Institute of Mathematics, EPFL MATH524

Chapter 3: Kernel density estimation

Lecturer: Rajita Chandak *Spring 2025*

We now consider the density (or distribution) estimation problem. If we make some parametric assumptions on the density function then the problem boils down to estimating the finite collection of parameters that identify the density, typically done with the maximum likelihood estimator (MLE).

If, however, we would like to make fewer restrictive assumptions on the family of distributions that the true density belongs to, the MLE no longer applies and we need a different set of tools to answer the question.

The natural first step in this case is to look at the empirical function. We start by considering the empirical CDF.

1 Empirical density estimation

Assume we have n i.i.d. data points, $X_1, \dots, X_n$ generated from some probability density function f. The CDF is then defined as $F(x) = \int_{-infty}^{x} f(t)\mathrm{d}t$. The empirical CDF (ECDF) is then defined as

$$F_n(x) = \frac{1}{n} \sum \mathbf{1}(X_i \leq x).$$

By the strong law of large numbers, it can be shown that $F_n(x) \xrightarrow{a.s.} F(x)$ for all $x \in \mathbb{R}$ as $n \to \infty$. Thus, F_n is a consistent estimator of the CDF. Now, the question is: How can we estimate the PDF from the empirical CDF estimator? From our understanding of derivatives, the natural solution is to approximate the derivative of the CDF,

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

for sufficiently small $h > 0$. Then, we plug-in F_n for F,

$$\hat{f}_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

$\hat{f}_n$ is known as the *Rosenblatt estimator*. An equivalent formulation of the estimator is

$$\hat{f}_n^R(x) = \frac{1}{2nh} \sum_i \mathbf{1}(x-h < X_i \leq x+h) = \frac{1}{nh} \sum_i K\Big(\frac{X_i - x}{h}\Big),$$

where $K(u) = \frac{1}{2}\mathbf{1}(-1 < u \leq 1)$ (uniform kernel).

2 Kernel density estimation

The Rosenblatt estimator can be generalized to a collection of estimators, known as the *kernel density estimators*,

$$\hat{f}_n(x) = \frac{1}{n}\sum_i K_h(X_i, ; x),$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function that satisfies $\int K(u)\mathrm{d}u = 1$. Note that the structure of this estimator is very similar to the local linear regression estimator.

Importantly, if K is non-negative and $X_1, \ldots, K_n$ are fixed, the mapping of $x \mapsto \hat{f}_n(x)$ is a valid probability density.

2.1 Kernel functions

We have already seen some common kernel functions in the previous lecture. Here we list a few more kernel functions.

- Biweight: $K(u) = \frac{15}{16}(1 - u^2)\mathbf{1}(|u| \leq 1)$
- Silverman: $K(u) = \frac{1}{2}\exp(-|u|/\sqrt{2})\sin(|u|/\sqrt{2} + \pi/4)$

2.2 MSE and Bias-Variance trade-off

Now, we investigate the reliability (or accuracy) of the kernel density estimator by evaluating its MSE:

$$\mathrm{MSE}(x) = \mathbb{E}_f[(\hat{f}_n(x) - f(x))^2].$$

Note that the expectation is with respect to the true density, f. We start with the bias-variance decomposition of the MSE:

$$\mathrm{MSE}(x) = \mathrm{Bias}^2(\hat{f}_n(x)) + \mathbb{V}(\hat{f}_n(x))$$

where

$$\mathrm{Bias}(\hat{f}_n(x)) = \mathbb{E}_f[\hat{f}_n(x)] - f(x)$$

and

$$\mathbb{V}(\hat{f}_n(x)) = \mathbb{E}_f[(\hat{f}_n(x) - \mathbb{E}_f[\hat{f}_n(x)])^2].$$

We will analyze the bias and variance separately.

