This is a draft version of work in progress, content will be revisited in subsequent versions.

A General Framework for Robust Local Polynomial Regression with Similarity Kernels

Yaniv Shulman

yaniv@shulman.info

Abstract

Local Polynomial Regression (LPR) is a widely used nonparametric method for modeling complex relationships due to its flexibility and simplicity. It estimates a regression function by fitting low-degree polynomials to localized subsets of the data, weighted by proximity. However, traditional LPR is sensitive to outliers and high-leverage points, which can significantly affect estimation accuracy. This paper revisits the kernel function used to compute regression weights and proposes a novel framework that incorporates both predictor and response variables in the weighting mechanism. The focus of this work is a conditional density kernel that robustly estimates weights by mitigating the influence of outliers through localized density estimation. A related joint density kernel is also discussed in an appendix. The proposed method is implemented in Python and is publicly available at `https://github.com/yaniv-shulman/rsklpr`, demonstrating competitive performance in synthetic benchmark experiments. Compared to standard LPR, the proposed approach consistently improves robustness and accuracy, especially in heteroscedastic and noisy environments, without requiring multiple iterations. This advancement provides a promising extension to traditional LPR, opening new possibilities for robust regression applications.

1. Introduction

Local polynomial regression (LPR) is a powerful and flexible statistical technique that has gained increasing popularity in recent years due to its ability to model complex relationships between variables. Local polynomial regression generalizes the polynomial regression and moving average methods by fitting a low-degree polynomial to a nearest neighbors subset of the data at the location. The polynomial is fitted using weighted ordinary least-squares, giving more weight to nearby points and less weight to points farther away. The value of the regression function for the point is then obtained by evaluating the fitted local polynomial using the predictor variable value for that data point. LPR has good accuracy near the boundary and performs better than all other linear smoothers in a minimax sense [2]. The biggest advantage of this class of methods is not requiring a prior specification of a function i.e. a parameterized model. Instead, only a small number of hyperparameters need to be specified such as the type of kernel, a smoothing parameter and the degree of the local polynomial. The method is therefore suitable for modeling complex processes such as non-linear relationships, or complex dependencies for which no theoretical models exist. These two advantages, combined with the simplicity of the method, makes it one of the most attractive of the modern regression methods for applications that fit the general framework of least-squares regression but have a complex deterministic structure.

Local polynomial regression incorporates the notion of proximity in two ways. The first is that a smooth function can be reasonably approximated in a local neighborhood by a simple function such as a linear or low order polynomial. The second is the assumption that nearby points carry more importance in the calculation of a simple local approximation or alternatively, that closer points are more likely to interact in simpler ways than far away points. This is achieved by a kernel which produces values that diminish as the distance between the explanatory variables increase to model stronger relationship between closer points.

Methods in the LPR family include the Nadaraya-Watson estimator [11, 16] and the estimator proposed by Gasser and Müller [9] which both perform kernel-based local constant fit. These were improved on in terms of asymptotic bias by the proposal of the local linear and more general local polynomial estimators [14, 3, 5, 4, 6]. For a review of LPR methods the interested reader is referred to [2].

LPR is however susceptible to outliers, high leverage points and functions with discontinuities in their derivative which often cause an adverse impact on the regression due to its use of least-squares based optimization [15]. The use of unbounded loss functions may result in anomalous observations severely affecting the local estimate. Substantial work has been done to develop algorithms to apply LPR to difficult data. To alleviate the issue [13] employs variable bandwidth to exclude observations for which residuals from the resulting estimator are large. In [3] an iterated weighted fitting procedure is proposed that assigns in each consecutive iteration smaller weights to points that are farther then the fitted values at the previous iteration. The process repeats for a number of iterations and the final values are considered the robust parameters and fitted values. An alternative common approach is to replace the squared prediction loss by one that is more robust to the presence of large residuals by increasing more slowly or a loss that has an upper bound such as the Tukey or Huber loss. These methods however require specifying a threshold parameter for the loss to indicate atypical observations or standardizing the errors using robust estimators of scale [10]. For a recent review of robust LPR and other nonparametric methods see [15, 12]

The main contribution of this paper is to revisit the kernel used to produce regression weights. The simple yet effective idea is to generalize the kernel such that both the predictor and the response are used to calculate weights. Within this framework, a non-negative kernel based on conditional density estimation is proposed that assigns robust weights to mitigate the adverse effect of outliers in the local neighborhood. Note the proposed framework does not preclude the use of robust loss functions, robust bandwidth selectors and standardization techniques. In addition the method is implemented in the Python programming language and is made publicly available. Experimental results on synthetic benchmarks demonstrate that the proposed method achieves competitive results and generally performs better than LOWESS using only a single training iteration.

The remainder of the paper is organized as follows: In Section 2, a brief overview of the mathematical formulation of local polynomial regression is provided. In Section 3, a framework for robust weights and the specific conditional density kernel are proposed. Section 4 provides an analysis of the estimator and a discussion of its properties. In Section 5, implementation notes and experimental results are provided. Finally, in Section 6, the paper concludes with directions for future research.

2. Local Polynomial Regression

This section provides a brief overview of local polynomial regression and establishes the notation subsequently used. We adopt the following standing assumptions: the training data $\mathcal{D}_T = \{(X_i, Y_i)\}_{i=1}^T$ are an i.i.d. sample from a continuous joint density f_{XY}; the error terms ϵ_i satisfy $\mathbb{E}[\epsilon_i|X_i] = 0$ and $\mathbb{E}[\epsilon_i^2|X_i] = \sigma^2(X_i) < \infty$; the density of the predictors $f_X(x)$ is positive in the region of interest; and any kernel function K is a non-negative, symmetric probability density function with finite second moments.

