ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 17**

Wavelet Approximation

Lecturer: Rob Nowak *Scribe: Changfang Zhu*

1 Review of Wavelet Approximation

Assume $f \in L_2([0,1])$, *i.e.,* $\int_0^1 f^2(t)dt < \infty$. Then f has a wavelet series representation:

$$f = C_0 + \sum_{j \geq 0}^{\infty} \sum_{k=1}^{2^j} \langle f, \psi_{j,k} \rangle \psi_{j,k}$$

, where $C_0 = \int_0^1 f(t)dt$. $j \sim$ scale, $k \sim$ position.

Key properties:

1. vanishing moments: We say ψ has M vanishing moments when $\int_0^1 t^m \psi_{j,k}(t)dt = 0$ for $m = 0, 1, \ldots, M-1$ This means the wavelet is "blind" to polynomial segment with degree $\alpha \leq M-1$
2. compact support: a daubechies wavelet with M vanishing moments has support $\propto$ 2M.

Together, these properties imply that only $O(l \log n)$ nonzero wavelet coefficients are needed to represent a piecewise polynomial function with l pieces and degree $\alpha \leq M-1$ on each piece.

2 Wavelet Approximation for Sampled Data

We can also use DWT to analyze and represent discrete, sampled functions. Suppose

$$\vec{f} = [f(\frac{1}{n}), f(\frac{2}{n}), \ldots, f(\frac{n}{n})]$$

then we can write $\vec{f}$ as

$$\vec{f} = C_0 + \sum_{j=0}^{\log_2 n - 1} \sum_{k=1}^{2^j} \langle \vec{f}, \vec{\psi}_{j,k} \rangle \vec{\psi}_{j,k}$$

where

$$\vec{\psi}_{j,k} = [\psi_{j,k}(1), \psi_{j,k}(2), \ldots, \psi_{j,k}(n)]$$

is a discrete time analog of the continuous time wavelets we considered before. In particular,

$$\sum_{i=1}^{n} i^l \psi_{j,k}(i) = 0$$

where $l = 0, 1, \ldots, N-1$ for the Daubechies-N discrete wavelets.

$$\langle \vec{f}, \vec{\psi}_{j,k} \rangle = \vec{f}^T \vec{\psi}_{j,k}$$

Thus we also have an analogous approximation result: If $\vec{f}$ are samples from a piecewise polynomial function of degree $\leq N$, with a finite number n of discontinuities, then $\vec{f}$ has $O(nJ)$ non-zero wavelet coefficients.