2.2.1 Bias

Let us first write out the bias in its most simplified form:

$$\mathrm{Bias}(\hat{f}_n(x)) = \mathbb{E}_f[\hat{f}_n(x)] - f(x) = \frac{1}{h}\int K\left(\frac{u - x}{h}\right) f(u)\mathrm{d}u - f(x).$$

To make any progress in understanding how this bias changes with h (the only hyper-parameter), we need to make some assumptions about the class of functions f that and the kernel function K. Consider the following:

Definition 1 (Hölder function class)**.** *Suppose $T \subset \mathbb{R}$ and β and L are two positive numbers. The class of functions $\mathcal{H}(\beta, L)$ is called the Hölder class if the set of $l = \lfloor \beta \rfloor$ times differentiable functions $f : T \to \mathbb{R}$ whose l-th derivative satisfies*

$$|f^{(l)}(x) - f^{(l)}(x')| \leq L|x - x'|^{\beta - l} \quad \forall x, x' \in T$$

Definition 2 (*l-th* order kernels)**.** *Assume $l \geq 1$ is an integer. K is an l-th order kernel if the functions $u \mapsto u^j K(u)$ for $j = 0, \ldots, l$ are integrable and satisfy*

$$\int K(u)\mathrm{d}u = 1 \qquad \int u^j K(u)\mathrm{d}u = 0, \quad j = 1, \ldots, l-1,$$

and

$$\int u^l K(u)\mathrm{d}u > 0.$$

Definition 3 (Symmetric kernels)**.** *A kernel is symmetric if $K(u) = K(-u)$. In this case, all odd moments of the kernel are necessarily 0. Therefore, the order of a symmetric kernel is always an even number.*

Many of the most common kernels (like the examples in Section 2.1 and in Chapter 2) are second-order kernels (symmetric, non-negative kernels are second-order kernels). We will see what higher-order kernel functions look like later.

Now, we define $f \in F(\beta, L)$, to mean:

$$F(\beta, L) = \{f \mid f \geq 0,\ \textstyle\int f(x)\mathrm{d}x = 1,\ \text{and} f \in \mathcal{H}(\beta, L)\}$$

Proposition 1 (Bias rate)
Suppose K is an order $l = \lfloor \beta \rfloor$ kernel and $f \in F(\beta, L)$. Furthermore, assume

$$\int |u|^\beta |K(u)|\mathrm{d}u < \infty.$$

Then, for all $x \in \mathbb{R}$, $h > 0$ and $n \geq 1$,

$$|Bias(\hat{f}_n(x))| \leq Ch^\beta$$

where

$$C = \frac{L}{l!} \int |u|^\beta |K(u)|\mathrm{d}u.$$

2.2.2 Variance

The following proposition provides control on the variance.

Proposition 2 (Variance bound)
Suppose $f(x) < f_{\max} < \infty$ *for all* $x \in \mathbb{R}$. *Let* $K : \mathbb{R} \to \mathbb{R}$ *be a function such that*

$$\int K^2(u)\mathrm{d}u < \infty.$$

Then, for all $x \in \mathbb{R}$, $h > 0$ *and* $n \geq 1$,

$$\mathbb{V}(\hat{f}_n(x)) \leq \frac{C}{nh}$$

where

$$C = f_{\max} \int K^2(u)\mathrm{d}u.$$

Note that this result shows that if the bandwidth is chosen to depend on n, $h = h_n$ such that $nh_n \to \infty$ as $n \to \infty$ and $h_n \to 0$, the variance will converge to 0.

Putting together the bias and variance bounds, we see that the MSE is bounded by

$$\mathrm{MSE} \leq C_1 h^{2\beta} + C_2 \frac{1}{nh}.$$

It is easy to see from this bound that the bias and variance terms depend on the bandwidth in opposite ways. This is precisely what leads to the parabolic shape of the MSE curve (plotted against h), and mathematically describes the bias-variance trade-off. As we have described previously, large h leads to increased bias, which is often termed **over-smoothing**.

