do not know $\langle x, g_m \rangle$ exactly, we can estimate this quantity from the data. Note

$$\langle y, g_m \rangle = \langle x, g_m \rangle + \langle w, g_m \rangle$$

and if we regard x as deterministic and unknown, then

$$E\left[|\langle y, g_m \rangle|^2 \right] = |\langle x, g_m \rangle|^2 + \sigma^2$$

$$\Rightarrow$$

$$|\langle x, g_m \rangle|^2 = E\left[|\langle y, g_m \rangle|^2 \right] - \sigma^2$$

$$\approx |\langle y, g_m \rangle|^2 - \sigma^2$$

Furthermore, to avoid the illogical situation where $|\langle y, g_m \rangle|^2 - \sigma^2$ is negative, we settle for the following estimate

$$|\langle x, g_m \rangle|^2 \approx \left(|\langle y, g_m \rangle|^2 - \sigma^2 \right)_+$$

where $(z)_+ = \begin{cases} z, & z \geq 0 \\ 0, & z < 0 \end{cases}$

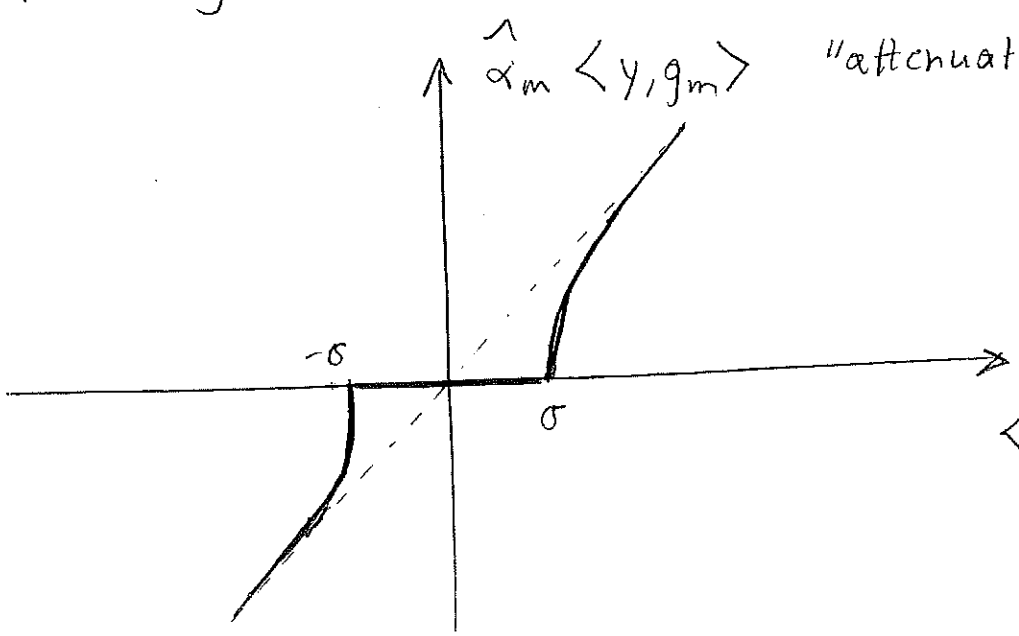
Plugging this estimate into the ideal coefficient attenuation produces

$$\hat{\alpha}_m = \frac{\left(|\langle y, g_m \rangle|^2 - \sigma^2\right)_+}{|\langle y, g_m \rangle|^2}$$

Hence, our estimate of x is

$$\hat{x} = \sum_{m=0}^{N-1} \frac{\left(|\langle y, g_m \rangle|^2 - \sigma^2\right)_+}{|\langle y, g_m \rangle|^2} \langle y, g_m \rangle \, g_m$$

The action of the estimator is revealed by plotting $\langle y, g_m \rangle$ vs. $\hat{\alpha}_m \langle y, g_m \rangle$



If $|\langle y, g_m \rangle| < \sigma$, $\hat{\alpha}_m \langle y, g_m \rangle = 0$.

If $|\langle y, g_m \rangle| > \sigma$, $\hat{\alpha}_m \langle y, g_m \rangle \approx \langle y, g_m \rangle$

From this we see that the "practical" coefficient attenuation scheme effectively acts like a threshold nonlinearity; coefficients with large magnitude are kept, small coefficients are set to zero (discarded).

This suggests a coefficient <u>selection</u> estimator, instead of attenuation. That is, let's estimate the signal x by <u>deciding</u> which of the noisy coefficients probably are important and contain significant signal information and use only these significant coefficients to estimate x.