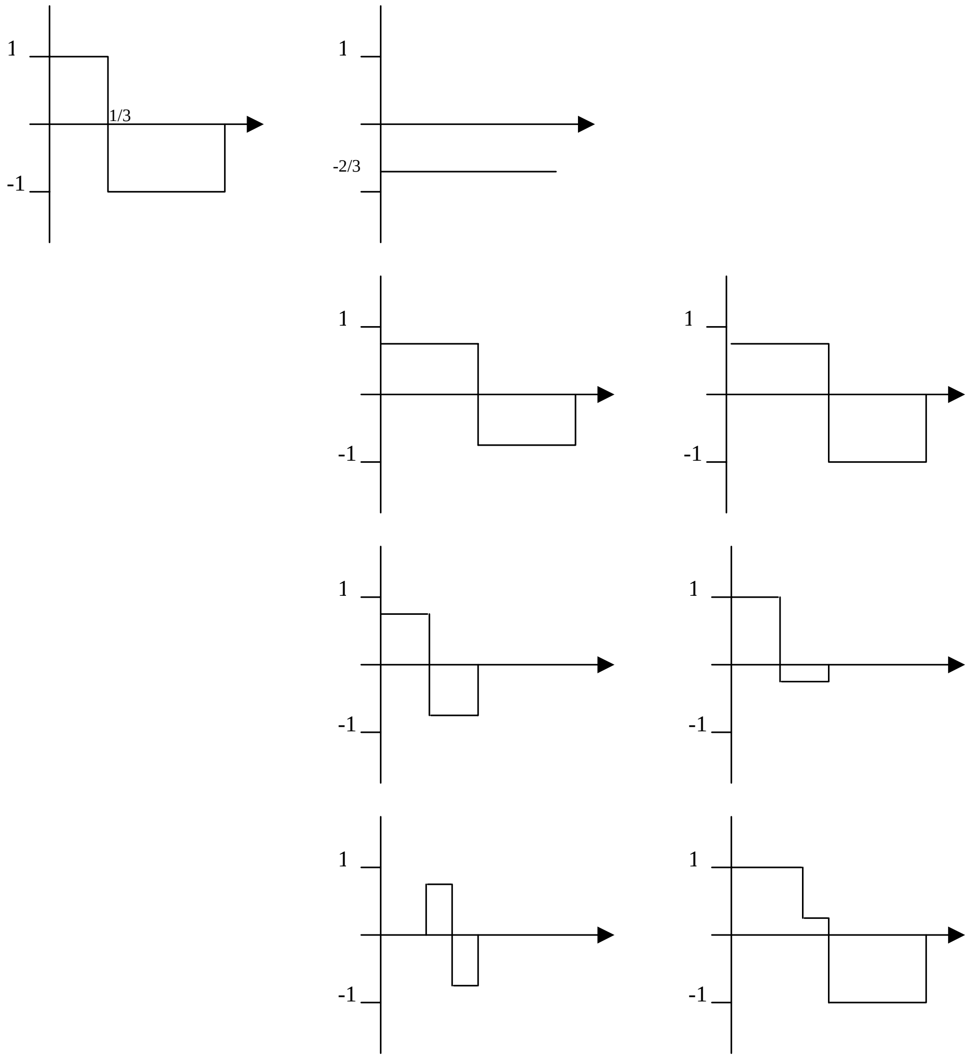

Figure 1: wavelet illustration: 1 break point in f $\Rightarrow$ 1 non-zero wavelet coefficient per scale

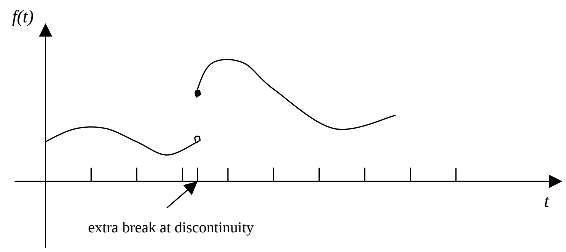

Figure 2: Partition on function with finite number of discontinuity

3 Approximating B^α functions with wavelets

Suppose $f \in B^\alpha(C_\alpha)$ and has a finite number of discontinuities. Let f_p denote a piecewise degree-N ($N = \lceil \alpha \rceil - 1$) approximation to f with $O(k)$ pieces; a uniform partition into k equal length intervals followed by addition splits at the point of discontinuity.

Then

$$|f(t) - f_p(t)|^2 = O(k^{-2\alpha}),\ \forall t \in [0,1]$$

$$\Longrightarrow |f(\frac{i}{n}) - f_p(\frac{i}{n})|^2 = O(k^{-2\alpha}),\ i = 1, 2, \ldots, n$$

$$\Longrightarrow \frac{1}{n} \parallel \vec{f} - \vec{f_p} \parallel_{l_2}^2 = O(k^{-2\alpha})$$

and $\vec{f_p}$ has $O(k \log_2(n))$ non-zero wavelet coefficients according to our previous analysis.

In the last lecture we saw that we could approximate well a function in $B^\alpha(C_\alpha)$ using a piecewise polynomial function, defined over a RDP, and that approximation could be represented in a tree structure using a small number of leafs (therefore an efficient representation).

In the current lecture we observed that the wavelets can also be used to represent efficiently some piecewise polynomial functions.

4 Denoising in Smooth Function Space III

note: $\longrightarrow$ Beyond Holder.

Let $f^* \in B^\alpha(C_\alpha)$ unknown ($\alpha \leq \alpha_{max}$), and function $f : [0,1] \rightarrow$ R consisting of a finite number of "piece" that are $H^\alpha(C_\alpha)$. Assume also that $|f^*(t)| \leq M,\ \forall t \in [0,1]$.

Consider our usual observation model:

$$Y_i = f^*(\frac{i}{n}) + W_i,\ i = 1, \ldots, n$$

$W_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ and $x_i = \frac{i}{n}$

Given a set of observations $\{Y_i\}$, we would like to estimate $f^*(\frac{i}{n})$. Let our candidate estimate be $f(\frac{i}{n})$, where f is a function that is piecewise polynomial. We are interested essentially only in the values of f at the sample points, *i.e.,* $f(\frac{i}{n}) : i = 1, \ldots, n$.