To choose the optimal bandwidth, we can minimize the MSE bound with respect to h:

$$\begin{aligned}
\hat{h} &= \operatorname*{argmin}_{h>0} C_1 h^{2\beta} + C_2 \frac{1}{nh} \\
&= \left(\frac{C_2}{2\beta C_1}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}} \\
&= O\left(n^{-\frac{1}{2\beta+1}}\right)
\end{aligned}$$

This optimal bandwidth gives

$$\mathrm{MSE}(x) = O(n^{-\frac{2\beta}{2\beta+1}}).$$

It turns out that this bound can be established uniformly (i.e., over all values of x) as well, through the following theorem:

Theorem 1 (Uniform MSE bound)
Suppose K is an order $l = \lfloor \beta \rfloor$ kernel and $f \in F(\beta, L)$. Furthermore, assume that

$$\int |u|^{\beta} |K(u)| \mathrm{d}u < \infty$$

and

$$\int K^2(u) \mathrm{d}u < \infty.$$

For some known constant $c > 0$, set $h = cn^{-\frac{1}{2\beta+1}}$. Then, for $n \geq 1$, $\hat{f}_n$ satisfies

$$\sup_{x \in \mathbb{R}} \sup_{f \in F(\beta, L)} \mathbb{E}_f[(\hat{f}_n(x) - f(x))^2] \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where C is a positive constant that depends on β, c, L and K.

The proof of the theorem relies on verifying that the assumptions of Proposition 2 are satisfied uniformly and then applying both Proposition 1 and 2.

The conclusion of Theorem 1 establishes the **rate of convergence** of the kernel density estimator as $n^{-\frac{\beta}{2\beta+1}}$.

2.3 Higher-order kernels

Theorem 1 assumes that bounded, order-l kernels exist. We have already seen examples of kernel for $l = 2$. Now, we provide a method for constructing higher-order kernels.

Start with the orthonormal basis of Legendre polynomials in $L_2([-1, 1], \mathrm{d}x)$ defined by

$$\varphi_0(x) = 1, \qquad \varphi_m(x) = \sqrt{\frac{2m+1}{2}} \frac{1}{2^m m!} \frac{\mathrm{d}^m}{\mathrm{d}x^m} \left[(x^2 - 1)^m \right], \ m = 1, 2, \ldots$$

This basis satisfies the property that

$$\int_{-1}^{1} \varphi_m(u) \varphi_k(u) \mathrm{d}u = \mathbf{1}(m = k)$$

Proposition 3 (Constructing *l-th* order kernel)
The function defined by

$$K(u) = \sum_{m=0}^{l} \varphi_m(0) \varphi_m(u) \mathbf{1}(|u| \leq 1)$$

is an l-th order kernel.

Remark 1 (Positivity constraint) It follows from Definition 2 that some kernels may take negative values on a set of positive Lebesgue measure. As a result, the estimators $\hat{f}_n$ based on such kernels can also take negative values. This property is sometimes emphasized as a drawback of estimators with higher order kernels, since the true density f itself will always be nonnegative. However, this remark is of minor importance because we can always use the positive part estimator:

$$\hat{f}_n^+(x) \triangleq \max\left\{0, \hat{f}_n(x)\right\}$$

whose risk is smaller than or equal to the risk of $\hat{f}_n$:

$$\mathbb{E}_f\left[(\hat{f}_n^+(x) - f(x))^2\right] \leq \mathbb{E}_f\left[(\hat{f}_n(x) - f(x))^2\right], \quad \forall x \in \mathbb{R}$$

In particular, Theorem 1 remains valid if we replace $\hat{f}_n$ by $\hat{f}_n^+$. Thus, the estimator $\hat{f}_n^+$ is nonnegative and attains the fast convergence rates associated with higher order kernels.

3 Multivariate extension

The collection of kernel density estimators we have studied so far can also be extended to the multidimensional case ($d > 1$). For example, if $d = 2$, supposed $\{(X_i^1, X_i^2)\}_{i=1}^n \in \mathbb{R}^2$ are i.i.d. with a joint density $f(\mathbf{x}) = f(x^1, x^2)$, the kernel estimator takes the form:

$$\hat{f}_n(\mathbf{x}) = \frac{1}{nh^2} \sum_i K(X_i^1; x^1) K(X_i^2; x^2)$$

for some kernel K and bandwidth $h > 0$. This multiplicative nature of the kernel applied to each dimension is referred to as the **product kernel**.