Ideal Coefficient Selection

Coefficient selection is equivalent to restricting the weights α_m to the set $\{0,1\}$. If $\alpha_m = 0$, we discard a coefficient. If $\alpha_m = 1$, we keep the coefficient $\langle y, g_m \rangle$. Under this restriction, the MSE

$$E \| x - \sum \alpha_m \langle y, g_m \rangle g_m \|^2$$

is minimized by

$$\alpha_m = \begin{cases} 1, & \text{if } |\langle x, g_m \rangle| \geq \sigma \\ 0, & \text{if } |\langle x, g_m \rangle| < \sigma \end{cases}$$

This rule sets the coefficients in which the signal component is less than the noise std to zero.

The MSE is equal to

$$\epsilon_S = E \| x - \sum \alpha_m \langle y, g_m \rangle g_m \|^2 = \sum_{m=0}^{N-1} \min\left(|\langle x, g_m \rangle|^2, \sigma^2 \right)$$

The ideal selection MSE is slightly larger than the ideal attenuation MSE:

$$\min\left(|\langle x, g_m\rangle|^2, \sigma^2\right) \geq \frac{|\langle x, g_m\rangle|^2 \sigma^2}{|\langle x, g_m\rangle|^2 + \sigma^2} \geq \frac{1}{2}\min\left(|\langle x, g_m\rangle|^2, \sigma^2\right)$$

$$\Rightarrow$$

$$\epsilon_s \geq \epsilon_a \geq \frac{\epsilon_s}{2}$$

Hence, ideal selection is nearly as good as ideal attenuation, but it is a much more simple estimation rule.

Both estimators depend on the basis $\{g_m\}_{m=0}^{N-1}$ being employed, and consequently their performances depend on how efficiently the basis represents the signal x. This is easily seen in the ideal coefficient selection estimator.

Let M be the number of coefficients satisfying $|\langle x, g_m \rangle| \geq \sigma$. We know from earlier discussions that the best nonlinear, signal-adapted approximation of x using M vectors from $\{g_m\}$ is

$$x_M = \sum_{m:\, |\langle x, g_m \rangle| \geq \sigma} \langle x, g_m \rangle \, g_m$$

We also know that the error is

$$\|x - x_M\|^2 = \sum_{m:\, |\langle x, g_m \rangle| < \sigma} |\langle x, g_m \rangle|^2$$

This shows that the MSE of the ideal selection rule is

$$\epsilon_S = E\left[\|x - \hat{x}\|^2\right] = \overbrace{\|x - x_M\|^2}^{\text{Bias}} + \overbrace{M\sigma^2}^{\text{Variance}}.$$

The error is small if and only if both terms are small

$\iff$ M is small and x_M is a good approximation to x.

$\iff$ the basis provides a very efficient representation of x.

Thresholding Estimators

Ideal coefficient selection is also impractical. A practical coefficient selection rule is to threshold the noisy coefficients $\langle Y, g_m \rangle$, similar to the practical coefficient attenuation we looked at earlier. (p. 364)

To be specific, let's study the following estimator.

$$\hat{X} = \sum_{m=0}^{N-1} \delta_T(\langle Y, g_m \rangle)\, g_m$$

where $\delta_T(\cdot)$ is a "hard" threshold function

$$\delta_T(z) = \begin{cases} z, & \text{if } |z| > T \\ 0, & \text{if } |z| \leq T \end{cases}$$

T is the threshold level.

The estimation error is

$$\epsilon = \sum_{m=0}^{N-1} E\left[\left| \langle x, g_m \rangle - \delta_T\left(\langle y, g_m \rangle \right) \right|^2 \right]$$

with

$$\left| \langle x, g_m \rangle - \delta_T\left(\langle y, g_m \rangle \right) \right|^2 = \begin{cases} \left| \langle w, g_m \rangle \right|^2, & \text{if } \left| \langle y, g_m \rangle \right| > T \\ \left| \langle x, g_m \rangle \right|^2, & \text{if } \left| \langle y, g_m \rangle \right| \leq T \end{cases}$$

Hence,

$$\epsilon \geq \epsilon_S = \sum_{m=0}^{N-1} \min\left(\left| \langle x, g_m \rangle \right|^2, \sigma^2 \right)$$

ϵ_S: ideal selection error

since the threshold does not guarantee that the min of $|\langle x, g_m \rangle|^2$ and σ^2 is selected. Nonetheless, the following theorem proves that by choosing an appropriate threshold T, we can guarantee that ϵ remains within a $2 \log_e N$ factor of ϵ_S.