Let's introduce some notations:

$$\vec{Y} = [Y_1, \ldots, Y_n]^T$$

$$\vec{f^*} = [f^*(\frac{1}{n}), \ldots, f^*(\frac{n}{n})]^T$$

$$\vec{f} = [f(\frac{1}{n}), \ldots, f(\frac{n}{n})]^T$$

We are assuming our Gaussian model:

$$-\log P_f(Y_i) = \frac{(Y_i - f(\frac{i}{n}))^2}{2\sigma^2} + const.$$

therefore

$$\hat{f}_n = \arg\min_{f\in\mathcal{F}}\{\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - f(\frac{i}{n}))^2}{2\sigma^2} + \frac{2c(f)\log 2}{n}\}$$

or using vector notation:

$$\begin{aligned}\hat{f}_n &= \arg\min_{f\in\mathcal{F}}\{\frac{(\vec{Y}-\vec{f})^T(\vec{Y}-\vec{f})}{2\sigma^2} + 2c(f)\log 2\} \\ &= \arg\min_{f\in\mathcal{F}}\{\frac{\parallel \vec{Y}-\vec{f}\parallel^2}{2\sigma^2} + 2c(f)\log 2\}\end{aligned}$$

We need to choose $\tilde{f}$ and $C(f)$. This is where wavelet come

Consider a wavelet transform with $\lceil \alpha_{max} \rceil$ vanishing moments (capable of representing efficiently piecewise polynomial function with polynomial pieces of degree $\lceil \alpha_{max} \rceil - 1$). This is a linear and orthogonal transform. Represent it by a matrix $\mathbf{W}$ (note that $\mathbf{W}^{-1} = \mathbf{W}^T$).

Let $\vec{\theta} = \mathbf{W}^T\vec{f}$, we will use $\vec{\theta}$ instead of $\vec{f}$. Now let

$$\Theta = \{\vec{\theta} : \textit{such that } \theta_i \textit{ is quantized using } \frac{1}{2}\log_2(n) \textit{ bits};\ |\theta_i| \leq M\}$$

Now we want to choose $C(\vec{\theta})$ to somehow emphasize the fact that a good candidate should be a piecewise polynomial with few pieces, hence only a few wavelet coefficients are non-zero.

5 Encoding the Wavelet Coefficients

Prefix-free encoding: Encode the non-zero coefficients one-by-one. Use one bit to indicate if there are still any non-zero coefficients to encode, if so encode one of them using $\log_2(n)$ bits for the location and $\frac{1}{2}\log_2(n)$ bits for the magnitude.

Hence in total we have:

$$\begin{aligned}C(\vec{\theta}) &= (1 + \frac{3}{2}\log_2(n))\#\{none - zero\ coefficients\} \\ &\approx \frac{3}{2}log_2(n)\#\{i : \theta_i \neq 0\}\end{aligned}$$

Since it is a prefix code, it satisfies the Kraft inequality:

$$\sum_{\vec{\theta}\in\Theta} 2^{-C(\vec{\theta})} \leq 1$$

Now we are ready to analyze our estimation using the complexity penalized result:

$$\hat{\theta}_n = \arg\min_{\vec{\theta}\in\Theta}\{\frac{\parallel \vec{Y} - W^T\vec{\theta}\parallel^2}{2\sigma^2} + 2C(\vec{\theta})\log 2\}$$

$$\hat{f}_n = W^T\hat{\theta}_n$$

And our bound is:

$$\frac{1}{n}\sum_{i=1}^{n}E[(\hat{f}_n(\frac{i}{n}) - f^*(\frac{i}{n}))^2] = \frac{1}{n}E[\parallel \vec{\hat{f}}_n - \vec{f^*}\parallel^2] \leq \min_{\vec{\theta}\in\Theta}\{\frac{2}{n}\parallel W^T\vec{\theta}\parallel^2 + \frac{8\sigma^2 C(\vec{\theta})\log 2}{n}\}$$

Suppose $f^* \in B^\alpha(C_\alpha)$ for $\alpha \leq \alpha_{max}$. Let $\tilde{f}$ be the samples of the best k piece piecewise polynomial approximation. As we saw in the beginning of the class

$$\frac{1}{n} \parallel \vec{\tilde{f}}_n - \vec{f^*} \parallel^2 \leq C_\alpha^2 k^{-2\alpha} = O(k^{-2\alpha})$$

as $k \to \infty$

Also the wavelet transform of $\tilde{f}$, defined as $\tilde{\theta} = W\tilde{f}$ has $O(k \log n)$ non-zero entries, as k, $n \to \infty$.

