ECE901 Spring 2004 Statistical Regularization and Learning Theory **Lecture: 16**

NonLinear Approximation and Wavelet Analysis

Lecturer: Rob Nowak *Scribe: Usman Khan*

1 Review

In Lecture 4 and 15, we investigated the problem of denoising a smooth signal in additive white noise. In Lecture 4, we considered Lipschitz functions and showed that by filling constants on a uniform partition of width $n^{-1/3}$ we can achieve an $n^{-2/3}$ rate of MSE convergence.

In Lecture 15, we considered Holder-α smooth functions, and we demonstrated that by automatically selecting partition width and using polynomial fits we can obtain a MSE convergence rate of $n^{-2\alpha/2\alpha+1}$, substantially better when $\alpha > 1$. Also important is the fact that we don't need to know the value of α a priori. The estimator $\widehat{f}_n$ is fundamentally different than its counterpart in Lecture 4.

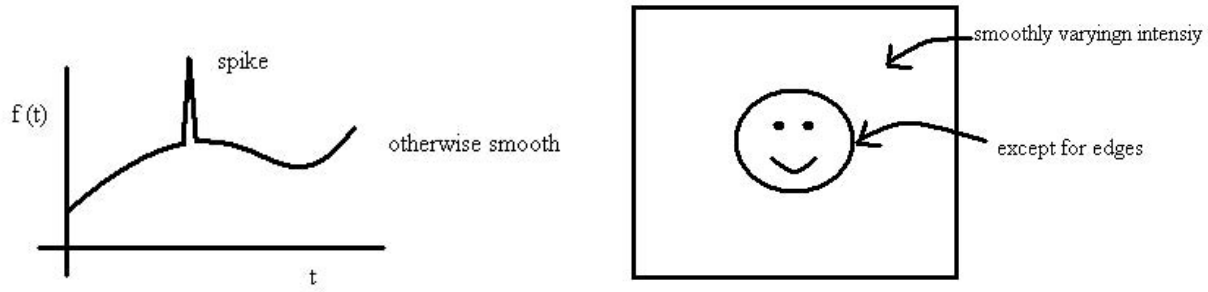
In both cases $\widehat{f}_n(t)$ is a linear function (polynomial on constant fit) of the data in each interval of the underlying partition. In Lecture 4, the partition was independent of the data, and so the overall estimator is a **linear function of the data**.

However, in Lecture 15 the partition itself was selected based on the data. Consequently, $\widehat{f}_n(t)$ is a **non-linear function of the data**. Linear estimators (linear functions of the data) cannot adapt to unknown degrees of smoothness. In this lecture, we lay the groundwork for one more important extension in the denoising application - spatial adaptivity. That is, we would like to construct estimators that not only adapt to unknown degrees of global smoothness, but that also adapt to **spatially varying** degrees of smoothness.

We will focus on the approximation theoretic aspects of the problem in this lecture, considering tree-based approximations and wavelet expansions. In the next lecture, we will apply these results to the denoising problem, this will bring us up to date with the current state-of-the-art in denoising and non-parametric estimation.

Recall that Holder spaces contain smooth functions that are well approximated with polynomials or piecewise polynomial functions. Holder spaces are quite large and contain many interesting signals. However, Holder spaces are still inadequate in many applications. Often, we encounter functions that are not smooth everywhere; they contain discontinuities, jumps, spikes, etc. Indeed, the "singularities" (or non-smooth points) can be the most interesting and informative aspects of the functions.

