ECE 830 Fall 2010 Statistical Signal Processing

instructor: R. Nowak , **scribe:** C. Hall

Lecture 25: Nonparametric Signal Estimation

Nonparametric Signal Estimation Setup

Recall in previous lectures we considered parametric signal estimation or "denoising" problems
Parametric Signal Estimation Setting

$$x = \underbrace{H\theta}_{f} + w \ , \ w \sim \mathcal{N}(0, I)$$

$$H_{n\times k} \ known$$

$$\theta_{k\times 1} \ unknown \, k \leq n$$

The signal is a vector in $\mathbb{R}^n$ that is described by $k \leq n$ parameters.
Now we are investigating a nonparametric version of this problem. Suppose we collect noisy samples of a function $f : [0,1] \to \mathbb{R}$, the setting becomes
Nonparametric Signal Estimation Setting

$$x_i = f(t_i) + w_i \ , \ i = 1, ..., n$$

Where:
$f : [0,1] \to \mathbb{R}$, *unknown*
$t_1, t_2, ..., t_n$ (the sampling locations) are uniformly spaced on the unit interval; e.g. $t_i = \frac{i-1}{n}$
w_i are iid noise, with $\mathbb{E}[w_i] = 0$ and $\mathbb{E}[w_i^2] = \sigma^2$, but otherwise unknown distribution

We know from classical Shannon-Nyquist sampling theory that the spacing between samples must be inversely proportional to the highest frequency of f. In other words the sampling rate should be inversely proportional to the "wiggliness" or "roughness" of the signal, the smoother the signal the fewer samples are needed. Sample signals are reconstructed by interpolating between the sampled values. For example, linear or polynomial interpolation is quite common. The classic theory doesn't address how the interpolation should be modified if noise is present in the samples, the topic of this lecture.

Hölder Smoothness

Since linear or polynomial interpolation is commonly used, that is the approach we will adopt. It is natural to ask: what types of signals or functions can be accurately interpolated/approximated by polynomials?

Recall the definition of Lipschitz Continuous, Equation 1.

$$|f(t) - f(s)| \leq L|t - s| \tag{1}$$

$$|f'(t) - f'(s)| \leq L|t - s| \tag{2}$$

Equation 1 implies it is Hölder 1 smooth with a Hölder constant $\alpha = 1$, if Equations 1 and 2 both hold then it is Hölder 2 smooth with a Hölder constant $\alpha = 2$, and so it continues as we take derivatives. More formally we say:

Definition A function $f : [0,1] \to \mathbb{R}$ with k continuous derivatives is said to be Hölder smooth with parameter and constant $L_\alpha \geq 0$ if

$$|f(t) - p(t;t_0)| \leq L_\alpha |t - t_0|^\alpha$$

where $p(t;t_0)$ is the degree k Taylor series approximation to f at t_0, and $k = \lceil \alpha \rceil - 1$

Examples

$$\alpha = 1 \Rightarrow k = 0, \ \textit{Lipschitz smoothness}$$

$$\alpha = 2 \Rightarrow k \geq 1 \ \textit{and linear (degree 1) approximation}$$

Smoother $f \Leftrightarrow$ *Larger* α

Approximating Hölder Smooth Functions

A Hölder α-smooth function can be well approximated by a piecewise polynomial function as follows. Divide the interval [0,1) into m disjoint subintervals,

$$\left[0, \frac{1}{m}\right), \left[\frac{1}{m}, \frac{2}{m}\right), ..., \left[\frac{m-1}{m}, 1\right)$$

Denote the j^{th} subinterval $I_j := \left[\frac{j-1}{m}, \frac{j}{m}\right)$. Let p(t;t') be the degree $k = \lceil \alpha \rceil - 1$ Taylor polynmial of f at some $t' \in I_j$. Then

$$|f(t) - p(t;t')| \leq L_\alpha |t - t'|$$
$$\leq L_\alpha m^{-\alpha} \ , \ \forall \ t, t' \in I_j$$

Now consider the sample points $t_i \in I_j$ There are $\frac{n}{m}$ sample points in I_j. Let p_j denote the polynomial of degree k that fits best to these points; i.e.,

$$p_j = arg \min_{p \in poly(k)} \frac{1}{n/m} \sum_{i:t_i \in I_j} |f(t_i) - p(t_i)|^2 = arg \min_{\theta \in \mathbb{R}^k} \frac{1}{n/m} \sum_{i:t_i \in I_j} |f(t_i) - \sum_{\ell=0}^{k} \theta_\ell t_i^\ell|^2$$