Let (X, Y) be a random pair and $\mathcal{D}_T = \{(X_i, Y_i)\}_{i=1}^T \subseteq \mathcal{D}$ be a training set comprising a sample of T data pairs. Suppose that $(X, Y) \sim f_{XY}$ a continuous density and $X \sim f_X$ the marginal distribution of X. Let $Y \in \mathbb{R}$ be a continuous response and assume a model of the form $Y_i = m(X_i) + \epsilon_i,\ i \in 1, \ldots, T$ where $m(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is an unknown function and ϵ_i are independently distributed error terms having zero mean such that $\mathbb{E}[Y \mid X = x] = m(x)$. There are no global assumptions about the function $m(\cdot)$ other than that it is smooth and that locally it can be well approximated by a low degree polynomial as per Taylor's theorem. The local p-th order Taylor expansion for $x \in \mathbb{R}^d$ near a point X_i yields:

$$m(X_i) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(X_i - x)^j := \sum_{j=0}^{p} \beta_j(x)(X_i - x)^j \tag{1}$$

For notational simplicity, we present the one-dimensional case ($d = 1$). The formulation extends to the multivariate case ($d > 1$) by replacing powers with multi-indices (see, e.g., [7], §3.2). To find an estimate $\hat{m}(x)$ of $m(x)$ the low-degree polynomial is fitted to the N nearest neighbors using weighted least-squares such to minimize the empirical loss $\mathcal{L}_{\text{lpr}}(x; \mathcal{D}_N, h)$:

$$\mathcal{L}_{\text{lpr}}(x; \mathcal{D}_N, h) := \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(X_i - x)^j \right)^2 K_h(X_i - x) \tag{2}$$

where $\beta(x) \in \mathbb{R}^{p+1}$ are the polynomial coefficients to be estimated. The minimizer is

$$\hat{\beta}(x) := \arg\min_{\beta(x)} \mathcal{L}_{\text{lpr}}(x; \mathcal{D}_N, h) \tag{3}$$

Where $K_h(\cdot) = h^{-d}K(\cdot/h)$ is a scaled kernel, $h \in \mathbb{R}_{>0}$ is the bandwidth parameter and $\mathcal{D}_N \subseteq \mathcal{D}_T$ is the subset of N nearest neighbors of x in the training set where the distance is measured on the predictors only. Having computed $\hat{\beta}(x)$ the estimate of $m(x)$ is taken as $\hat{m}(x) = \hat{\beta}_0(x)$. The term kernel carries here the meaning typically used in the context of nonparametric regression i.e. a non-negative real-valued weighting function that is typically symmetric, unimodal at zero, and integrates to one. Higher degree polynomials and smaller N generally increase the variance and decrease the bias of the estimator and vice versa [2]. For derivation of the local constant and local linear estimators for the multidimensional case see [8].

Remark on Nearest Neighbors and Bandwidth.. In the following, the local neighborhood is defined by taking the N nearest neighbors to x. Thus, $\mathcal{D}_N \subseteq \mathcal{D}_T$ contains exactly N points. A distance-based kernel K_h is then used to weight those neighbors. In our implementation, we follow a common practical approach where distances within the neighborhood are first normalized to the interval $[0, 1]$, and then a kernel (e.g., Laplacian) is applied. This effectively makes the bandwidth adaptive to the local density of predictors, combining a fixed-size local subset (via N) with a variable kernel scaling to ensure stable local fits. The asymptotic properties discussed later are conditional on the sequence of nearest-neighbor distances [5].

3. Robust Weights with Similarity Kernels

The main idea presented is to generalize the kernel function used in equation (2) to produce robust weights. This is achieved by using a similarity kernel function defined on the data domain $\mathcal{K}_{\mathcal{D}} : \mathcal{D}^2 \to \mathbb{R}_+$ that enables weighting each point and incorporating information on the data in the local neighborhood in relation to the local regression target (x, y).

The proposed empirical loss function is:

$$\mathcal{L}_{\text{rsk}}(x, y; \mathcal{D}_N, \mathcal{H}) := \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x, y)(X_i - x)^j \right)^2 \mathcal{K}_{\mathcal{D}}\left((x, y), (X_i, Y_i); \mathcal{H}\right) \tag{4}$$

The estimated coefficients are found by minimizing this loss:

$$\hat{\beta}(x, y; \mathcal{D}_N, \mathcal{H}) := \arg\min_{\beta(x,y)} \mathcal{L}_{\text{rsk}}(x, y; \mathcal{D}_N, \mathcal{H}) \tag{5}$$

Where $\mathcal{H}$ is the set of bandwidth parameters. There are many possible choices for such a similarity kernel to be defined within this general framework. However, used as a local weighting function, such a kernel should have the following attributes:

1. Non-negative, $\mathcal{K}_{\mathcal{D}}((x, y), (x', y')) \geq 0$.
2. Symmetry in the inputs, $\mathcal{K}_{\mathcal{D}}((x, y), (x', y')) = \mathcal{K}_{\mathcal{D}}((x', y'), (x, y))$.
3. Tending toward decreasing as the distance in the predictors increases. That is, given a similarity function on the response $s(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}_+$, if $s(y, y')$ indicates high similarity the weight should decrease as the distance between the predictors grows, $s(y, y') > \alpha \implies \mathcal{K}_{\mathcal{D}}((x, y), (x + u, y')) \geq \mathcal{K}_{\mathcal{D}}((x, y), (x + v, y'))\ \ \forall\ ||u|| \leq ||v||$ and some $\alpha \in \mathbb{R}_+$.

In this work a useful non-negative kernel is proposed. Similarly to the usual kernels used in (2), these tend to diminish as the distance between the explanatory variables increases to model stronger relationship between closer points. In addition, the weights produced by the kernels also model the "importance" of the pair (x, y). This is useful for example to down-weight outliers to mitigate their adverse effect on the ordinary least square based regression. Note that for the Reproducing Kernel Hilbert Space (RKHS) interpretation discussed in Section 4, the kernel $\mathcal{K}_{\mathcal{D}}$ must also be positive-definite, but this condition is not required for the main results of this paper. Formally let $\mathcal{K}_{\mathcal{D}}$ be defined as:

$$\mathcal{K}_{\mathcal{D}}\left((x,y),(x',y');\mathcal{H}_1,\mathcal{H}_2\right) = K_1(x,x';\mathcal{H}_1)K_2\left((x,y),(x',y');\mathcal{H}_2\right) \tag{6}$$

Where $K_1 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ and $K_2 : \mathcal{D}^2 \to \mathbb{R}_+$ are non-negative kernels and $\mathcal{H}_1, \mathcal{H}_2$ are the sets of bandwidth parameters. The purpose of K_1 is to account for the distance between a neighbor to the local regression target and therefore may be chosen as any of the kernel functions that are typically used in equation (2). The role of K_2 is to perform robust regression by detecting local outliers in an unsupervised manner and assigning them with lower weights.

The material below gives a kernel-agnostic lemma that shows when the optimisation for the empirical estimator $\hat{\beta}(x)$ is invariant to the (unknown) response value y at the regression location x. A corollary then specialises this result to the conditional-density kernel, which is the focus of this paper.