Now let $\underline{\theta}$ be quantized version of $\tilde{\theta}$, then

$$\begin{aligned} \parallel W^T\underline{\theta} - \vec{f^*} \parallel^2 &= \parallel W^T\underline{\theta} - W^T\tilde{\theta} + W^T\tilde{\theta} - \vec{f^*} \parallel^2 \\ &= \underbrace{\parallel W^T\underline{\theta} - W^T\tilde{\theta} \parallel^2}_{\|\underline{\theta}-\tilde{\theta}\|} + \parallel W^T\tilde{\theta} - \vec{f^*} \parallel^2 + 2\langle W^T\underline{\theta} - W^T\tilde{\theta}, W^T\tilde{\theta} - \vec{f^*}\rangle \end{aligned}$$

so

$$\begin{aligned} &\leq \parallel \underline{\theta} - \tilde{\theta} \parallel^2 + \parallel W^T\tilde{\theta} - \vec{f^*} \parallel^2 + 2 \parallel \underline{\theta} - \tilde{\theta} \parallel \cdot \parallel W^T\tilde{\theta} - \vec{f^*} \parallel \\ &\leq M^2 n^{-1} + C_\alpha^2 k^{-2\alpha} + 2MC_\alpha k^{-\alpha} n - \frac{1}{2} \end{aligned}$$

Also

$$\begin{aligned} C(\underline{\theta}) &= 1 + (\frac{3}{2}\log_2(n) + 1)\#\{i : \theta_i \neq 0\} \\ &= O(k \log^2 n) \end{aligned}$$

Plugging all this into the bound we get:

$$\frac{1}{n} E[\parallel \hat{f}_n - \vec{f^*} \parallel^2] \leq O(\max\{\frac{1}{n}, k^{-2\alpha}, k^{-2\alpha} n^{-\frac{1}{2}}, \frac{k \log^2 n}{n}\})$$

Choosing $k = \lfloor n^{\frac{1}{2\alpha+1}} \rfloor$ yields

$$\frac{1}{n} E[\parallel \hat{f}_n - \vec{f^*} \parallel^2] = O(n^{-\frac{2\alpha}{2\alpha+1}} \log^2 n)$$

This is almost the same performance as if we had $f^* \in H^\alpha(C_\alpha)$, with known α!!

So at the expense of an extra $\log^2 n$ factor we get adaptability to unknown $\alpha \leq \alpha_{max}$ (recall episode II of "Denoising ..."), and also perform almost as well as if we know where the discontinuities of f^* are!!

All this is very appealing, but can we compute $\hat{\theta}_n$ and therefore $\hat{f}_n = W^T\hat{\theta}_n$ easily?

The answer is "yes".

$$\hat{\Theta}_n = \arg\min_{\bar{\theta} \in \Theta} \{\frac{\parallel \bar{Y} - W^T\bar{\theta} \parallel^2}{2\sigma^2} + 2C(\bar{\theta}) \log 2\}$$

Let $\bar{Z}$ be the wavelet transform of $\bar{Y}$, $\bar{Z} = W\bar{Y}$, then

$$\begin{aligned} \hat{\theta}_n &= \arg\min_{\bar{\theta} \in \Theta} \left\{ \frac{\parallel W^T\bar{Z} - W^T\bar{\theta} \parallel^2}{2\sigma^2} + 2 \cdot \frac{3}{2} \log_2 n \#\{i : \theta_i \neq 0\} \log 2 \right\} \\ &= \arg\min_{\bar{\theta} \in \Theta} \left\{ \frac{\sum_{i=1}^n (Z_i - \theta_i)^2}{2\sigma^2} + 3 \log n \sum_{i=1}^n 1_{\{\theta_i \neq 0\}} \right\} \\ &= \arg\min_{\bar{\theta} \in \Theta} \sum_{i=1} n \left\{ \underbrace{\frac{(Z_i - \theta_i)^2}{2\sigma^2} + 1_{\{\theta_i \neq 0\}} 3 \log n}_{U_i} \right\} \end{aligned}$$

We can optimize each term in the summation independently: If $\theta_i = 0$, then $U_i = \frac{Z_i^2}{2\sigma^2}$, If $\theta_i \neq 0$, we might as well take $\theta_i = Z_i$ and so $U_i = 3\log n$. Therefore

$$\hat{\theta}_i = \begin{cases} Z_i & \text{if } |Z_i| \geq \sqrt{6\sigma^2 \log n} \\ 0 & \text{otherwise} \end{cases} = h(Z_i) \tag{1}$$

Thus we just need to compute the wavelet coefficients of $\bar{Y}$ ($\bar{Z} = W\bar{Y}$), threshold according to (1), and compute the inverse wavelet transform $\hat{f} = W^T h(W\bar{Y})$. Since the wavelet transform is very fast ($O(n)$ operation) this denoising scheme is very efficient.