Then

$$|f(t) - p_j(t)| \leq L_\alpha m^{-\alpha} \ , \ \forall t \in I_j$$

p_j has a simple parametric form

$$p_j(t) = \theta_{0j} + \theta_{1j} t + \ldots + \theta_{kj} t^k = \theta_j^T v$$

where

$$\theta_j = \begin{bmatrix} \theta_{0j} \\ \theta_{1j} \\ \vdots \\ \theta_{kj} \end{bmatrix} \quad v = \begin{bmatrix} 1 \\ t \\ \vdots \\ t^k \end{bmatrix}$$

$$\theta_j = arg \min_{\theta \in \mathbb{R}^{k+1}} \sum_{i:t_i \in I_j} |f(t_i) - \theta^T v_i|^2$$

We can express this in matrix-vector notation.
Let f_j be a vector of $\frac{n}{m}$ samples of $f(t_i)$, $t_i \in I_j$.
Let V_j be the Vandermonde matrix with rows $\{v_i^T\}_{i:t_i \in I_j}$.
Then

$$\theta_j = arg \max_{\theta \in \mathbb{R}^{k+1}} ||f_j - V_j \theta||_2^2$$

$$(V_j^T V_j)^{-1} V_j^T f_j \ , \ if \ V_j \ has \ full-rank$$

Fact: V_j has full-rank iff $\frac{n}{m} \geq k+1$
Now consider the piecewise polynomial approximation

$$\bar{f}(t) = \sum_{j=1}^{m} p_j(t) 1_{\{t \in I_j\}}$$

The L_2 error of this approximation is

$$||f - \bar{f}||_2^2 = \int_0^1 |f(t) - \bar{f}(t)| dt = \sum_{j=1}^{m} \int_{I_j} |f(t) - p_j(t)|^2 dt \leq \sum_{j=1}^{m} \int_{I_j} L_\alpha^2 m^{-2\alpha} dt = L_\alpha^2 m^{-2\alpha}$$

$$\frac{n}{m} \geq k+1 \Rightarrow m \leq \frac{n}{k+1}$$

Estimating a Hölder Smooth Function from Noisy Data

To estimate f from data

$$x_i = f(t_i) + w_i \ , \ i = 1, ..., n \ , \ t_i = \frac{i-1}{n}$$

We will assume that the noises are iid with $\mathbb{E}[w] = 0$, $\mathbb{E}[w_i^2] = \sigma^2$. We will make no further assumptions about the noise distribution.
Here is our approach. We will "fit" a polynomial of degree $\lceil \alpha \rceil - 1$ to the observations falling in each of the subintervals. On subinterval I_j we obtain

$$\hat{\theta}_j := \min_{\theta \in \mathbb{R}^{k+1}} \frac{1}{n_j} \sum_{i: t_i \in I_j} |x_i - p_\theta(t_i)|^2$$

$$= \min_{\theta} \frac{1}{n_j} \sum_{i: t_i \in I_j} |x_i - \theta^T V_i|^2$$

where $n_j = \# t_i$ in I_j $= \frac{n}{m}$ and $v_i = \begin{bmatrix} 1 \\ t_i \\ t_i^2 \\ \vdots \\ t_i^k \end{bmatrix}$

This has a simple solution. Let x_j be a vector of the samples $\{x_i\}_{i: t_i \in I_j}$ and let V_j be the Vandermonde matrix with rows $\{v_i^T\}_{i: t_i \in I_j}$ Then

$$\hat{\theta} = \min_{\theta \in \mathbb{R}^{k+1}} ||x_j - V_j \theta||_2^2$$

$$= (V_j^T V_j)^{-1} V_j^T x_j$$

assuming the matrix V_j is full-rank.
Fact: V_j has full-rank iff $\frac{n}{m} \geq k+1$
Let $\hat{p}_j(t) := \hat{\theta}_j^T v = \hat{\theta}_{0j} + \hat{\theta}_{1j} t + ... + \hat{\theta}_{dj} t^d$
and define our estimator to be

$$\hat{f}(t) := \sum_{j=1}^{m} \hat{p}_j(t) 1_{\{t \in I_j\}}$$

Note:

$$\mathbb{E}[\hat{\theta}_j] = (V_j^T v_j)^{-1} V_j^T \mathbb{E}[x_j]$$

$$= (V_j^T V_j)^{-1} V_j \begin{bmatrix} f(t_{i1}) \\ \vdots \\ f(t_{in_j}) \end{bmatrix} = \theta_j$$

$$\Rightarrow \mathbb{E}[\hat{p}_j] = p_j$$

Bounding the Error (MSE)

The error we would like to bound is

$$\mathbb{E}[||f - \hat{f}||_2^2] = \mathbb{E}\left[\int |f(t) - \hat{f}(t)|^2 dt\right] = \mathbb{E}\left[\sum_{j=1}^{m} \int_{I_j} |f(t) - \hat{p}_j(t)|^2 dt\right]$$