Lemma 1 (Invariance under separable similarity kernels)**.** *Let the similarity kernel be*

$$\mathcal{K}_{\mathcal{D}}\Big((x,y),(x',y');\mathcal{H}\Big) = K_1\Big(x,x';\mathcal{H}_1\Big)\,K_2\Big((x,y),(x',y');\mathcal{H}_2\Big),$$

with K_1 being any non-negative kernel function on $\mathbb{R}^d \times \mathbb{R}^d$, and let K_2 be separable:

$$K_2\Big((x,y),(x',y');\mathcal{H}_2\Big) = c(x,y)\,w(x',y'), \qquad \textit{where } c(x,y) > 0 \textit{ and } w(x',y') \geq 0.$$

Then the empirical loss (4) *becomes*

$$\mathcal{L}_{rsk}(x,y;\mathcal{D}_N,\mathcal{H}) = c(x,y)\sum_{i=1}^{N}\Big(Y_i - \textstyle\sum_{j=0}^{p}\beta_j(x)(X_i-x)^j\Big)^2 K_1(x,X_i;\mathcal{H}_1)\,w(X_i,Y_i),$$

so the minimiser $\hat{\beta}(x,y)$ with respect to β_j is independent of y and will be denoted $\hat{\beta}(x) = (\hat{\beta}_0(x),\ldots,\hat{\beta}_p(x))^T$. For $\mathcal{K}_{\mathcal{D}}$ to be symmetric (a requirement for a Mercer kernel), we must have $c(x,y) = w(x,y)$. However, symmetry is not required for the minimization problem itself.

Proof. The term $c(x,y)$, which is positive and constant with respect to the summation index i, is a scalar factor multiplying the entire sum. Since scaling an objective by a positive constant does not affect its minimizer, the vector minimizer $\hat{\beta}(x,y)$ is independent of y and thus can be simply denoted as $\hat{\beta}(x)$. □

Conditional Density Kernel

The primary method proposed for K_2 is proportional to the estimated localized conditional marginal distribution of the response variable at the location. This corresponds to choosing the components of a separable K_2 as follows:

$$K_2\Big((x,y),(x',y');\mathcal{H}_2\Big) = \hat{f}_{Y|X}(y \mid x;\mathcal{H}_2)\,\hat{f}_{Y|X}(y' \mid x';\mathcal{H}_2),$$

where $\hat{f}_{Y|X}(\cdot \mid \cdot; \mathcal{H}_2)$ is a kernel conditional-density estimator with bandwidth(s) $\mathcal{H}_2$. The nonparametric conditional density estimation is performed using the Parzen–Rosenblatt window (kernel density estimator):

$$\hat{f}(y \mid x; \mathcal{H}_2) = \hat{f}(x, y; \mathcal{H}_2) / \hat{f}(x; \mathcal{H}_2) \tag{7}$$

$$= \frac{|\mathbf{H}_v|^{-1/2} \sum_{i=1}^{N} K_v \left(\mathbf{H}_v^{-1/2}(v - V_i)\right)}{|\mathbf{H}_x|^{-1/2} \sum_{i=1}^{N} K_x \left(\mathbf{H}_x^{-1/2}(x - X_i)\right)} \tag{8}$$

Where $v = [x, y] \in \mathbb{R}^{d+1}$ is the concatenated vector of the predictors and the response; and $\mathbf{H}_v, \mathbf{H}_x$ are bandwidth matrices.

Corollary 1 (Conditional–density kernel objective)**.** *Choose $K_1(x, x'; \mathcal{H}_1)$ to be a standard kernel for local polynomial regression, such as $K_{h_1}(x - x')$, and let K_2 be the conditional density kernel defined above. Then, we can identify*

$$c(x, y) = \hat{f}_{Y|X}(y \mid x; \mathcal{H}_2), \qquad \textit{and} \qquad w(X_i, Y_i) = \hat{f}_{Y|X}(Y_i \mid X_i; \mathcal{H}_2).$$

Assuming $\hat{f}_{Y|X}(y \mid x; \mathcal{H}_2) > 0$, Lemma 1 yields the simplified weighted least-squares objective for $\hat{\beta}(x)$:

$$\tilde{\mathcal{L}}(x) = \sum_{i=1}^{N} \left(Y_i - \sum_{j=0}^{p} \beta_j(x)(X_i - x)^j \right)^2 K_{h_1}(x - X_i)\, \hat{f}_{Y|X}(Y_i \mid X_i; \mathcal{H}_2),$$

which is the empirical objective function whose properties are analysed in the next section.

Figure 1: Loss function surface, shown as a function of the residual (horizontal axis) and the response variable's value (depth axis). The plot assumes a standard quadratic loss in the residual, a standard normal density for the response (as a proxy for K_2), and excludes the K_1 distance kernel scaling. The vertical axis represents a value proportional to loss × density.

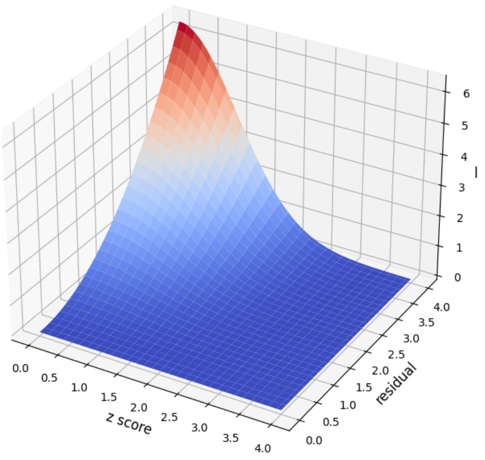
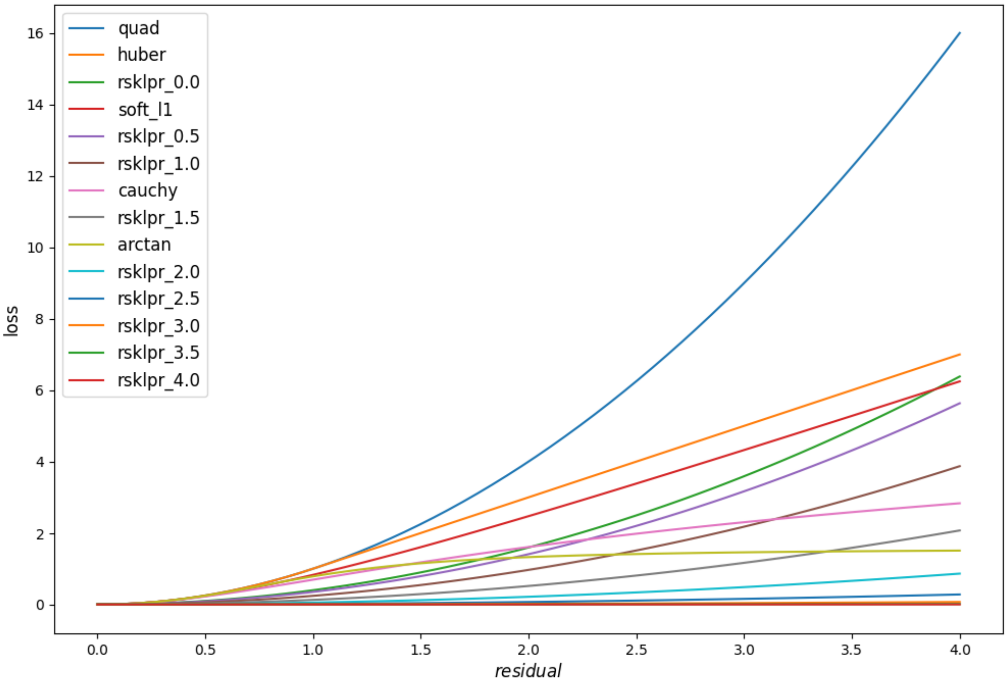


