Kernel Density Estimation - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

Kernel Density Estimation (KDE) is a nonparametric method to estimate an unknown probability density f on $\mathbb{R}^d$ from i.i.d. samples. From first principles, KDE is the convolution of the empirical measure with a *smoothing kernel*, yielding a continuous density estimator that trades bias and variance through a *bandwidth* parameter. We derive the estimator, prove normalization, compute its bias/variance, obtain optimal (asymptotic) bandwidths via MISE/AMISE analysis, and give multivariate and algorithmic formulations.

1.2 Data and Notation

Let the data be i.i.d.

$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n, \qquad \mathbf{x}_i \in \mathbb{R}^d,$$

with true density $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. Write the empirical measure

$$\widehat{P}_n \;=\; \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}.$$

A *kernel* is a nonnegative function $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ satisfying

$$\int_{\mathbb{R}^d} K(\mathbf{u})\,d\mathbf{u} = 1, \qquad \int_{\mathbb{R}^d} \mathbf{u}\,K(\mathbf{u})\,d\mathbf{u} = \mathbf{0}, \qquad \int_{\mathbb{R}^d} \|\mathbf{u}\|^2 K(\mathbf{u})\,d\mathbf{u} = \mu_2(K) < \infty. \tag{1}$$

The *bandwidth* $h > 0$ controls smoothing. Define the scaled kernel

$$K_h(\mathbf{u}) \;=\; h^{-d}K(\mathbf{u}/h).$$

1.3 Model Formulation (Convolution View)

The KDE at $\mathbf{x} \in \mathbb{R}^d$ is the empirical convolution with K_h:

$$\widehat{f}_h(\mathbf{x}) \;=\; (K_h * \widehat{P}_n)(\mathbf{x}) \;=\; \int K_h(\mathbf{x}-\mathbf{z})\, d\widehat{P}_n(\mathbf{z}) \;=\; \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{x}_i) \;=\; \frac{1}{nh^d}\sum_{i=1}^{n} K\Big(\frac{\mathbf{x}-\mathbf{x}_i}{h}\Big). \tag{2}$$

Normalization (proof). Using Eqn. (2) and $\int K_h = 1$,

$$\int_{\mathbb{R}^d} \widehat{f}_h(\mathbf{x})\, d\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n} \int K_h(\mathbf{x}-\mathbf{x}_i)\, d\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n} 1 = 1.$$

Thus $\widehat{f}_h$ is a valid density. ■

1.4 Pointwise Bias and Variance (Univariate First, $d = 1$)

Assume f is twice continuously differentiable and K satisfies Eqn. (1).

Bias. By the change of variables $u = (x-\xi)/h$,

$$\mathbb{E}\,\widehat{f}_h(x) = \int K_h(x-\xi) f(\xi)\, d\xi = \int K(u)\, f(x-hu)\, du.$$

A Taylor expansion of $f(x-hu)$ at x gives

$$f(x-hu) = f(x) - huf'(x) + \tfrac{h^2u^2}{2} f''(x) + o(h^2).$$

Using $\int K = 1$ and $\int uK(u)\, du = 0$,

$$\mathrm{Bias}\big[\widehat{f}_h(x)\big] = \mathbb{E}\,\widehat{f}_h(x) - f(x) = \frac{h^2\,\mu_2(K)}{2}\, f''(x) + o(h^2), \qquad \mu_2(K) = \int u^2 K(u)\, du. \tag{3}$$

Variance. Since the summands are i.i.d.,

$$\mathrm{Var}\big[\widehat{f}_h(x)\big] = \frac{1}{n}\,\mathrm{Var}\Big(K_h(x-X)\Big) = \frac{1}{n}\Big(\mathbb{E}\,K_h^2(x-X) - (\mathbb{E}\,K_h(x-X))^2\Big).$$

Compute $\mathbb{E}\,K_h^2(x-X) = \int K_h^2(x-\xi) f(\xi)\, d\xi = \frac{1}{h}\int K^2(u)\, f(x-hu)\, du = \frac{R(K)}{h} f(x) + o(h^{-1})$, where

$$R(K) = \int K^2(u)\, du.$$

Hence

$$\mathrm{Var}\big[\widehat{f}_h(x)\big] = \frac{f(x)\, R(K)}{n\,h} + o\Big(\frac{1}{nh}\Big). \tag{4}$$