Theorem (Donoho and Johnstone)

If $T = \sigma\sqrt{2\log_e N}$, then the MSE of the hard thresholding estimator satisfies

$$\epsilon = E\left[\|x - \hat{x}\|^2\right] \leq (2\log_e N + 1)\left(\sigma^2 + \underbrace{\sum_{m=0}^{N-1} \min\left(|\langle x, g_m\rangle|^2, \sigma^2\right)}_{\epsilon_s}\right)$$

proof:

For a Gaussian random variable Z of mean μ and variance σ^2, define

$$\rho(T, \mu, \sigma) = E\left[\left(Z\, I_{|Z|>T} - \mu\right)^2\right]$$

In our estimation problem $\langle Y, g_m\rangle$ is a Gaussian r.v. of mean $\langle x, g_m\rangle$ and variance σ^2. The hard thresholding function $\delta_T(z) = z\, I_{|z|>T}$.

With this notation we can express the MSE as

$$\epsilon = E\left[\|x - \hat{x}\|^2\right] = \sum_{m=0}^{N-1} \rho\left(T, \langle x, g_m \rangle, \sigma\right).$$

Next define

$$L(T, \mu, \sigma) = \frac{\rho(T, \mu, \sigma)}{\sigma^2 N^{-1} + \min(\sigma^2, \mu^2)}$$

and

$$\Lambda(T, \sigma) = \sup_{\mu \in \mathbb{R}} L(T, \mu, \sigma)$$

Then using $\Lambda(T, \sigma)$ we can bound the error by

$$\epsilon \leq \sum_{m=0}^{N-1} \Lambda(T, \sigma)\left[\sigma^2 N^{-1} + \min(\sigma^2, |\langle x, g_m \rangle|^2)\right]$$

$$\leq \Lambda(T, \sigma)\left[\sigma^2 + \sum_{m=0}^{N-1} \min(\sigma^2, |\langle x, g_m \rangle|^2)\right]$$

Hence, to prove the theorem it suffices to show that

$$\Lambda(T, \sigma) \leq 2 \log_e N + 1$$

Equivalently, we must show that

$$L(T, \mu, \sigma) \leq 2 \log_e N + 1$$

for all $\mu \in \mathbb{R}$. Also note that $L(T, \mu, \sigma) = L(T, -\mu, \sigma)$, so we can restrict attention to $\mu \geq 0$. We can also assume that $\sigma = 1$ since

$$\ell(T, \mu, \sigma) = \sigma^2 \ell\left(\frac{T}{\sigma}, \frac{\mu}{\sigma}, 1\right).$$

Thus, we must show that

$$L(T, \mu, 1) \leq 2 \log_e N + 1 \qquad (\star)$$

for $\mu \geq 0$ and $T = (2 \log_e N)^{\frac{1}{2}}$.

To establish this inequality we will need to make use of the Gaussian distribution function.

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

$$F(z) = \int_{-\infty}^{z} f(u)\, du$$

$$G(z) = \int_{z}^{\infty} f(u)\, du = 1 - F(z)$$

The following lemma is used to prove (A).

Lemma :

(a) $\rho(T,\mu,1) \le T^2+1$, if $T \ge 1$ and $\mu \in \mathbb{R}$

(b) $\rho(T,\mu,1) \le \mu^2+1$, if $\mu \in \mathbb{R}$

(c) $\rho(T,\mu,1) \le \rho(T,0,1) + 1.2\mu^2$, if $0 \le \mu \le T$

(d) $\rho(T,\mu,1) \le 2f(T)(T+1)$, if $\mu = 0$ and $T \ge 1$

proof of (b) :

Observe that

$$(z I_{|z|>T} - \mu)^2 \le (z-\mu)^2 + \mu^2$$

Taking expectation proves (b) and (a) for $T \ge \mu$.

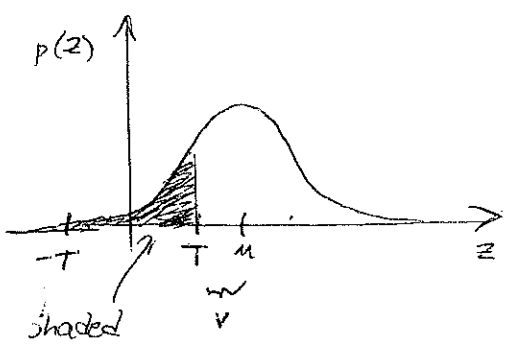
proof of (a): For $T < \mu$

$$E\left[(z I_{|z|>T} - \mu)^2\right] = P(|z|>T)\, E\left[(z-\mu)^2 \mid |z|>T\right] + P(|z| \le T)\,\mu^2$$

$$\le E\left[(z-\mu)^2\right] + P(|z| \le T)\,\mu^2$$

$$= 1 + P(|z| \le T)\,(T+v)^2 \qquad v = \mu - T$$

$$\le 1 + (T+v)^2 G(v).$$



If $T \geq 1$, then the exponential decay of $G(v)$ guarantees that

$$\frac{(T+v)^2}{T^2} G(v) = \left(1 + \frac{v}{T}\right)^2 G(v) \leq 1$$

for all $v \geq 0$, and hence

$$(T+v)^2 G(v) \leq T^2$$ proving (a) for $T < \mu$.