Figure 2: This figure compares the proposed loss function (rsklpr) at various standard deviation levels with common robust losses (e.g., Huber, Cauchy) and the standard quadratic loss. The attenuation of loss in areas with low-density data demonstrates the enhanced robustness of the proposed method. It is assumed that K_2 is equivalent to the standard Gaussian density and the K_1 distance kernel scaling is excluded. The numbers appended to ”rsklpr” indicate the number of standard deviations away from the mean. The vertical axis represents a value proportional to loss × density, while the horizontal axis represents the residual value.



Regardless of the choice of kernel, the hyperparameters of this model are similar in essence to the standard local polynomial regression and comprise the span of included points, the kernels and their associated bandwidths. Note that this estimator can be replaced with other robust density estimators and better results are anticipated by doing so however exploring this option is left for future work.

4. Properties

This section discusses the properties of the proposed estimator, beginning with its interpretation on a finite sample and then moving to its asymptotic behaviour. Note the notation in this section is simplified by excluding explicit mentions of $\mathcal{D}_N$ and $\mathcal{H}$, however the analysis is conditional on the nearest neighbors in the sample, $\mathcal{D}_N$.

4.1. Finite-Sample Interpretation as a Re-weighted LPR

At the sample level, the proposed estimator can be understood as a direct re-weighting of the terms in the standard LPR loss function. The weights are determined by the local conditional density of the response.

Proposition 1 (Equivalence to a Re-weighted LPR Objective)**.** *Minimizing the proposed empirical loss from Corollary 1,*

$$\tilde{\mathcal{L}}(x) = \sum_{i=1}^{N}\Big(Y_i - \sum_{j=0}^{p}\beta_j(x)(X_i - x)^j\Big)^2 K_{h_1}(x - X_i)\,\hat{f}_{Y|X}(Y_i \mid X_i),$$

is equivalent to minimizing a weighted average of the standard LPR loss terms, where each term's contribution is scaled by its estimated conditional density $\hat{f}_{Y|X}(Y_i \mid X_i)$.

Proof. Let $w_i = \hat{f}_{Y|X}(Y_i \mid X_i)$. Assuming that not all weights are zero (i.e., $\sum_{i=1}^{N} w_i > 0$, which holds if the conditional density estimate is non-zero for at least one neighbor), we can divide the objective by this sum without changing the resulting $\hat{\beta}(x)$:

$$\hat{\beta}(x) = \arg\min_{\beta(x)} \frac{1}{\sum_{k=1}^{N} w_k} \sum_{i=1}^{N} w_i \left[\left(Y_i - \sum_{j=0}^{p}\beta_j(x)(X_i - x)^j\right)^2 K_{h_1}(x - X_i)\right]. \tag{9}$$

The term in the square brackets is the i-th term of the standard LPR loss function from Equation (2). The expression is therefore a weighted arithmetic mean of these standard LPR terms. This interpretation makes it clear that points with a low estimated conditional density (i.e., response outliers) are down-weighted in a single, non-iterative step. Note that if a kernel with bounded support (e.g., Epanechnikov) is used for density estimation, it is theoretically possible for all weights w_i in a neighborhood to be zero, although this is not an issue with unbounded kernels like the Gaussian. □

4.2. Asymptotic Properties

We now analyze the behavior of the estimator as the sample size $N \to \infty$.

Proposition 2 (Population Objective and the Intercept Term)**.** *Let $f_{X,Y}(u, v)$ denote the joint density of (X, Y) where $X \in \mathbb{R}^d, Y \in \mathbb{R}$. For a chosen regression point $x \in \mathbb{R}^d$, define the population objective function as*

$$\mathcal{J}(x;\beta) = \iint_{\mathbb{R}^d\times\mathbb{R}} \big(v - g_x(u;\beta)\big)^2 K_{h_1}(x - u)\, w(u, v)\, f_{X,Y}(u, v)\,\mathrm{d}v\,\mathrm{d}u,$$

where $g_x(u;\beta) = \sum_{j=0}^{p}\beta_j(x)(u-x)^j$, K_{h_1} is a kernel function (assumed to be radial and symmetric, hence $K_{h_1}(x - u) = K_{h_1}(\|x - u\|)$), and $w(u, v)$ is a population-level non-negative weight function. Let $\beta^\star(x) = \mathrm{argmin}_\beta\, \mathcal{J}(x;\beta)$. The first component, $\beta_0^\star(x)$, represents the local intercept of the polynomial fit at x.

For local constant regression ($p = 0$), or for local polynomial regression ($p \geq 1$) under Assumption A1, $\beta_0^\star(x)$ is given by:

$$\beta_0^\star(x) = \frac{\displaystyle\iint v\, K_{h_1}(x - u)\, w(u, v)\, f_{X,Y}(u, v)\,\mathrm{d}v\,\mathrm{d}u}{\displaystyle\iint K_{h_1}(x - u)\, w(u, v)\, f_{X,Y}(u, v)\,\mathrm{d}v\,\mathrm{d}u}. \tag{10}$$

Proof. Define the kernel-tilted measure aggregate weight at (u, v) as

$$\omega_x(u, v) = K_{h_1}(x - u)\, w(u, v)\, f_{X,Y}(u, v).$$

Let $\mathcal{N}(x)$ denote the denominator of (10), and write the monomials centered at x as $q_j(u; x) = (u - x)^j$. Using the shorthand $\langle H(u, v)\rangle := \iint H(u, v)\, \omega_x(u, v)\, \mathrm{d}v\, \mathrm{d}u$, the objective function is:

$$\mathcal{J}(x;\beta) = \left\langle v^2 \right\rangle - 2 \sum_{j=0}^{p} \beta_j(x) \left\langle v\, q_j(u; x) \right\rangle + \sum_{j,k=0}^{p} \beta_j(x)\beta_k(x) \left\langle q_j(u; x)\, q_k(u; x) \right\rangle.$$

