Local Polynomial Approximation with Adaptive Support: The Intersection of Confidence Intervals Rule

Lecture Notes on Advanced Signal Processing

July 16, 2025

Contents

1 Introduction to Adaptive Local Polynomial Approximation

1.1 Motivation and Overview

The Local Polynomial Approximation (LPA) represents a fundamental technique in non-parametric regression and signal processing. While traditional LPA methods employ fixed-size support regions, this lecture extends the framework to incorporate *adaptive support selection* through the Intersection of Confidence Intervals (ICI) rule, addressing the fundamental bias-variance trade-off inherent in statistical estimation.

Definition 1.1 (Signal Model)**.** *Consider a discrete signal model:*

$$y_i = f(x_i) + \eta_i, \quad i = 1, 2, \ldots, n \tag{1}$$

where:

- $f : \mathbb{R} \to \mathbb{R}$ *represents the unknown ground truth signal*
- $\eta_i \sim \mathcal{N}(0, \sigma^2)$ *denotes independent and identically distributed (i.i.d.) Gaussian noise*
- y_i *constitutes the observed noisy measurements*

The central challenge in signal estimation involves recovering $f(x)$ from the noisy observations $\{y_i\}_{i=1}^n$ while optimally balancing estimation accuracy and stability.

1.2 Fixed-Support LPA: Review and Limitations

In the standard LPA framework, estimation of $f(x_0)$ proceeds by fitting a polynomial of degree p within a fixed interval $[x_0 - h, x_0 + h]$:

$$\hat{f}(x) = \sum_{j=0}^{p} \beta_j (x - x_0)^j \tag{2}$$

The coefficients $\{\beta_j\}_{j=0}^p$ are obtained through weighted least squares:

$$\{\hat{\beta}_j\} = \arg\min_{\{\beta_j\}} \sum_{i:|x_i - x_0| \le h} w_i \left(y_i - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j \right)^2 \tag{3}$$

where w_i represents the weight assigned to observation i, typically chosen as a kernel function $K((x_i - x_0)/h)$ that emphasizes observations near x_0.

Remark 1.2. *The estimate $\hat{f}(x_0) = \hat{\beta}_0$ can be expressed as a linear combination of observations:*

$$\hat{f}(x_0) = \sum_{i=1}^{n} h_i(x_0) y_i = \mathbf{h}(x_0)^T \mathbf{y} \tag{4}$$

where $\mathbf{h}(x_0)$ represents the equivalent kernel arising from the polynomial fitting procedure.

2 Statistical Properties and the Bias-Variance Decomposition

2.1 Mean Squared Error Analysis

The performance of any estimator $\hat{f}(x_0)$ is commonly quantified through the Mean Squared Error (MSE):

Theorem 2.1 (Bias-Variance Decomposition)**.** *For any estimator $\hat{f}(x_0)$ of $f(x_0)$, the MSE admits the decomposition:*

$$\mathrm{MSE}[\hat{f}(x_0)] = \mathbb{E}[(\hat{f}(x_0) - f(x_0))^2] = \mathrm{Bias}^2[\hat{f}(x_0)] + \mathrm{Var}[\hat{f}(x_0)] \tag{5}$$

where:

$$\mathrm{Bias}[\hat{f}(x_0)] = \mathbb{E}[\hat{f}(x_0)] - f(x_0) \tag{6}$$

$$\mathrm{Var}[\hat{f}(x_0)] = \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] \tag{7}$$

Proof. Starting from the MSE definition:

$$\begin{aligned} \mathrm{MSE}[\hat{f}(x_0)] &= \mathbb{E}[(\hat{f}(x_0) - f(x_0))^2] && (8) \\ &= \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)] - f(x_0))^2] && (9) \end{aligned}$$

Adding and subtracting $\mathbb{E}[\hat{f}(x_0)]$:

$$\begin{aligned} &= \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2] + 2\mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])(\mathbb{E}[\hat{f}(x_0)] - f(x_0))] && (10) \\ &\quad + (\mathbb{E}[\hat{f}(x_0)] - f(x_0))^2 && (11) \end{aligned}$$

Since $\mathbb{E}[\hat{f}(x_0)] - f(x_0)$ is deterministic:

$$= \mathrm{Var}[\hat{f}(x_0)] + 2(\mathbb{E}[\hat{f}(x_0)] - f(x_0))\mathbb{E}[\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]] + \mathrm{Bias}^2[\hat{f}(x_0)] \tag{12}$$