proof of (c):

We will show that $\frac{\partial^2 \rho(T, \mu, 1)}{\partial \mu^2} \leq 2.4$ which, according to Taylor's remainder Theorem, shows that for $0 \leq \mu \leq T$

$$\rho(T, \mu, 1) \leq \rho(T, 0, 1) + 2.4 \frac{\mu^2}{2}.$$

Note that

(#) $$\rho(T, \mu, 1) = \mu^2 \left[F(T-\mu) - F(-T-\mu)\right] + G(T-\mu) + G(T+\mu) + (T-\mu) f(T-\mu) + (T+\mu) f(T+\mu).$$

$\left.\begin{array}{l} \\ \end{array}\right\} P(|Z| \leq T) \cdot \mu^2$

$\left.\begin{array}{l} \\ \\ \\ \end{array}\right\} P(|Z| > E[(Z-\mu^2 \; |Z|$

Differentiating this expression twice, and using the inequalities

$$T(T \pm 2u) \leq (T \pm u)^2$$

and

$$4u\, f(T+u) - 4u f(T-u) \leq 0, \quad \text{for } 0 \leq T \leq u$$

and finally substituting $s = T+u$ and then $s = T-u$, we verify that

$$\frac{\partial^2}{\partial u^2} \rho(T,u,1) \leq 2 + 2 \sup_{s>0} \left[f(s)(s^3 - 2s) - 2F(-s) \right] \leq 2.4$$

proof of (d):

Again, due to the exponential decay of $G(T)$, we verify that

$$G(T) \leq T^{-1} f(T), \quad \text{for } T \geq 1$$

Using this inequality (and setting $u=0$) in (#) proves (d).

Now we can finish the proof of the theorem. by showing that for all $\mu \geq 0$ and $T = \sqrt{2 \log_e N}$ we have

$$L(T, \mu, 1) = \frac{\rho(T, \mu, 1)}{N^{-1} + \min(\mu^2, 1)} \leq 2 \log_e N + 1$$

First, <u>for $\mu \geq T$</u> (a) implies

$$L(T, \mu, 1) \leq \frac{T^2 + 1}{N^{-1} + 1} \leq 2 \log_e N + 1$$

For <u>$1 \leq \mu \leq T$</u> (b) implies

$$L(T, \mu, 1) \leq \frac{\mu^2 + 1}{N^{-1} + 1} \leq 2 \log_e N + 1$$

For <u>$0 \leq \mu \leq 1$</u> (c) implies

$$L(T, \mu, 1) \leq \frac{\rho(T, 0, 1)}{N^{-1}} + \frac{\rho(T, \mu, 1) - \rho(T, 0, 1)}{\mu^2}$$

$$\leq N \rho(T, 0, 1) + 1.2$$

and (d) shows that

$$\leq N \, 2 f\left(\sqrt{2 \log_e N}\right)\left(\sqrt{2 \log_e N} + 1\right) + 1.2$$

$$= \frac{N \cdot 2}{\sqrt{2\pi}} e^{-\frac{2 \log_e N}{2}} \left(\sqrt{2 \log_e N} + 1\right) + 1.2$$

Finally, for $N \geq 4$ we have

$$L(T, u, 1) \leq 2 \log_e N + 1$$

for the $\underline{0 \leq u \leq 1}$ case as well, completing the proof. ∎

The theorem shows that with $T = \sqrt{2 \log_e N} \cdot \sigma$, the hard thresholding estimator's MSE is bounded by

$$\epsilon \leq \left(2 \log_e N + 1\right)\left(\sigma^2 + \underbrace{\sum_{m=0}^{N-1} \min\left(|\langle x, g_m \rangle|^2, \sigma^2\right)}\right)$$

Recall this is ϵ_S: the MSE of the ideal coefficient selection estimator.

So,

$$\epsilon \leq \left(2 \log_e N + 1\right)\left(\sigma^2 + \epsilon_S\right)$$

Hence the hard thresholding error is within a $2\log_e N$ factor of $\epsilon_s + \sigma^2$. The extra σ^2 is on the order of the increase in the error of the ideal selection estimator if we had $N+1$ instead of N parameters, so it is more or less negligible.

Bottom Line:

MSE of hard thresholding $\leq 2\log_e N \times$ MSE of ideal coefficient selection

and since $\epsilon_s \geq \epsilon_a \geq \frac{\epsilon_s}{2}$,

MSE of hard thresholding $\leq 4\log_e N \times$ MSE of ideal coefficient attenuation.

So, hard thresholding performs almost as well as ideal (but impractical) methods. But, how well are the ideal methods? This depends on the basis $\{g_m\}$ and the structure of the signal underlying the observations.