Differentiating with respect to each $\beta_\ell(x)$ and setting the gradient to zero gives the system of $p+1$ normal equations:

$$\sum_{k=0}^{p} \beta_k^\star(x) \left\langle q_k(u; x)\, q_\ell(u; x) \right\rangle = \left\langle v\, q_\ell(u; x) \right\rangle, \qquad \ell = 0, \ldots, p. \tag{11}$$

The first equation (for $\ell = 0$), noting $q_0(u; x) \equiv 1$, is:

$$\beta_0^\star(x)\mathcal{N}(x) + \sum_{k=1}^{p} \beta_k^\star(x) \left\langle q_k(u; x) \right\rangle = \left\langle v \right\rangle. \tag{12}$$

Assumption A1 (Symmetry in weighted moments). For $p \geq 1$, we assume that the weighted moments of odd order are zero, i.e.,

$$\left\langle q_j(u; x) \right\rangle = \iint (u - x)^j\, K_{h_1}(x - u)\, w(u, v)\, f_{X,Y}(u, v)\, \mathrm{d}v\, \mathrm{d}u = 0, \qquad \text{for odd } j \in \{1, \ldots, p\}. \tag{13}$$

This condition is standard in LPR analysis [see, e.g., 6, Sec. 3.2] and holds if the kernel K_{h_1} is symmetric and the effective weight function $W_0(u) = \int_v w(u, v) f_{Y|X}(v|u) f_X(u) dv$ is locally even in a neighborhood of x. For local linear regression ($p = 1$), this simplifies to requiring $\langle q_1(u; x)\rangle = 0$.

Under Assumption A1, the terms for odd k in Equation (12) vanish. For local linear regression ($p = 1$), this is sufficient to isolate $\beta_0^\star(x)$. For $p \geq 2$, standard LPR results show that this leads to a block-diagonal system, from which the formula for $\beta_0^\star(x)$ holds. For local constant regression ($p = 0$), the sum in (12) is empty, so the result holds without needing Assumption A1. □

Corollary 2 (Population Target with Conditional Density Weights)**.** *Consider the specific case where the weight function is the true conditional density,* $w(u, v) = f_{Y|X}(v \mid u)$*, and assume* $f_{Y|X}(v \mid u) > 0$*. Under the same conditions as Proposition 2, the population intercept* $\beta_0^\star(x)$ *from* (10) *takes the explicit form:*

$$\beta_0^\star(x) = \frac{\iint v\, K_{h_1}(x - u)\, [f_{Y|X}(v \mid u)]^2\, f_X(u)\, \mathrm{d}v\, \mathrm{d}u}{\iint K_{h_1}(x - u)\, [f_{Y|X}(v \mid u)]^2\, f_X(u)\, \mathrm{d}v\, \mathrm{d}u}.$$

This expression can be rewritten as a locally weighted average of $\mu'(u)$*:*

$$\beta_0^\star(x) = \frac{\int K_{h_1}(x - u)\, \mu'(u)\, C(u)\, f_X(u)\, \mathrm{d}u}{\int K_{h_1}(x - u)\, C(u)\, f_X(u)\, \mathrm{d}u},$$

where $\mu'(u) = \frac{\int v[f_{Y|X}(v|u)]^2 dv}{\int [f_{Y|X}(v|u)]^2 dv}$ *and* $C(u) = \int [f_{Y|X}(v \mid u)]^2 dv$*. This shows that as the bandwidth* $h_1 \to 0$ *(under standard rate conditions, e.g.,* $Nh_1^d \to \infty$*, and with* h_2 *either fixed or not shrinking faster than* h_1*),* $\beta_0^\star(x)$ *converges to* $\mu'(x)$*. This target is generally different from the true conditional mean* $m(x) = \mathbb{E}[Y|X = x]$*. This result provides the formal basis for the asymptotic bias discussion.*

4.3. Asymptotic Target and Conditions for Unbiasedness

Corollary 2 establishes that the proposed estimator asymptotically targets $\mu'(x)$. A crucial question is under what conditions this target coincides with the true regression function, $m(x) = \mathbb{E}[Y|X = x]$. The two targets are equivalent, $\mu'(x) = m(x)$, if and only if the conditional distribution $f(Y|X)$ is symmetric about its mean $m(x)$.

The most important instance of such a symmetric distribution is the normal distribution. However, the property holds for any symmetric conditional density (e.g., Laplace, Student's t). If we assume that for each fixed x, the conditional density $f(Y|X = x)$ is symmetric around $m(x)$, then $[f(Y|X)]^2$ is also symmetric around $m(x)$. The expectation with respect to this squared density remains $m(x)$, and therefore $\mu'(x) = m(x)$. Minimizing the expected loss of the proposed method becomes equivalent to minimizing the expected loss of standard LPR. This demonstrates that under the ideal condition of conditional symmetry, the proposed estimator is asymptotically unbiased.

Conversely, when the conditional distribution $f(Y \mid X)$ is asymmetric, the mean under the squared density $\mu'(X)$ will differ from the true mean $m(X)$, introducing an asymptotic bias of $\mathrm{Bias}(x) = \mu'(x) - m(x)$. An example quantifying this bias for the asymmetric exponential distribution is provided in Appendix B.

4.4. Comparison with Standard and Iterative Robust LPR

While the proposed robust method builds on the LPR framework, its weighting mechanism introduces key differences.

4.4.1. The Core Difference vs. Standard LPR: The Weighting Function

The fundamental difference lies in what determines the "importance" of a neighboring data point (u, v) when estimating the regression function at a point x. For the standard LPR The population objective aims to minimize:

$$\mathcal{J}_{\mathrm{std}}(x;\beta) = \iint \left(v - g_x(u;\beta)\right)^2 K_{h_1}(x-u) f_{X,Y}(u,v)\,\mathrm{d}v\,\mathrm{d}u$$

The weight is determined by the kernel $K_{h_1}(x-u)$ and the data-generating process, but it is linear in the conditional density term $f_{Y|X}(v|u)$.

For the proposed method (with $w(u, v) = f_{Y|X}(v|u)$), the population objective is:

$$\mathcal{J}_{\mathrm{rsk}}(x;\beta) = \iint \left(v - g_x(u;\beta)\right)^2 K_{h_1}(x-u)[f_{Y|X}(v|u)]^2 f_X(u)\,\mathrm{d}v\,\mathrm{d}u$$

The proposed method's key innovation is the squaring of the conditional density term, $[f_{Y|X}(v|u)]^2$. This change amplifies the weighting effect, more strongly down-weighting observations (u, v) where the response v is unlikely given the predictor u.