Asymptotics and consistency. If $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then the bias $\to 0$ by Eqn. (3) and the variance $\to 0$ by Eqn. (4), giving pointwise consistency. Moreover,

$$\sqrt{nh}\,\big(\widehat{f}_h(x) - \mathbb{E}\,\widehat{f}_h(x)\big) \;\xrightarrow{d}\; \mathcal{N}\big(0,\; f(x)R(K)\big).$$

1.5 MISE and AMISE; Optimal Bandwidth (Univariate)

Define the Mean Integrated Squared Error

$$\mathrm{MISE}(h) = \mathbb{E}\int \big(\widehat{f}_h(x) - f(x)\big)^2\,dx = \int \mathrm{Bias}^2(x)\,dx + \int \mathrm{Var}(x)\,dx.$$

Using Eqns. (3)–(4) and neglecting higher-order terms,

$$\mathrm{AMISE}(h) \;=\; \frac{R(K)}{nh} + \frac{h^4\,\mu_2(K)^2}{4}\,R\big(f''\big), \qquad R(g) = \int g(x)^2\,dx.$$

Minimizing AMISE(h) over $h > 0$ yields

$$h^{\star}_{\mathrm{AMISE}} = \left(\frac{R(K)}{\mu_2(K)^2\,R(f'')}\right)^{1/5} n^{-1/5}. \tag{5}$$

In practice $R(f'')$ is unknown; *plug-in* or *rules of thumb* approximate it. For Gaussian K and approximately normal data with standard deviation $\hat{\sigma}$, Silverman's rule takes

$$h_{\mathrm{Silv}} = 0.9\,\min\big(\hat{\sigma},\,\tfrac{\mathrm{IQR}}{1.34}\big)\,n^{-1/5}. \tag{6}$$

1.6 Bandwidth Selection by Cross-Validation (Univariate)

Two classical data-driven criteria are:

Least-Squares Cross-Validation (LSCV). Minimize

$$\mathrm{LSCV}(h) = \int \widehat{f}_h(x)^2\,dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i,h}(x_i),$$

where $\widehat{f}_{-i,h}$ is the leave-one-out estimator. Using convolution identities,

$$\int \widehat{f}_h^2 = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n}(K*K)\Big(\frac{x_i - x_j}{h}\Big), \quad \widehat{f}_{-i,h}(x_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K\Big(\frac{x_i - x_j}{h}\Big).$$

Biased Cross-Validated Log-Likelihood. Maximize

$$\mathrm{CVLL}(h) = \frac{1}{n}\sum_{i=1}^{n} \log \widehat{f}_{-i,h}(x_i),$$

which targets predictive fit; numerically it is robust for unimodal densities.

1.7 Multivariate KDE ($d \geq 1$)

Let a positive-definite bandwidth matrix $\mathbf{H} \in \mathbb{R}^{d\times d}$ and define

$$K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{u}), \qquad \widehat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i).$$

Common choices: (i) *isotropic* $\mathbf{H} = h^2\mathbf{I}_d$; (ii) *diagonal* $\mathbf{H} = \mathrm{diag}(h_1^2, \ldots, h_d^2)$; (iii) full $\mathbf{H}$.

Bias/variance (isotropic). Under smoothness (with Laplacian Δf),

$$\mathrm{Bias}\big[\widehat{f}_h(\mathbf{x})\big] \approx \frac{h^2\mu_2(K)}{2}\,\Delta f(\mathbf{x}), \qquad \mathrm{Var}\big[\widehat{f}_h(\mathbf{x})\big] \approx \frac{f(\mathbf{x})R(K)}{n\,h^d}. \tag{7}$$

Therefore

$$\mathrm{AMISE}(h) \;\approx\; \frac{R(K)}{n\,h^d} + \frac{h^4\mu_2(K)^2}{4}\,R(\Delta f),$$

whose minimizer scales as

$$h^\star_{\mathrm{AMISE}} \;\propto\; n^{-1/(d+4)}. \tag{8}$$

Consistency requires $h \to 0$ and $nh^d \to \infty$.