Let $\bar{p_j}(t) := \mathbb{E}[\hat{p}_j(t)] = \mathbb{E}[\hat{\theta}_{0j}] + \mathbb{E}[\hat{\theta}_{1j}]t + ... + \mathbb{E}[\hat{\theta}_{dj}]t^d$ and $\bar{f}(t) = \sum_{j=1}^m \bar{p}_j(t) 1_{\{t \in I_j\}}$.
Decompose the error as follows:

$$\mathbb{E}[||f - \hat{f}||_2^2] = \mathbb{E}[||f - \bar{f} + \bar{f} - \hat{f}||_2^2] \leq ||f - \bar{f}||_2^2 + 2\underbrace{\mathbb{E}\left[\int_{[0,1]} |f - \bar{f}||\bar{f} - \hat{f}| dt\right]}_{=0 \ since \ \mathbb{E}[\hat{f}] = \bar{f}} + \mathbb{E}[||\bar{f} - \hat{f}||_2^2]$$

$$= ||f - \bar{f}||_2^2 + \mathbb{E}[||\bar{f} - \hat{f}||_2^2]$$

$$\leq L_\alpha^2 m^{-2\alpha} + \mathbb{E}[||\bar{f} - \hat{f}||_2^2]$$

Therefore

$$\mathbb{E}[||f - \hat{f}||_2^2] \leq L_\alpha^2 m^{-2\alpha} + \mathbb{E}[||\bar{f} - \hat{f}||_2^2] \tag{3}$$

Bounding $\mathbb{E}[||\bar{f} - \hat{f}||_2^2]$

$$\mathbb{E}[||\bar{f} - \hat{f}||_2^2] = \mathbb{E}\left[\sum_{j=1}^{m} \int_{I_j} |p_j - \hat{p}_j|^2 dt\right]$$

$$= \mathbb{E}\left[\sum_j \int_{I_j} |(\theta_j - \hat{\theta}_j)^T v|^2 dt\right]$$

$$\leq \sum_j \mathbb{E}\left[\int_{I_j} ||\theta_j - \hat{\theta}_j)||_2^2 \, ||v||_2^2 dt\right] \ , \ by \ applying \ Cauchy - Schwarz$$

$$\leq \sum_j^m \int_{I_j} ||v||_2^2 dt \, \mathbb{E}\left[||\theta_j - \hat{\theta}_j||_2^2\right]$$

Since $\hat{\theta}_j$ is an unbiased estimator of θ_j

$$\mathbb{E}\left[||\theta_j - \hat{\theta}_j||_2^2\right] = var(\hat{\theta}) \leq C_1' \frac{k+1}{n/m}$$

where $C_1' > 0$ is a constant depending on v_j and σ^2. Therefore

$$||\bar{f} - \hat{f}||_2^2 \leq \sum_j \int_{I_j} C_1' \frac{m(k+1)}{n} ||v||_2^2 dt$$

$$\leq C_1 \frac{m(k+1)}{n} \ , \ for \ some \ C_1 > 0 \ since \int_{[0,1]} ||v||_2^2 dt = constant$$

Final Bound

$$\mathbb{E}||f - \hat{f}||_2^2 = ||f - \bar{f}||_2^2 + \mathbb{E}[||\bar{f} - \hat{f}||_2^2$$
$$\leq L_\alpha^2 m^{-2\alpha} + C_1 \frac{m(k+1)}{n}$$
$$= L_\alpha^2 m^{-2\alpha} + C_2 \frac{m}{n}$$

Taking $m = n^{+\frac{1}{2\alpha+1}}$ yields

$$\mathbb{E}[||f - \hat{f}||_2^2 \leq C \times n^{-\frac{2\alpha}{2\alpha+1}} \;, \; for\ some\ C > 0 \tag{4}$$

Note that as smoothness α increases so does the rate of convergence.
This analysis is easily extended to Hölder smooth functions on $[0,1]^d$
<u>Estimating d-dimensional Hölder Smooth Functions</u>
if:
  $f[0,1]^d \rightarrow \mathbb{R}$, is a Hölder α-smooth function
then:
  n noiseless samples$\rightarrow ||f - \bar{f}||_2^2 = \int_{[0,1]^d} |f(t) - \bar{f}(t)|^2 dt \leq Cn^{-\frac{2\alpha}{d}} \;, \; C > 0$
  n noisy samples$\rightarrow \mathbb{E}[||f - \hat{f}||_2^2] \leq Cn^{-\frac{-2\alpha}{2\alpha+d}}$, $C > 0$
So the final error bound is:

$$\mathbb{E}[||f - \hat{f}||_2^2] \leq Cn^{-\frac{2\alpha}{2\alpha+d}}$$

Thus we see that the "blessing of smoothness" can offset the "curse of dimensionality"