4.4.2. The True Counterpart: Iterative Robust Methods

The direct practical counterpart to the proposed method is iterative robust LPR, such as the procedure used in LOWESS. These methods use an *iterative approach* by repeatedly fitting the data and adjusting weights. After each fit, residuals are calculated, and new ”robustness weights” are assigned to each point, typically by down-weighting points with large residuals. In contrast, the proposed method is a single-step procedure where the weights are derived from an explicit estimate of the data-generating distribution itself.

It is understood that many robust estimators can introduce some bias as a price for their resilience to outliers. While iterative robust LPR is also subject to such biases, this aspect often receives insufficient attention in the literature, largely due to the analytical challenges involved. In contrast, the proposed method, by virtue of its non-iterative nature and direct link to the data distribution, makes this trade-off explicit. The bias towards $\mu'(x)$ is clearly defined and can be analyzed, offering a degree of theoretical transparency that is not readily available for its iterative counterparts.

4.5. Trade-off Between Robustness and Bias via the K_2 Kernel and Bandwidth Selection

The proposed estimator utilizes the K_2 kernel to adjust data point weights based on both predictors and responses, controlling the trade-off between robustness and bias. The bandwidth $\mathcal{H}_2$ of the K_2 kernel plays a crucial role in this mechanism.

In the loss function, each data point is weighted by $w_i = K_{h_1}(x - X_i)\hat{f}(Y_i \mid X_i; \mathcal{H}_2)$. The K_2 component assigns lower weights to less probable responses, effectively down-weighting outliers.

The bandwidth $\mathcal{H}_2$ controls the sensitivity of K_2 to variations in the response. For very small $\mathcal{H}_2$ values the density estimator $\hat{f}(Y_i \mid X_i; \mathcal{H}_2)$ becomes sharply peaked at each Y_i, and the weights become nearly uniform after normalization, diminishing robustness. Conversely, for very large $\mathcal{H}_2$ the density estimator becomes nearly constant across different Y_i, and the estimator approaches standard LPR. An intermediate bandwidth $\mathcal{H}_2$ achieves a balance. The optimal $\mathcal{H}_2$ can be selected using methods like cross-validation. This adaptive capability opens the door for more sophisticated, context-dependent bandwidth selection strategies but is left for future work.

4.6. Relationship to Kernel Methods and RKHS

The use of positive definite kernels in defining the weights $\mathcal{K}_\mathcal{D}$ allows the proposed estimator to be interpreted within the Reproducing Kernel Hilbert Spaces (RKHS) framework. If $\mathcal{K}_\mathcal{D}$ is chosen to be a positive definite kernel (e.g., by ensuring both K_1 and K_2 are positive definite), it induces a feature map $\phi : \mathcal{D} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space, such that:

$$\mathcal{K}_\mathcal{D}\left((x,y),(x',y')\right) = \langle \phi(x,y), \phi(x',y') \rangle_\mathcal{H}. \tag{14}$$

The weights $\mathcal{K}_\mathcal{D}((x,y),(X_i,Y_i))$ can be interpreted as inner products in the feature space $\mathcal{H}$. Consequently, the loss function can be viewed as a weighted least-squares problem where the weights are determined by the similarity between the feature representations of the data points and the point of interest.

Furthermore, consider the role of the Kernel Density Estimator (KDE) in the proposed method. The KDE at a point (x,y) using a positive definite kernel K_2 is given by:

$$\hat{f}(x,y) = \frac{1}{N}\sum_{i=1}^{N} K_2\left((x,y),(X_i,Y_i);\mathcal{H}_2\right). \tag{15}$$

Letting K_2 be positive definite, there exists a feature mapping $\psi : \mathcal{D} \to \mathcal{G}$ such that the KDE at (x, y) can be expressed as:

$$\hat{f}(x, y) = \left\langle \psi(x, y), \frac{1}{N} \sum_{i=1}^{N} \psi(X_i, Y_i) \right\rangle_{\mathcal{G}} . \tag{16}$$

This expression shows that the KDE measures how closely the feature representation $\psi(x, y)$ aligns with the average feature representation of the data in the space induced by K_2. In the proposed method, this alignment influences the weights in the regression, as the density estimates derived from K_2 directly affect the overall weights. By leveraging positive definite kernels, the method inherently operates within the RKHS framework, where weights represent similarities in feature space. This perspective highlights the connection between the kernel-based weighting and the feature mappings, offering insights into the estimator's flexibility and robustness.

5. Experiments and Implementation Notes

This section presents an evaluation of the proposed method (RSKLPR), implemented in Python and published as an open source package https://github.com/yaniv-shulman/rsklpr. The experiments focus on comparing the performance of RSKLPR against existing local regression techniques under synthetic settings with different noise characteristics.

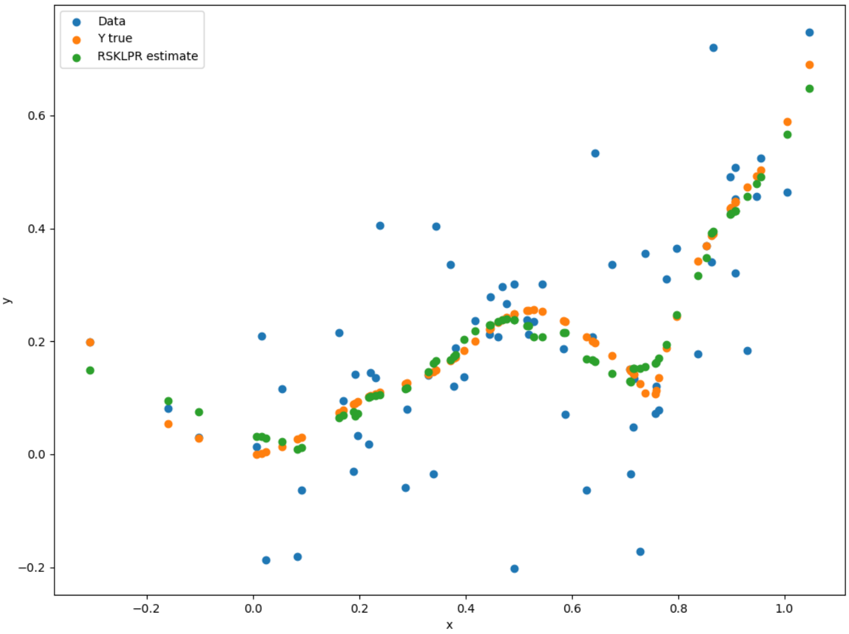
Implementation Details

The implementation normalizes distances in each neighborhood to the range $[0, 1]$, consistent with the approach in [3], effectively making the bandwidth for K_1 adaptive. For the kernel $K_1(x, x')$, a Laplacian kernel $e^{-\|x-x'\|}$ was selected. Note that this is an un-normalized kernel; since weights within a neighborhood are scaled, the normalization constant does not affect the final estimate. For density estimation in K_2, a factorized multidimensional Kernel Density Estimator (KDE) with scaled Gaussian kernels was used. This factorization is a simplification that ignores potential covariance between predictors and response but is computationally efficient. This approximation may under-down-weight observations where the predictors and response are strongly correlated; a full joint bandwidth matrix could be used to address this without changing the underlying theory. Bandwidth selection for density estimation was explored using several standard methods. Scaling constants within neighborhoods, such as those in $\hat{f}(y \mid x)$ and $\hat{f}(x, y)$, were excluded for computational efficiency, as they do not impact the local regression results. The experiments were done only with the local linear estimator i.e. $p = 1$ as it is well known to be superior.