Example 1 *Functions not smooth everywhere.*

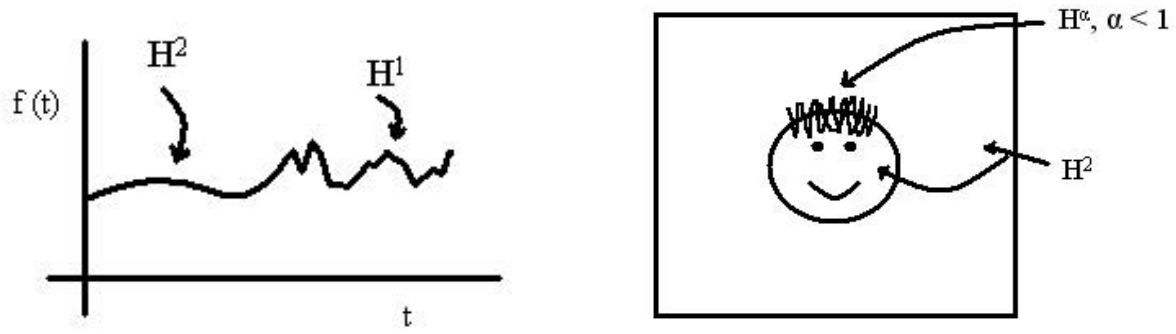


(a) (b)

Figure 1: Example of functions not smooth everywhere. (a) 1-D Case (b) 2-D Case

Furthermore, functions of interest may possess different degrees of smoothness in different regions.

Example 2 *Functions with different degrees of smoothness.*



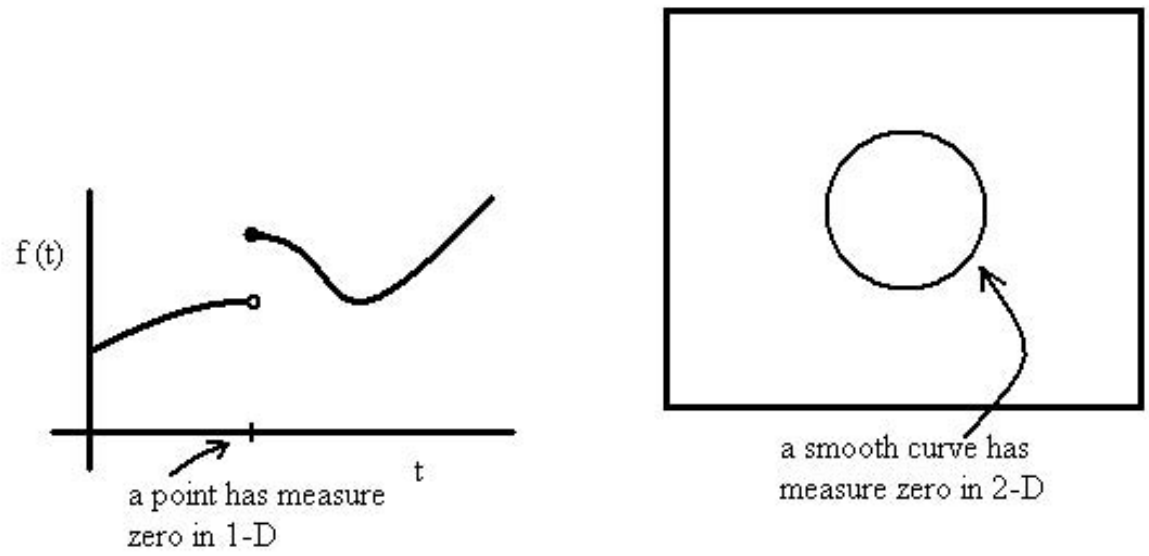
(a) (b)

Figure 2: Example of functions having different degrees of smoothness. (a) 1-D Case (b) 2-D Case

2 NonLinear Approximation via Trees

Let $B^\alpha(C_\alpha)$ denote the set of all functions that are $H^\alpha(C_\alpha)$ everywhere except on a set of measure zero. To simplify the notation, we won't explicitly identify the domain (e.g., $[0,1]$ or $[0,1]^d$); that will be clear from the context.

Example 3 *Sets of measure zero.*



(a) (b)

Figure 3: Sets of measure zero. (a) 1-D Case (b) 2-D Case

Let's consider a 1-D case first.

Let $f \in B^{\alpha}(C_{\alpha})$ and consider approximating f by a piecewise polynomial function on a uniform partition.