For more general d, we can write the estimator as

$$\hat{f}(x) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^n \left\{ \prod_{j=1}^d K\left(\frac{X_i^j - x^j}{h_j}\right) \right\}.$$

Note that here one can choose a different kernel and bandwidth based on the covariate, if desired. The risk of this estimator is given by

$$\frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^d h_j^4 \int f_{jj}^2(x)dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jj} f_{kk} dx \right] + \frac{\left(\int K^2(x)dx\right)^d}{nh_1 \cdots h_d}$$

where f_{jj} is the second partial derivative of f. The optimal bandwidth for this estimator is given by $h_i \equiv h = O\left(n^{-1/(4+d)}\right)$, leading to a risk of order $O\left(n^{-4/(4+d)}\right)$. We see that the risk increases rapidly with dimension. From this formulation we can see the effect of increased dimensionality on

the estimator. This curse of dimensionality implies that the accuracy of the estimator deteriorates quickly as dimension increases.

To get a sense of how serious this problem is, consider the following table from Silverman (1986) which shows the sample size required to ensure a MSE less than 0.1 at 0 when the true density is a multivariate normal and the optimal bandwidth is used.

Dimension	Sample Size
1	4
2	19
3	67
4	223
5	768
6	2790
7	10,700
8	43,700
9	187,000
10	842,000

This is clearly not good for practical purposes. As a result of this, the confidence intervals (constructed analogously to the one-dimensional case) get increasingly wide as d increases. It is important to highlight here that the problem is not the method of estimation, but rather, the wide bands correctly reflect the difficulty of the problem. We will discuss this curse of dimensionality and other estimation tools that may be used to resolve it later in the course.

4 Connecting density estimation and regression

There is a useful trick for converting a density estimation problem into a regression problem. This trick was made rigorous by Nussbaum (1996). By converting to regression, we can use all the tools we developed in the previous chapter, including the method for constructing confidence intervals.

Suppose, as we have thus far, $X_1, \ldots, X_n \sim F$ with density $f = F'$. WLOG, suppose the data are on $[0, 1]$. Divide the interval $[0, 1]$ into k equal width bins where $k \approx n/10$. Define

$$Y_j = \sqrt{\frac{k}{n}} \times \sqrt{N_j + \frac{1}{4}}$$

where N_j is the number of observations in bin j. Then,

$$Y_j \approx r\left(t_j\right) + \sigma\varepsilon_j$$

where $\varepsilon_j \sim N(0,1), \sigma = \sqrt{\frac{k}{4n}}, r(x) = \sqrt{f(x)}$ and t_j is the midpoint of the j^{th} bin. To see why, let B_j denote the j^{th} bin and note that

$$N_j \approx \text{Poisson}\left(n \int_{B_j} f(x)dx\right) \approx \text{Poisson}\left(\frac{nf\left(t_j\right)}{k}\right)$$

such that $\mathbb{E}[N_j] = \mathbb{V}(N_j) \approx nf(t_j)/k$. Applying the delta method, we see that $\mathbb{E}[Y_j] \approx \sqrt{f(t_j)}$ and $\mathbb{V}(Y_j) \approx k/(4n)$.

We have thus converted the density estimation problem into a non-parametric regression problem with equally spaced X_i's and constant variance. We can now apply our favorite non-parametric regression method to generate an estimator $\hat{r}_n$ and set

$$\hat{f}_n(x) = \frac{(r^+(x))^2}{\int_0^1 (r^+(s))^2 \mathrm{d}s},$$

with $r^+(x) = \max\{\hat{r}_n(x), 0\}$. In particular, we can construct confidence intervals.