This leads us to the most interesting result of all. Let's revisit the piecewise polynomial signal estimation problem. Recall that the ideal (but unrealizable) nonlinear estimator that assumes the breakpoints are known produces an error that decays like $\frac{1}{N}$. In contrast, the best linear estimator (without knowledge of breakpoints) has an error that decays like $\frac{1}{\sqrt{N}}$.

Proposition: (Donoho & Johnstone)

Suppose that x is a piecewise polynomial over $[0, N-1]$ and that each polynomial segment is of degree $\leq d$. Then a hard thresholding estimator, formulated with a Daubechies wavelet basis with $d+1$ vanishing moments has an error satisfying

$$\frac{E\left[\|x-\hat{x}\|^2\right]}{N} \leq \left(1+2K(d+1)\log N\right)\frac{(2\log N+1)\sigma^2}{N}$$

$$\sim O\left(\frac{\log^2 N}{N}\right)$$

proof: The previous theorem proves that in any basis

$$E\left[\|x-\hat{x}\|^2\right] \le (2\log N + 1)\left(\sigma^2 + \sum_{m=0}^{N-1} \min\left(|\langle x, g_m\rangle|^2, \sigma^2\right)\right)$$

and in particular, we can take $\{g_m\}_{m=0}^{N-1}$ to be a discrete wavelet basis. $\{\psi_{j,k}\}_{j,k}$.

If M is the number of inner products $\langle x, \psi_{j,m}\rangle \neq 0$, then

$$E\left[\|x-\hat{x}\|^2\right] \le (2\log N + 1)(M+1)\sigma^2 \quad (\star)$$

Now assume that the wavelets have $d+1$ vanishing moments. If the support of $\psi_{j,k}$ does not include one of the breakpoints, then $\langle x, \psi_{j,k}\rangle = 0$.

Also note that Daubechies wavelets with $d+1$ vanishing moments have support (width) of $2^j(2d+2)$.

Hence, if the signal has K pieces/segments, then at each scale 2^j there are <u>at most</u>

$K(2d+2)$ wavelets that overlap with the breakpoints. Therefore, we have at most $K(2d+2)$ non-zero wavelet coefficients at each scale.

We also have at most $\log N$ scales, so in total there are

$$M \leq K(2d+2)\log N$$

Using this bound, we find

$$\frac{E[\|x-\hat{x}\|^2]}{N} \leq \left(1 + 2K(d+1)\log N\right)\frac{(2\log N+1)\sigma^2}{N}$$

$$\sim O\left(\frac{\log^2 N}{N}\right).$$

So, this proposition shows that for piecewise polynomial signal estimation the wavelet-based hard thresholding error is on the order of $\frac{\log^2 N}{N}$,

much smaller than $\frac{1}{\sqrt{N}}$ for large N,

and very close to $\frac{1}{N}$ for large N.

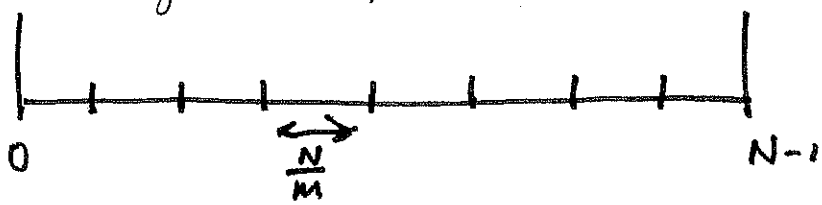
$\Rightarrow$ even if we knew the breakpoints we could not do much better.

$\Rightarrow$ nonlinear wavelet-based thresholding is much better than any linear estimator, even the Wiener filter.

Estimators

Linear piecewise polynomial:

↳ not signal adaptive

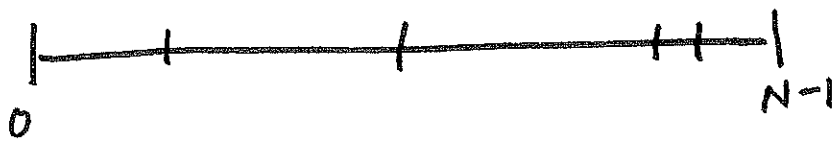


- M equal width pieces
- polynomial fit on each piece

degrees of freedom = $M \cdot (d+1)$ $\Rightarrow$ $M(d+1)$-dimension subspace

minimum MSE with $M \sim \sqrt{N}(d+1)$

Clairvoyant (piecewise polymial with known breakpoints)