If f is Holder-α smooth everywhere, then by using an appropriate partition width k^{-1} and fitting degree $\lceil \alpha \rceil$ polynomials on each interval we have an approximation f_k satisfying

$$|f(t) - f_k(t)| \leq C_{\alpha} k^{-\alpha}$$

and

$$||f - f_k||_{L_2}^2 = O(k^{-2\alpha})$$

.

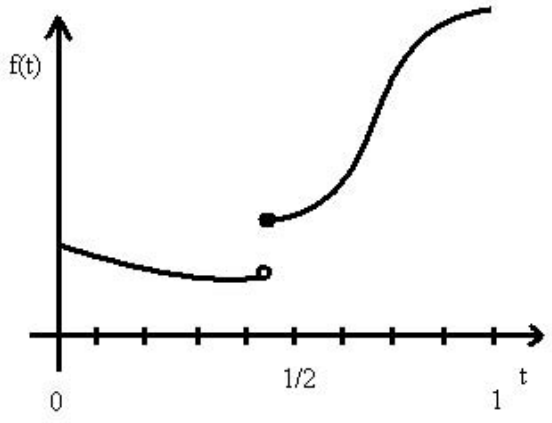


Figure 4: Smooth curve with a discontinuity.

However, if there is a discontinuity then for t in the interval containing the discontinuity the difference

$$|f(t) - f_k(t)|$$

will not be small.

Example 4 *Suppose f is piecewise Lipschitz and f_k ia a piecewise constant.*

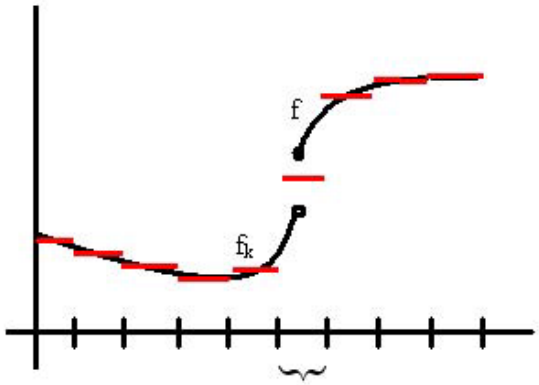


Figure 5:

$$|f(t) - f_k(t)| \approx \Delta$$

where Δ is a constant equal to average of f on right and left side of discontinuity in this interval.

$$\Rightarrow ||f - f_k||_{L_2}^2 = O(k^{-1}$$

where k^{-1} is the width of the interval. Notice this rate is quite slow.

This problem naturally suggests the following remedy: use very small intervals near discontinuities and larger intervals in smooth regions. Specifically, suppose we use intervals of width $k^{-2\alpha}$ to contain the discontinuities and the intervals of width k^{-1} elsewhere. Then accordingly piecewise polynomial approximation $\tilde{f}_k$ satisfies

$$||f - \tilde{f}_k||_{L_2}^2 = O(k^{-2\alpha})$$

.

We can accomplish this need for "adaptive resolution" or "multiresolution" using recursive partitions and trees.

3 Recursive Dyadic Partitions

We discussed this idea already in our examination of classification trees. Here is the basic idea again, graphically.