5 Asymptotic (sub-)optimality

We have evaluated the point-wise MSE behavior of the kernel density estimator. However, another method of assessing the performance of the estimator could be to analyze the integrated mean squared-error (IMSE). This type of evaluation may be more relevant when one wants good control on the expected error as opposed to the error at any single evaluation point. The IMSE is defined as

$$\mathbb{E}_f[\int (\hat{f}_n(x) - f(x))^2 \mathrm{d}x],$$

which by the Tonelli-Fubini theorem and the bias-variance decomposition simplifies to

$$\int \text{MSE}(\hat{f}_n(x))\mathrm{d}x = \int \text{Bias}^2(\hat{f}_n(x))\mathrm{d}x + \int \mathbb{V}(\hat{f}_n(x))\mathrm{d}x.$$

Now in order to bound the IMSE, we repeat the exercise of bounding the integrated bias and variance terms separately under analogous assumptions as for bounding the MSE. We start with the variance bound:

Proposition 4 (Integrated variance bound)
Suppose $K : \mathbb{R} \to \mathbb{R}$ satisfies

$$\int K^2(u)\mathrm{d}u < \infty.$$

Then for any $h > 0, n \geq 1$ and any probability density function f,

$$\int \sigma^2(x)\mathrm{d}x \leq \frac{1}{nh} \int K^2(u)\mathrm{d}u.$$

Note that here that like in Proposition 2, Proposition 4 does not make any structural assumptions on f. However, the bias bound will require structural assumptions on f similar to the Hölder class restriction in Proposition 1.

Definition 4 (Sobolev class)**.** *Let $\beta \geq 1$ be an integer and $L > 0$. The Sobolev class $S(\beta, L)$ is the set of all $\beta - 1$ times differentiable functions $f : \mathbb{R} \to \mathbb{R}$ with $f^{(\beta-1)}$ being absolutely continuous and*

$$\int (f^{(\beta)}(x))^2 \mathrm{d}x \leq L.$$

Then, let $\mathcal{F}(\beta, L) = \{f \in S(\beta, L) : f \geq 0, \int f(x)\mathrm{d}x = 1\}$. Then, the bias bound is given by the following proposition:

Proposition 5 (Integrated bias bound)
Assume $f \in \mathcal{F}(\beta, L)$ and K be an order β kernel with

$$\int |u|^\beta |K(u)| \mathrm{d}u < \infty.$$

Then, for any $h > 0, n \geq 1$,

$$\int Bias^2(x)\mathrm{d}x \leq C^2 h^{2\beta}.$$

where

$$C = \frac{L}{l!} \int |u|^\beta |K(u)| \mathrm{d}u.$$

From these two results, we get the IMSE bound

$$\text{IMSE} \leq C^2 h^{2\beta} + \frac{1}{nh} \int K^2(u)\mathrm{d}u.$$

The IMSE is minimized at

$$h_n^* = \left(\frac{\int K^2}{2\beta C^2} \right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$$

with optimal order

$$\text{IMSE} = O\left(n^{-\frac{2\beta}{2\beta+1}}\right).$$

Notice that the behavior of the IMSE is very similar to that of the MSE.

The natural assumption from the MSE and IMSE bounds would be to select the bandwidth and kernel that minimizes the expressions for some chosen density f. Lets first look at the following result

Lemma 1

Suppose the kernel K satisfies: $\int K^2 < \infty$, $\int u^2|K(u)|\mathrm{d}u < \infty$ and define $S_K = \int u^2 K(u) \neq 0$. Assume further that f is differentiable, the first derivative is absolutely continuous and $\int (f^{(2)}(x))^2 \mathrm{d}x < \infty$. Then, for all $n \geq 1$,

$$IMSE = \left[\frac{1}{nh}\int K^2(u)\mathrm{d}u + \frac{h^4}{4}S_K^2 \int (f^{(2)}(x))^2 \mathrm{d}x\right](1 + o(1)),$$

where the $o(1)$ term depends on f and approaches zero as $h \to 0$.