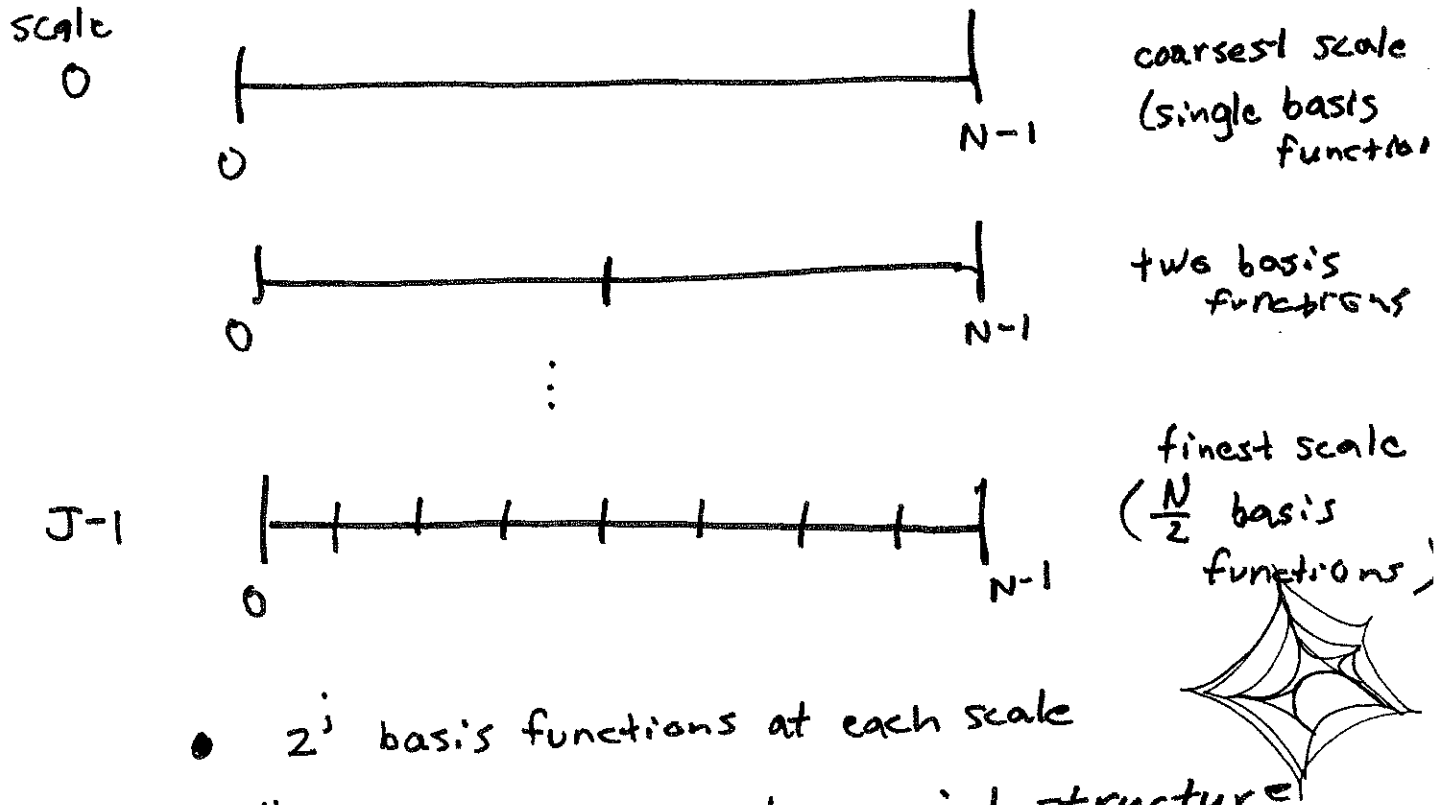


- K unequal width pieces
- polynomial fit to each piece

degrees of freedom = $K(d+1)$

$\Rightarrow$ $K(d+1)$ dimensional subspace

Wavelet-based:



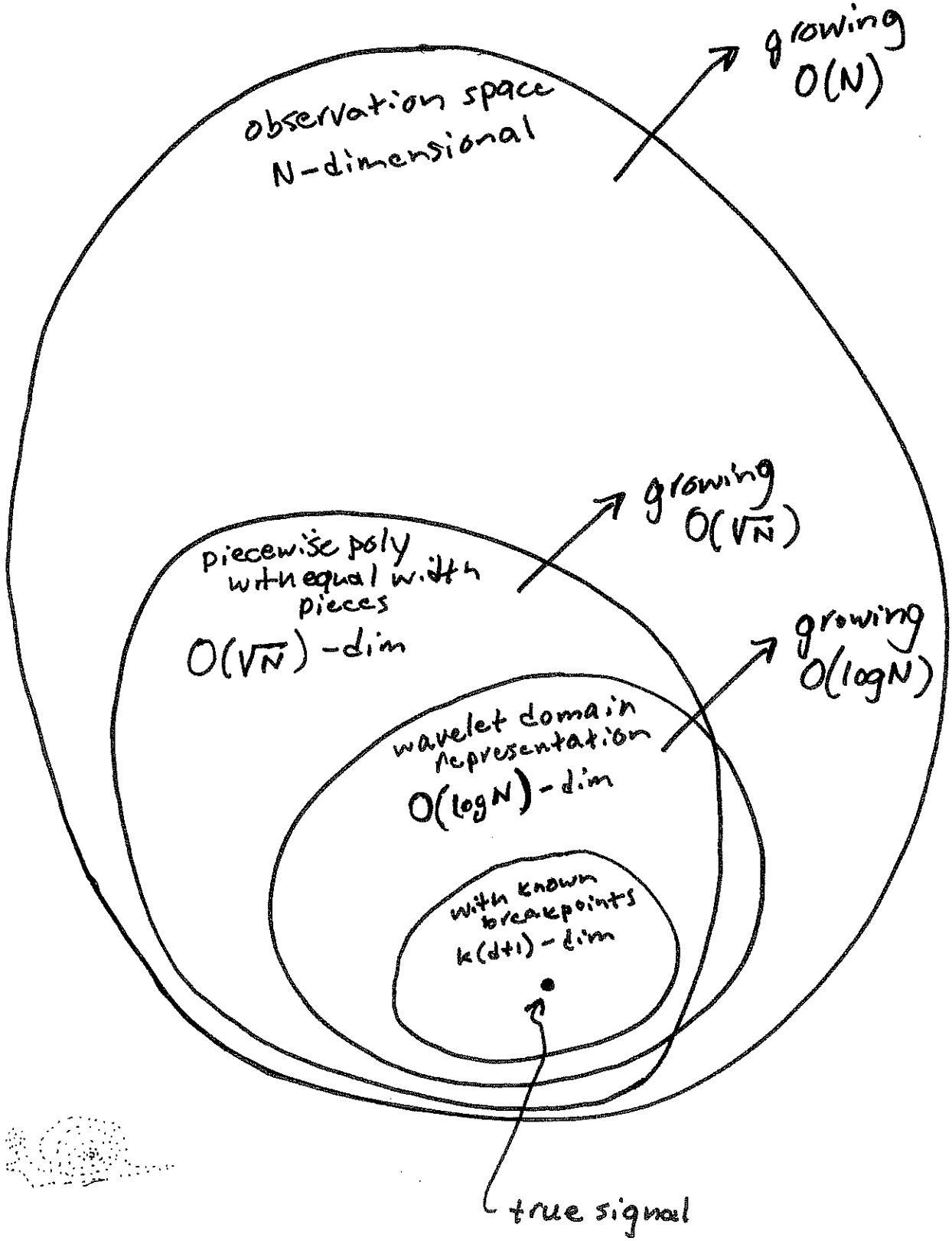
- 2^j basis functions at each scale
- "blind" to pure polynomial structure
- at most k true breakpoints

  $\Rightarrow$ at most $O(k)$ non-zero coefficients at each scale

$\Rightarrow$ true signal lies in an $O(K \cdot \log N)$ dimensional subspace of N-dimensional observation space

increases as N increases, but as $\log(N)$ (good)

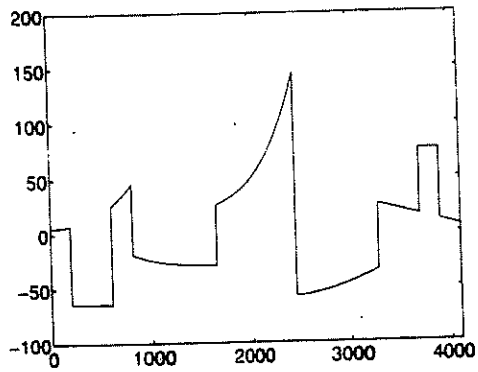
Subspace Perspective

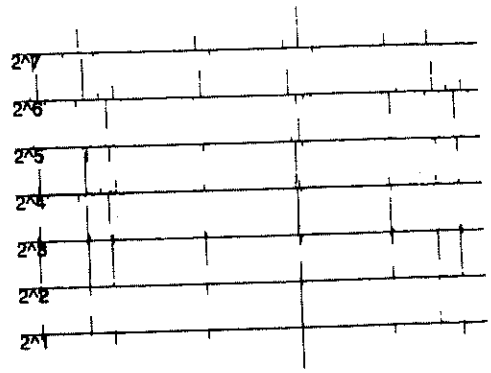


rate of error decay = $\frac{\text{subspace dim of signal representation}}{\text{observation space dim}}$

ex.

- signal
  piecewise
  polynomial
  x.