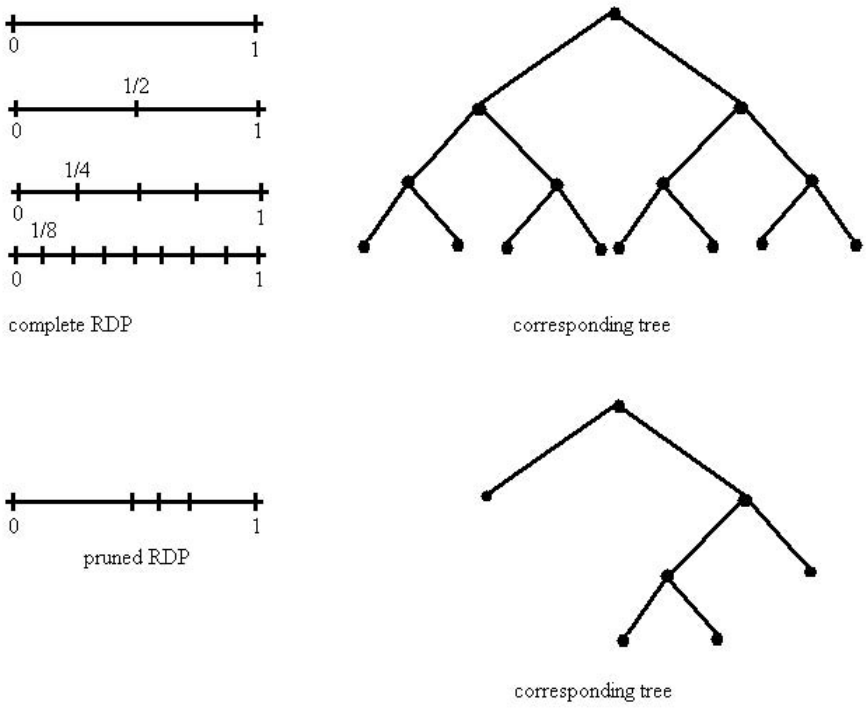


Figure 6: Complete and pruned RDP along with their correspnding tree structures.

Consider a function $f \in B^\alpha(C_\alpha)$ that contains no more than m points of discontinuity, and is $H^\alpha(C_\alpha)$ away from these points.

Lemma 1 *Consider a complete RDP with n intervals, then there exists an associated pruned RDP with $O(klogn)$ intervals, such that an associated piecewise degree $\lceil\alpha\rceil$ polynomial approximation $\tilde{(f)}_k$, has a squared approximation error of $O(min(k^{-2\alpha}, n^{-1}))$.*

Proof : Assume $n > k > m$. Divide $[0,1]$ into k intervals. If f is smooth on a particular interval I, then

$$|f(t) - \tilde{f}_k(t)| = O(k^{-2\alpha}) \forall t \in I$$

.

In intervals that contain a discontinuity, recursively subdivide into two until the discontinuity is contained in an interval of width n^{-1}. This process results in at most $log_2 n$ addition subintervals per discontinuity, and the squared approximation error is $O(k{-}2\alpha)$ on all of them accept the m intervals of width n^{-1} containing the discontinuities where the error is $O(1)$ at each point.

Thus, the overall squared L_2 norm is

$$||f - \tilde{f}_k||^2_{L_2} = O(min(k^{-2\alpha}, n^{-1}))$$

and there are at most $k + log_2 n$ intervals in the partition. Since $k \dot{\iota} m$, we can upperbound the number of intervals by $2klog_2 n$.

Note that if the initial complete RDP has $n \approx k^{2\alpha}$ intervals, then the squared error is $O(k^{-2\alpha})$.

Thus, we only incur a factor of $2\alpha log k$ additional leafs and achieve the same overall approximation error as in the $H^\alpha(C_\alpha)$ case. We will see that this is a small price to pay in order to handle not only smooth functions, but also piecewise smooth functions.

4 Wavelet Approximations

Let $f \in L^2([0,1])$; $\int f^2(t)dt < \infty$.

A wavelet approximation is a series of the form

$$f = c_o + \sum_{j \geq 0} \sum_{k=1}^{2^j} < f, \psi_{j,k} > \psi_{j,k}$$

where c_o is a constant ($c_o = \int_0^1 f(t)dt$),

$$< f, \psi_{j,k} >= \int_0^1 f(t)\psi_{j,k}(t)dt$$

and the basis functions $\psi_{j,k}$ are orthonormal, oscillatory signals, each with an associated scale 2^{-j} and position $k2^{-j}$. $\psi_{j,k}$ is called the wavelet at scale 2^{-j} and position $k2^{-j}$.