Experimental Design

Synthetic datasets were generated with both additive Gaussian noise and asymmetric data distributions to simulate various regression scenarios. The following characteristics were varied: noise types, including homoscedastic and heteroscedastic Gaussian noise as well as asymmetric noise distributions (Exponential, Log-normal, Gamma, and Weibull); data density, encompassing both sparse and dense data regimes; and regression complexity, modeling non-linear curves and surfaces. Performance was evaluated using Root Mean Square Error (RMSE) and sensitivity to neighborhood size.

Figure 3: Performance of RSKLPR on 1D synthetic data with heteroscedastic Gaussian noise. The proposed method effectively aligns with the true regression function while mitigating the influence of outliers and noise.

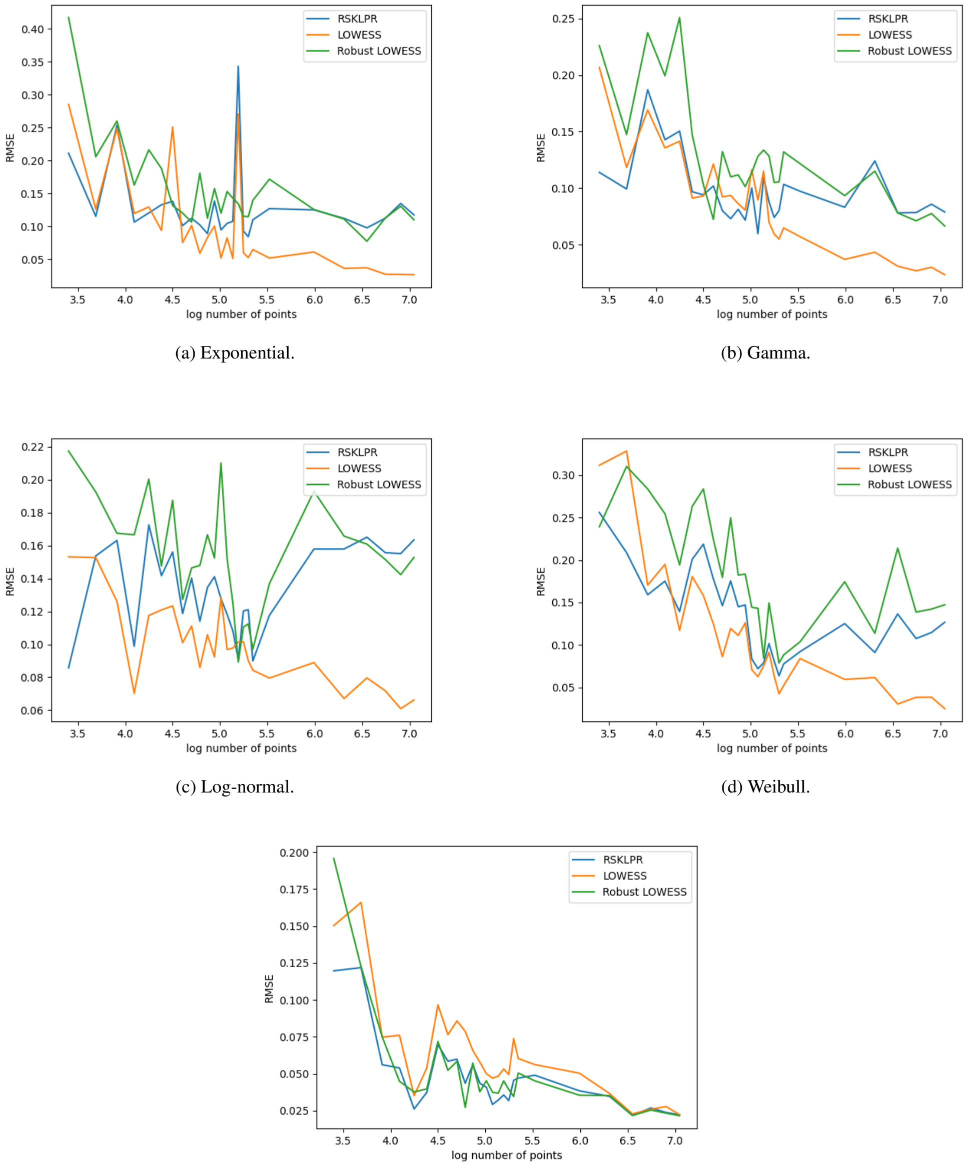


Results and Observations

Under Gaussian noise settings, the proposed method performed competitively. Unlike iterative robust variants, RSKLPR achieved these results with a single iteration. A regression example with heteroscedastic Gaussian noise is shown in Figure 3. The proposed method aligns with the true regression function while effectively mitigating the influence of noise and outliers.

Under asymmetric data distributions, RSKLPR exhibited robust performance in low density settings, often matching or outperforming standard LPR and the iterative robust variant. In high-density settings, the proposed method diverged from the true mean, confirming the theoretical results on asymptotic bias. However, it consistently outperformed the iterative robust LPR. Figure 4 presents RMSE trends for asymmetric noise distributions for the three methods.

The method was also significantly less sensitive to the neighborhood size making it an an attractive option for applications where robust regression is critical. Complete experimental results, including multivariate settings and bootstrap-based confidence intervals, are available at https://nbviewer.org/github/yaniv-shulman/rsklpr/tree/main/src/experiments as interactive Jupyter notebooks [1].



(a) Exponential.

(b) Gamma.

(c) Log-normal.

(d) Weibull.

(e) Gaussian.

Figure 4: These subplots compare RMSE as a function of data density for the proposed method (RSKLPR), standard LOWESS, and Robust LOWESS (5 iterations) across various noise distributions: (a) Exponential, (b) Gamma, (c) Log-normal, (d) Weibull, and (e) Gaussian. The results demonstrate the effectiveness of RSKLPR in low-density data and align well with theoretical expectations for denser data.