6 The Intuition Behind the Thresholding Approach

The observation are of the form $Y_i = f^*(\frac{i}{n}) + W_i$ where $(W_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$. Let

$$\vec{U} = (W_1, \ldots, W_n)^2$$

The i.i.d. assumption tells us that $E[\vec{U}] = 0$, and $Cov[\vec{U}] = E[\vec{U}\vec{U}^T] = \sigma^2\mathbf{I}$. Now let's analyze the wavelet transform of $\vec{Y}$:

$$W\vec{Y} = W(\vec{f^*} + \vec{U}) = W\vec{f^*} + W\vec{U}$$

and

$$E[W\vec{U}] = WE[\vec{U}] = O$$

$$\begin{aligned} Cov[W\vec{U}] &= E[(W\vec{U})(W\vec{U})^T] \\ &= WE[\vec{U}\vec{U}^T]W^T \\ &= \sigma^2 WW^T = \sigma^2\mathbf{I} \end{aligned}$$

Therefore the noise has the same characteristics both in the original domain and the wavelet domain. Why is this interesting?

Since f^* is a $B^\alpha(C_\alpha)$ function, it's representation by a wavelet transform has a few large magnitude coefficients and al the others are small.

Thus we keep the bulk of important coefficients, and remove the small ones that get "buried" in noise. Since only a few remain, most noise is removed.

7 Key Points

- Function in $B^\alpha(C_\alpha)$ have a sparse representation in the wavelet domain.
- The noise is white, and affects all the coefficients equally (for colored noise, much more sophisticated techniques must be used).

8 Wavelet in 2-D

Suppose f is a 2-D image that is piecewise polynomial:

Let f_p denote a 2-D piecewise polynomial (degree α) approximation to f using $O(k)$ pieces. Divide the image into k equal-sized bins, and divide to pixel level in bins near the boundary.

$$|f(s,t) - f_p(s,t)|^2 = \begin{cases} O(k^{-\alpha}) & (s,t) \in \text{big box} \\ O(n^{-\alpha}) & (s,t) \in \text{little box without boundary} \\ O(1) & (s,t) \in \text{little boundary box} \end{cases}$$

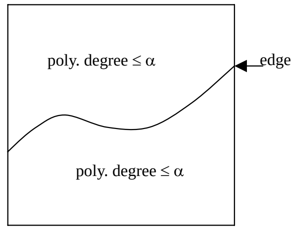

Figure 3: Approximate f with a 2-D piecewise polynomial of degree α

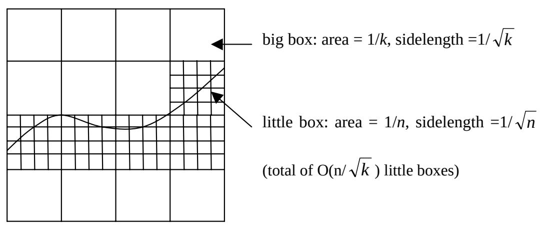

Figure 4: Divide image into K equally-sized bins, and divide to pixel level in bins near the boundary

$$\Rightarrow \| f - f_p \|_{L^2}^2 = O(k^{-\alpha} + 1/\sqrt{n}) \tag{2}$$

So if $\alpha \leq 1/2$, we can take advantage of the surface smoothness to get fast approx. err. decay, but if $\alpha > 1/2$, approximation error is dominated by the boundary.

How many non-zero wavelet coefficients?

At scale j, there are 4^j total coefficients, and boundary intersects support of 2^j (i.e. non-zero) $\Longrightarrow$ $\sum_{j=1}^{\log_4 n} 2^j \leq 4^{\frac{1}{2}\log_4 n} = n^{1/2}$.

- approximation error $\sim 1/\sqrt{n}$
- estimation error $\sim 1/\sqrt{n}$
- total error $\sim 1/\sqrt{n}$

This is why extensions such as curvelets, wedgelets, are so relevant.