Example 5 Haar Wavelets

$$\psi_{j,k}(t) = 2^{j/2}(\mathbf{1}_{\{t \in [2^{-j}(k-1), 2^{-j}(k-1/2)]\}} - \mathbf{1}_{\{t \in [2^{-j}(k-1/2), 2^{-j}k]\}})$$

$$\int_0^1 \psi_{j,k}(t)dt = 0$$

$$\int_0^1 \psi^2_{j,k}(t)dt = \int_{(k-1)2^{-j}}^{k2^{-j}} 2^j dt = 1$$

$$\int_0^1 \psi_{j,k}(t)\psi_{l,m}(t)dt = \delta_{j,l}.\delta_{k,m}$$

Note: *If f is constant on $[2^{-j}(k-1), 2^{-j}k]$, then*

$$\int f\psi_{j,k}(t) = 0$$

.

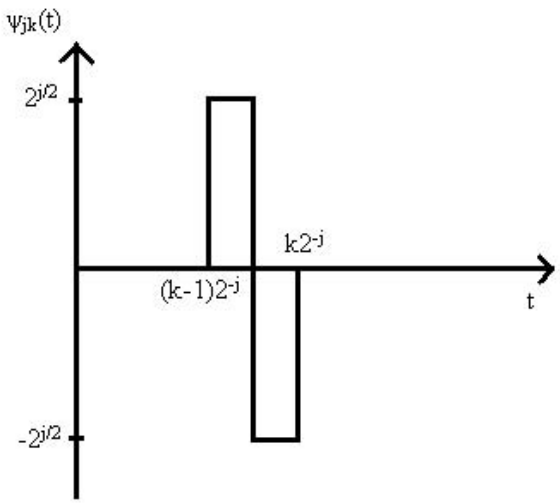


Figure 7: Haar Wavelet

Suppose f is piecewise constant with at most m discontinuities. Let

$$f_J = c_o + \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} < f, \psi_{j,k} > \psi_{j,k}$$

. Then, f_J has at most mJ non-zero wavelet coefficients; i.e., $< f, \psi_{j,k} >= 0$ for all but mJ terms, since at most one Haar Wavelet at each scale senses each point of discontinuity. Said another way, all but at most m of the wavelets at each scale have support over constant regions of f.

f_J itself will be piecewise constant with discontinuities only possible occurring at end points of the intervals $[2^{-J}(k-1), 2^{-J}k]$. Therefore, in this case

$$||f - f_J||_{L_2}^2 = O(2^{-J}).$$

Daubechies wavelets are the extension of the Haar wavelet idea. Haar wavelets have one "vanishing moment":

$$\int_0^1 \psi_{j,k} = 0.$$

Daubechies wavelets are "smoother" basis functions with extra vanishing moments. The Daubechies-N wavelet has N vanishing moments.

$$\int_0^1 t^l \psi_{j,k} dt = 0 for l = 0, 1, ..., N-1.$$

The Daubechies-1 wavelet is just the Haar case.

If f is a piecewise degree $\leq N$ polynomial with at most m pieces, then using the Daubechies-N wavelet system.

$$||f - f_J||_{L_2}^2 = O(2^{-J});$$

and

$$f_J(t) = c_o + \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} < f, \psi_{j,k} > \psi_{j,k}(t)$$

has at most $O(mJ)$ non-zero wavelet coefficients. f_J is called the **Discrete Wavelet Transform (DWT)** approximation of f. The key idea is the same as we saw with trees.