6. Future Work and Research Directions

This work introduces a new robust variant of Local Polynomial Regression (LPR), opening several avenues for further exploration and refinement. Since the proposed method generalizes the traditional LPR, there are opportunities to replace certain standard components in equation (4) with more robust alternatives. These could include approaches such as robust methods for bandwidth selection or substituting the conventional quadratic residual function with alternatives better suited for handling outliers.

An important research direction is to explore adaptive bandwidth selection strategies that respond dynamically to local data density. In regions where data are sparse, the bandwidth in K_2 could be fine-tuned to maintain robust down-weighting of potential outliers. Conversely, in denser regions, broader bandwidths may be adopted, causing the estimator to behave more like standard LPR and reduce any bias introduced by the robust weighting. Incorporating such adaptive bandwidths could further enhance the method's overall performance and flexibility.

Additionally, further development of this framework may involve exploring different kernel functions and assessing how robust density estimators influence overall performance. Extending the method within the RKHS framework presents another valuable direction. This could allow for the introduction of a regularization term in the loss function, enhancing control over estimator smoothness and mitigating the risk of overfitting. Through these future directions, the robustness and adaptability of the proposed method could be substantially advanced.

References

[1] Project jupyter is a non-profit, open-source project, born out of the ipython project in 2014 as it evolved to support interactive data science and scientific computing across all programming languages. `https://jupyter.org/`.

[2] M. Avery. Literature review for local polynomial regression. 2010.

[3] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[4] W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[5] J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004, 1992.

[6] J. Fan. Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21, 03 1993.

[7] J. Fan and I. Gijbels. *Local polynomial modelling and its applications*. Chapman & Hall/CRC, Boca Raton, Fla., 1996.

[8] E. García-Portugués. *Notes for Nonparametric Statistics*. 2023. Version 6.9.0. ISBN 978-84-09-29537-1.

[9] T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11:171–185, 1984.

[10] R. A. Maronna, D. Martin, V. J. Yohai, and Hardcover. Robust statistics: Theory and methods. 2006.

[11] E. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9:141–142, 1964.

[12] M. Salibian-Barrera. Robust nonparametric regression: Review and practical considerations. *Econometrics and Statistics*, 2023.

[13] V. G. Spokoiny. Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *The Annals of Statistics*, 26(4):1356 – 1378, 1998.

[14] C. J. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–620, 1977.

[15] P. Čížek and S. Sadıkoğlu. Robust nonparametric regression: A review. *WIREs Comput. Stat.*, 12(3), apr 2020.

[16] G. S. Watson. Smooth regression analysis. 1964.

Appendix A. Joint Density Kernel

An alternative kernel for K_2 can be defined that is proportional to the joint distribution of the random pair. This could be useful, for example, to also down-weight high-leverage points in the predictor space.

$$K_2\Big((x,y),(x',y');\mathcal{H}_2\Big) = \hat{f}(x,y;\mathcal{H}_2)\hat{f}(x',y';\mathcal{H}_2) \tag{A.1}$$

Where the joint density can be estimated using the Parzen-Rosenblatt window estimator. This choice also satisfies the conditions of Lemma 1, with $c(x,y) = \hat{f}(x,y;\mathcal{H}_2)$ and $w(X_i,Y_i) = \hat{f}(X_i,Y_i;\mathcal{H}_2)$. The simplified empirical objective function for a point (X_i,Y_i) in the neighborhood becomes:

$$\tilde{\mathcal{L}}(x) = \sum_{i=1}^{N}\Big(Y_i - \sum_{j=0}^{p}\beta_j(x)(X_i-x)^j\Big)^2 K_{h_1}(x-X_i)\,\hat{f}(X_i,Y_i;\mathcal{H}_2).$$

This formulation weights each point (X_i,Y_i) by its estimated joint density, in addition to the standard distance-based weight $K_{h_1}(x-X_i)$.

The mechanism by which this kernel provides robustness becomes clearer when we consider its population-level objective. The objective function involves an integral term weighted by $[f(X,Y)]^2$. We can decompose this squared joint density:

$$[f(X,Y)]^2 = [f(Y|X)\cdot f(X)]^2 = [f(Y|X)]^2\cdot[f(X)]^2.$$

This decomposition reveals a dual weighting mechanism. The $[f(Y|X)]^2$ term provides robustness to outliers in the response variable, operating identically to the conditional density kernel discussed in the main paper. Simultaneously, the $[f(X)]^2$ term directly addresses high-leverage points. Observations X that lie in low-density regions of the predictor space will have a small $f(X)$ value, and this effect is amplified by the squaring.

Therefore, the joint density kernel explicitly down-weights points that are unusual in either the response space (outliers) or the predictor space (high-leverage points). This provides a clear theoretical underpinning for its use in settings where both types of robust treatment are desired. A full investigation of its properties is left for future work.

Appendix B. Asymptotic Bias Example with an Exponential Conditional Distribution

This appendix illustrates how asymmetry in the conditional distribution $f(Y \mid X)$ can introduce asymptotic bias in the proposed estimator. The focus is on a standard exponential distribution.

Suppose that for each fixed X, the conditional distribution $f(Y \mid X)$ follows a standard exponential law with rate parameter $\lambda(X)$:

$$f(Y \mid X) \;=\; \lambda(X)\,\exp\!\Big(-\lambda(X)\,Y\Big), \qquad Y \geq 0,$$

so that the true regression function is

$$m(X) \;=\; \mathbb{E}[Y \mid X] \;=\; \frac{1}{\lambda(X)}.$$

When this density is squared, we obtain

$$[f(Y \mid X)]^2 \;=\; [\lambda(X)]^2 \exp\!\Big(-2\,\lambda(X)\,Y\Big), \quad Y \geq 0,$$

which is proportional to an exponential density with rate $2\,\lambda(X)$. The mean of Y under this squared density is

$$\mu'(X) \;=\; \frac{1}{2\,\lambda(X)}.$$

As established in the main text, the proposed estimator asymptotically converges to $\mu'(X)$ rather than $m(X)$. Consequently, at each point x, the asymptotic bias is

$$\mathrm{Bias}(x) \;=\; \mu'(x) \;-\; m(x) \;=\; \frac{1}{2\,\lambda(x)} \;-\; \frac{1}{\lambda(x)} \;=\; -\frac{1}{2\,\lambda(x)}.$$

This example illustrates how the asymmetry of an exponential distribution can steer the estimator toward $1/(2\,\lambda(X))$ rather than the true mean $1/\lambda(X)$. Although such a shift introduces asymptotic bias, the robust weighting can still be advantageous in practical situations where outliers or heavy-tailed noise are significant concerns.