The middle term vanishes as $\mathbb{E}[\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]] = 0$:

$$= \mathrm{Var}[\hat{f}(x_0)] + \mathrm{Bias}^2[\hat{f}(x_0)] \tag{13}$$

□

2.2 Variance Calculation for Linear Estimators

For the LPA estimator expressed as in Equation (4), we derive the variance explicitly:

Lemma 2.2 (Variance of LPA Estimator)**.** *Under the signal model (1) with i.i.d. Gaussian noise $\eta_i \sim \mathcal{N}(0, \sigma^2)$:*

$$\mathrm{Var}[\hat{f}(x_0)] = \sigma^2 \|\mathbf{h}(x_0)\|^2 = \sigma^2 \sum_{i=1}^{n} h_i^2(x_0) \tag{14}$$

Proof. Starting from the linear representation:

$$\operatorname{Var}[\hat{f}(x_0)] = \operatorname{Var}\left[\sum_{i=1}^{n} h_i(x_0) y_i\right] \tag{15}$$

$$= \operatorname{Var}\left[\sum_{i=1}^{n} h_i(x_0)(f(x_i) + \eta_i)\right] \tag{16}$$

Since $f(x_i)$ is deterministic and η_i are independent:

$$= \sum_{i=1}^{n} h_i^2(x_0) \operatorname{Var}[\eta_i] \tag{17}$$

$$= \sigma^2 \sum_{i=1}^{n} h_i^2(x_0) \tag{18}$$

□

3 The Adaptive Support Problem

3.1 Support Size Impact on Estimation Quality

The choice of support size h critically affects both bias and variance components:

Definition 3.1 (Support-Dependent Bias and Variance). *For a given polynomial degree p and support size h, define:*

$$B(h, x_0) = \mathbb{E}[\hat{f}_h(x_0)] - f(x_0) \quad \textit{(Bias)} \tag{19}$$

$$V(h, x_0) = \mathrm{Var}[\hat{f}_h(x_0)] \quad \textit{(Variance)} \tag{20}$$

Theorem 3.2 (Asymptotic Behavior). *Under suitable regularity conditions on f and the kernel weights:*

$$B(h, x_0) = O(h^{p+1}) \quad \textit{as } h \to 0 \tag{21}$$

$$V(h, x_0) = O(h^{-1}) \quad \textit{as } h \to 0 \tag{22}$$

This theorem reveals the fundamental trade-off: decreasing h reduces bias but increases variance, while increasing h has the opposite effect. The optimal choice depends on the local smoothness of f around x_0.

3.2 Signal-Adaptive Support Selection

Consider a signal with varying local characteristics:

- **Smooth regions**: Large derivatives are negligible, permitting larger support sizes without significant bias
- **High-frequency regions**: Rapid variations necessitate smaller support to capture local behavior accurately
- **Discontinuities**: Require asymmetric or minimal support to avoid crossing the discontinuity

Remark 3.3. *The optimal support size $h^*(x_0)$ that minimizes $\mathrm{MSE}[\hat{f}_h(x_0)]$ satisfies:*

$$\frac{\partial}{\partial h}\left[B^2(h, x_0) + V(h, x_0)\right] = 0 \tag{23}$$

This yields a location-dependent optimal support, motivating adaptive methods.

4 The Intersection of Confidence Intervals (ICI) Rule

4.1 Theoretical Foundation

The ICI rule provides a data-driven approach to support selection based on statistical confidence intervals:

Definition 4.1 (Confidence Interval for LPA). *For a given support size h and confidence parameter $\gamma > 0$, the confidence interval for $\hat{f}_h(x_0)$ is:*

$$CI_h(x_0) = \left[\hat{f}_h(x_0) - \gamma\sqrt{V(h, x_0)}, \hat{f}_h(x_0) + \gamma\sqrt{V(h, x_0)}\right] \tag{24}$$

For Gaussian noise with known variance σ^2, choosing $\gamma = z_{\alpha/2}$ (the $\alpha/2$ quantile of the standard normal distribution) yields a $(1 - \alpha)$ confidence interval.