1.8 Choice of Kernel and Efficiency

For fixed bandwidth and under AMISE, kernels differ only through $R(K)$ and $\mu_2(K)$. Among second-order kernels, the Epanechnikov kernel minimizes AMISE:

$$K_{\mathrm{Epa}}(u) = \tfrac{3}{4}(1-u^2)\mathbf{1}\{|u| \le 1\}.$$

In practice, the bandwidth dominates performance; Gaussian kernels are popular for their smoothness and FFT-friendly convolution on grids.

1.9 Boundary Bias and Remedies (Univariate on $[a, b]$)

When support is bounded, naive KDE underestimates near boundaries because mass leaks outside. Remedies include: (i) *reflection*: augment data with reflected points; (ii) *boundary kernels* that reweight/truncate K near edges; (iii) *transforms* (e.g., log or logit), estimate in transformed space, and back-transform with Jacobian.

1.10 Algorithm (Evaluation and Bandwidth Search)

1. **Input:** $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, kernel K, bandwidth (scalar h or matrix $\mathbf{H}$), query points $\{\mathbf{x}^{(q)}\}_{q=1}^Q$.
2. **Evaluate KDE:** For each q, compute

$$\widehat{f}(\mathbf{x}^{(q)}) = \frac{1}{n}\sum_{i=1}^n |\mathbf{H}|^{-1/2} K\big(\mathbf{H}^{-1/2}(\mathbf{x}^{(q)} - \mathbf{x}_i)\big).$$

3. **Select bandwidth (optional):** Optimize h (or $\mathbf{H}$) via AMISE plug-in Eqn. (5), rule Eqn. (6), LSCV, or CVLL over a search grid.

Complexity. Direct evaluation costs $O(nQd)$. For large n, Q, use tree-based pruning for compact-support kernels, FFT on grids for Gaussian kernels, or fast Gauss transforms.

1.11 Connections and Properties

- **Smoother of empirical measure.** $\widehat{f}_h = K_h * \widehat{P}_n$ is a linear smoother; as $h \downarrow 0$, $\widehat{f}_h$ approaches the empirical spikes; as $h \uparrow \infty$, it flattens.
- **Moment preservation.** With symmetric K of unit mass, $\int \mathbf{x}\,\widehat{f}_h(\mathbf{x})\,d\mathbf{x} = \frac{1}{n}\sum_i \mathbf{x}_i$ (sample mean).
- **CDF estimator.** The integrated KDE yields a smoothed empirical CDF: $\widehat{F}_h(x) = \frac{1}{n}\sum_i \int_{-\infty}^{(x-x_i)/h} K(u)\,du$.

1.12 Summary of Variables and Their Dimensions

- $\mathbf{x}_i \in \mathbb{R}^d$: ith observation; n: sample size; d: dimension.
- $K : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$: kernel with unit mass, zero mean, finite second moment $\mu_2(K)$; $R(K) = \int K^2$.
- $h > 0$ or $\mathbf{H} \in \mathbb{R}^{d\times d}$: bandwidth (scalar or matrix).
- $\widehat{f}_h(\mathbf{x}) = \frac{1}{nh^d} \sum_i K\big((\mathbf{x} - \mathbf{x}_i)/h\big)$: KDE (scalar density value).
- Bias/variance (isotropic): Bias $\approx \frac{h^2 \mu_2(K)}{2} \Delta f$, Var $\approx \frac{f\,R(K)}{nh^d}$.
- AMISE: AMISE$(h) \approx \frac{R(K)}{nh^d} + \frac{h^4 \mu_2(K)^2}{4} R(\Delta f)$; optimal $h^\star \propto n^{-1/(d+4)}$.

1.13 Summary

From first principles, KDE arises by convolving the empirical measure with a scaled kernel K_h, yielding a bona fide density estimator with unit integral. A second-order Taylor analysis delivers explicit bias and variance, exposing the bandwidth-driven bias–variance trade-off and leading to AMISE-optimal rates $h^\star \propto n^{-1/(d+4)}$. In higher dimensions, isotropic or full bandwidth matrices control anisotropic smoothing, with consistency guaranteed when $h \to 0$ and $nh^d \to \infty$. Practical performance hinges far more on bandwidth choice than on the specific (reasonable) kernel.