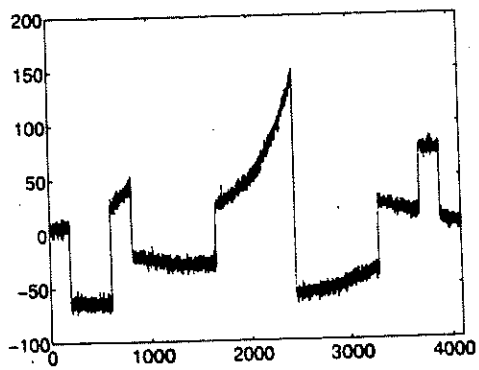


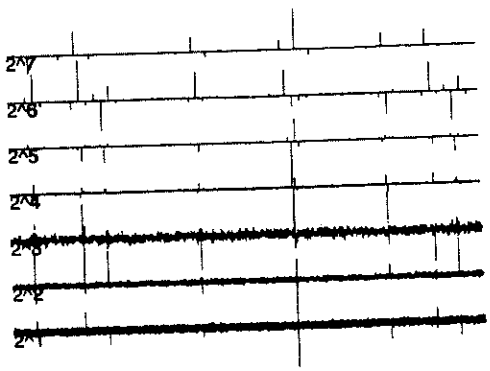
(a)

noisy
observation

$\frac{1}{SNR} = \frac{E[\|x-y\|^2]}{\|x\|^2}$

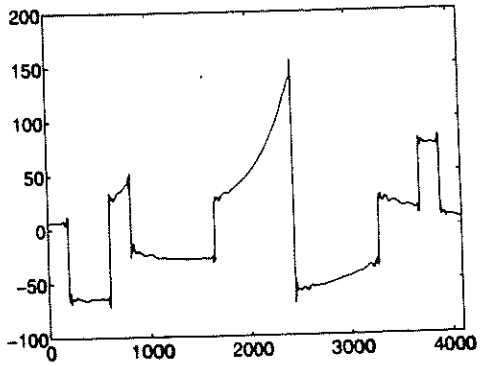
$= 21.9\,dB$

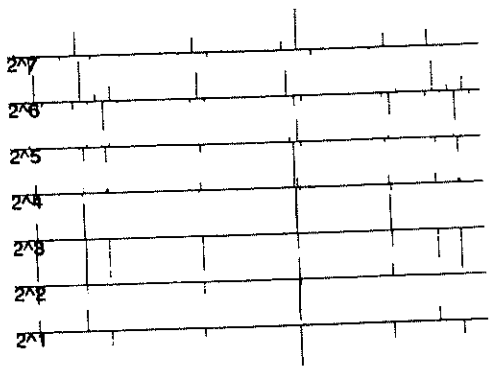




(b)

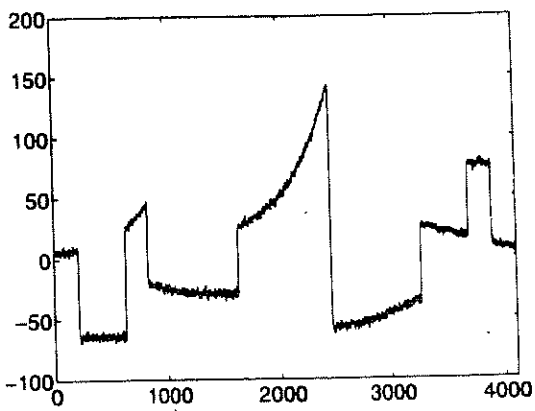
- hard threshold
  estimate
  $SNR = 30.8\,dB$





(c)

- Wiener
  filter
  estimate
  $SNR = 25.9\,dB$



(c)

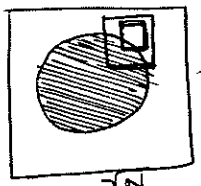
Covariance formed by computing cov of all possible circular shifts of x.

⇒ circular stationary signal model

⇒ DFT based Wiener filter

N pixels total

Ex.



$O(\sqrt{N})$ pixels on circle boundary
× $(\log(N))$ scales
$= O(\sqrt{N}\log N) = \#$ non zero W.C.

$\frac{\sqrt{N}\log N}{N} \sim O(1/\sqrt{N})$

(a) Image

(b) Image+Noise: SNR = 14.60 dB

(c) Denoised Image: SNR = 20.17 dB

(d) TI Denoised Image: SNR = 22.09 dB

Fig. 3. *Denoising using threshold rule* $\delta(z) = \frac{(|z|^2-3\sigma^2)_+}{z}$. *(a) Original image. (b) Noisy image. (c) Wavelet-based denoise (Haar basis). (d) TI wavelet-based denoise (Haar basis).*

Translation Invariant

denoise w/ shifted wavelet bases, average results

- what's the big deal about GLM?
- Bias vs. Variance tradeoff
- General polynom concavity
- General concavity
- Don's errors
- Multivar Steepest descent