Using this result, if we minimize for h and K, we obtain the "optimal" bandwidth

$$h^{IMSE}(K) = \left(\frac{\int K^2}{nS_K^2 \int (f^{(2)})^2}\right)^{1/5}$$

and the "optimal" kernel shape

$$K^{IMSE}(u) = \frac{3}{4}(1 - u^2)_+.$$

Note that this is actually just the Epanechnikov kernel. Of course, in practice using the IMSE-optimal bandwidth is not possible since it depends on the second derivative of the true density function. This idea of substituting in h^{IMSE} to the KDE, is called the **oracle estimator**, since it depends on unknown quantities.

From Lemma 1, the asymptotic IMSE, plugging in for the Epanechnikov kernel and the IMSE-optimal bandwidth, is

$$\lim_{n\to\infty} n^{4/5}\mathbb{E}[\int (f^E(x) - f(x))^2 \mathrm{d}x] = \frac{3^{4/5}}{5^{1/5}4}\left(\int (f^{(2)}(x))^2 \mathrm{d}x\right)^{1/5}.$$

It may seem reasonable to claim that this is the best IMSE that can be obtained and the oracle estimator is optimal, given that we have use IMSE optimal kernel and bandwidth. However, this is not accurate. The following result should explain why:

Lemma 2

Assume that f is differentiable, the first derivative is absolutely continuous and $\int (f^{(2)}(x))^2 \mathrm{d}x < \infty$. Let K be an order 2 kernel (i.e., $S_K = 0$) with $\int K^2 < \infty$. Then, for any $\varepsilon > 0$, the kernel estimator with $h = n^{-1/5}\varepsilon^{-1}\int K^2$ satisfies

$$\limsup_{n\to\infty} n^{4/5}\mathbb{E}[\int (\hat{f}_n(x) - f(x))^2 \mathrm{d}x] \leq \varepsilon.$$

The same is true for $\hat{f}_n^+ = \max(0, \hat{f}_n)$.

We see that for all $\varepsilon > 0$ small enough the estimators $\hat{f}_n$ and $\hat{f}_n^+$ of Lemma 2 have smaller asymptotic IMSE than the Epanechnikov oracle under the same assumptions on f. Note that $\hat{f}_n$, $\hat{f}_n^+$ are completely data-dependent estimators, not oracles. So, if the performance of estimators is

measured by their asymptotic IMSE for a fixed f there are several estimators that are strictly better than the Epanechnikov oracle. Furthermore, Lemma 2 implies

$$\inf_{T_n} \limsup_{n\to\infty} n^{4/5}\mathbb{E}_f[\int (T_n(x) - p(x))^2 \mathrm{d}x] = 0,$$

where $\inf_{T_n}$ is the infimum over all the kernel estimators. The positive part estimator $\hat{f}_n^+$ is included in Lemma 2 on purpose. In fact, it is often argued that one should use nonnegative kernels because the density itself is nonnegative. This argument would support the "optimality" of the Epanechnikov kernel because it is obtained from minimization of the asymptotic IMSE over nonnegative kernels.

Lemma 2 constructs a counterexample. The estimators $\hat{f}_n$ and $\hat{f}_n^+$ given in this lemma are by no means advocated for as being good. In fact, they can be rather counter-intuitive. For example, notice that the bandwidth h contains an arbitrarily large constant factor ε^{-1}. This factor is used to diminish the variance term, whereas, for fixed density f, the condition $\int u^2 K(u)\mathrm{d}u = 0$ eliminates the main bias term if n is large enough, that is, if $n \geq n_0$, starting from some n_0 that depends on f. This elimination of the bias is possible for fixed f but not uniformly over the Sobolev class with $\beta = 2$ (and more generally, for many large class of functions $f \in \mathcal{F}$). The message of the lemma, then, is to show that as soon as we consider the problem of asymptotic optimality for *a fixed density*, even counter-intuitive estimators can outperform the oracle.

This idea of fixed f asymptotics providing inconsistent ideas of optimality will be a strong motivator for the minimax theory and uniform optimality we will discuss in the next few chapters.

References

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and gaussian white noise. *The Annals of Statistics*, 24(6):2399–2430.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Routledge.