4.2 The ICI Algorithm

Theorem 4.2 (ICI Principle). *Consider a sequence of support sizes $h_1 < h_2 < \cdots < h_K$. Define:*

$$\mathcal{I}_k(x_0) = \bigcap_{j=1}^{k} CI_{h_j}(x_0) \tag{25}$$

The optimal scale $k^*(x_0)$ is chosen as:

$$k^*(x_0) = \max\{k : \mathcal{I}_k(x_0) \neq \emptyset\} \tag{26}$$

Input: Signal $\{y_i\}_{i=1}^n$, polynomial degree p, scales $\{h_k\}_{k=1}^K$, parameter γ
Output: Optimal scales $\{k^*(x_i)\}_{i=1}^n$ and estimates $\{\hat{f}(x_i)\}_{i=1}^n$

1. **Initialization:**
   - Set $L_0(x_i) = -\infty$, $U_0(x_i) = +\infty$ for all i
   - Set $k^*(x_i) = 0$ for all i
2. **For each scale** $k = 1, 2, \ldots, K$:
   (a) Compute LPA estimates $\hat{f}_{h_k}(x_i)$ for all i
   (b) Compute variances $V(h_k, x_i) = \sigma^2 \sum_j h_{j,k}^2(x_i)$
   (c) Calculate confidence bounds:
   $$l_k(x_i) = \hat{f}_{h_k}(x_i) - \gamma\sqrt{V(h_k, x_i)} \tag{27}$$
   $$u_k(x_i) = \hat{f}_{h_k}(x_i) + \gamma\sqrt{V(h_k, x_i)} \tag{28}$$
   (d) Update intersection bounds:
   $$L_k(x_i) = \max\{L_{k-1}(x_i), l_k(x_i)\} \tag{29}$$
   $$U_k(x_i) = \min\{U_{k-1}(x_i), u_k(x_i)\} \tag{30}$$
   (e) **If** $U_k(x_i) < L_k(x_i)$ and $k^*(x_i) = 0$:
      - Set $k^*(x_i) = k - 1$
3. **Final estimates:** $\hat{f}(x_i) = \hat{f}_{h_{k^*(x_i)}}(x_i)$

5 Extensions and Advanced Topics

5.1 Directional LPA for Discontinuous Signals

When signals contain discontinuities, symmetric kernels produce significant artifacts. The solution involves directional estimation:

Definition 5.1 (Directional Kernels)**.** *Define left and right directional kernels:*

$$h_i^L(x_0) = h_i(x_0) \cdot \mathbf{1}_{x_i \leq x_0} \quad \textit{(Left kernel)} \tag{31}$$

$$h_i^R(x_0) = h_i(x_0) \cdot \mathbf{1}_{x_i > x_0} \quad \textit{(Right kernel)} \tag{32}$$

5.2 Variance-Based Aggregation

Given multiple estimates $\{\hat{f}^{(m)}(x_0)\}_{m=1}^M$ with variances $\{V^{(m)}(x_0)\}_{m=1}^M$:

Theorem 5.2 (Optimal Linear Aggregation)**.** *The minimum-variance linear combination is:*

$$\hat{f}_{agg}(x_0) = \sum_{m=1}^{M} w_m(x_0)\hat{f}^{(m)}(x_0) \tag{33}$$

where the optimal weights are:

$$w_m(x_0) = \frac{1/V^{(m)}(x_0)}{\sum_{j=1}^M 1/V^{(j)}(x_0)} \tag{34}$$

Proof. The variance of the aggregated estimator is:

$$\text{Var}[\hat{f}_{agg}(x_0)] = \sum_{m=1}^{M} w_m^2(x_0)V^{(m)}(x_0) \tag{35}$$

Minimizing subject to $\sum_m w_m = 1$ yields the result via Lagrange multipliers. □

5.3 Practical Considerations

1. **Parameter Selection:**
   - Polynomial degree: $p \in \{0, 1, 2\}$ typically suffices
   - Scale progression: $h_k = h_1 \cdot \rho^{k-1}$ with $\rho \in [2, 3]$
   - Confidence parameter: $\gamma \in [2, 3]$ balances adaptation and stability
2. **Boundary Handling:** Near boundaries, support sizes must be reduced to maintain sufficient observations, naturally handled by the ICI rule.
3. **Computational Efficiency:** Pre-compute kernel matrices for each scale to enable efficient implementation.

6 Conclusion

The ICI rule provides a principled, data-driven approach to adaptive support selection in LPA, automatically balancing bias and variance based on local signal characteristics. This framework extends naturally to multidimensional signals and various kernel designs, forming a cornerstone of modern non-parametric estimation theory.

Remark 6.1 (Future Directions)**.** *Extensions include:*

- *Multiscale decompositions with wavelet-like properties*
- *Robust estimation under non-Gaussian noise*
- *Online/recursive implementations for streaming data*
- *Applications to image processing and computer vision*