5 Sampled Data

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

$$\underline{f} = [f(1/n), f(2/n), ..., f(n/n)]$$

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

$$\underline{f} = c_o + \sum_{j=0}^{log_2 n-1} \sum_{k=1}^{2^j} < \underline{f}, \underline{\psi}_{j,k} > \underline{\psi}_{j,k}$$

where

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

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

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

for the Daubechies-N discrete wavelets.

$$< \underline{f}, \underline{\psi}_{j,k} >= \underline{f}^T \underline{\psi}_{j,k}$$

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

6 Approximating B^α functions with wavelets

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

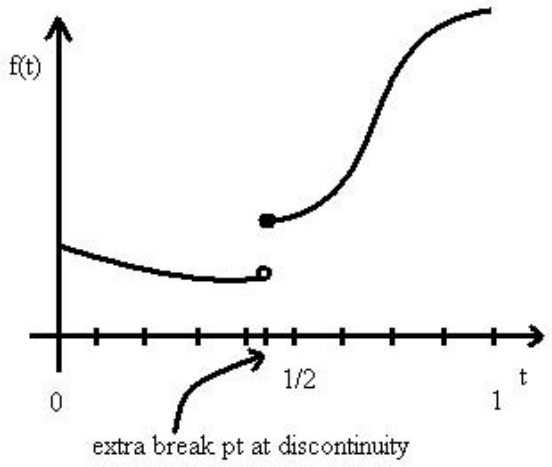


Figure 8:

Then

$$|f(t) - f_p(t)|^2 = O(k^(- 2\alpha)) \forall t \in [0, 1]$$
$$\Rightarrow |f(i/n) - f_p(i/n)|^2 = O(k^{-2\alpha}) i = 1, ..., n$$
$$\Rightarrow 1/n ||\underline{f} - \underline{f}_p||^2_{L_2} = O(k^{-2\alpha}))$$

and $\underline{f}_p$ has $O(klog_2 n)$ non-zero coefficients according to our previous analysis.

7 Wavelets in 2-D

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

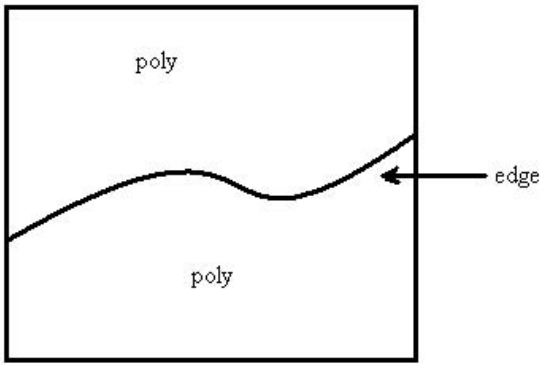


Figure 9:

A pruned RDP of k squares decorated with polyfits gives

$$||f - f_k||_{L_2}^2 = O(k^{-1})$$

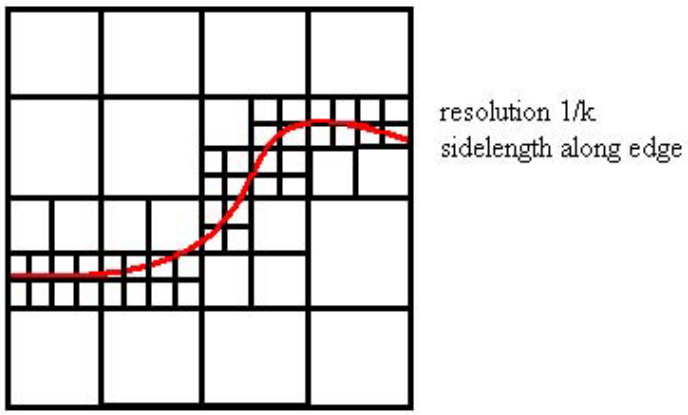


Figure 10:

Let $\underline{f} = [f(i/k, j/k)_{i,j=1}^n$ sample range.

$$f_n(t) = \sum_{i,j=1}^{k} f(i/k, j/kk) \mathbf{1}_{\{t \in [i-1/k, i/k) x [j-1/k, j/k)\}}$$

then

$$||f - f_n||_{L_2}^2 = O(k^{-1})$$

$O(1)$ error on k of the k^2 pixels, near zero elsewhere. The DWT of $\underline{f}$ has $O(k)$ non-zero wavelet coefficients. $O(2^j)$ at scale $2^{-j}, j = 0, 1, ..., logn$.