CS861: Theoretical Foundations of Machine Learning Lecture 14 - 10/04/2024

University of Wisconsin–Madison, Fall 2023

Lecture 14: Nonparametric Regression and Density Estimation

Lecturer: Kirthevasan Kandasamy *Scribed by: Ziyang Cai, Jakob Juergens*

Disclaimer 2: *These notes are an updated version of existing notes created in 2023 by Haoran Xiong, Zhihao Zhao, Haoyue Bai, Zexuan Sun, Elliot Pickens, and Yuya Shimizu to reflect the contents covered in 2024.*

In this lecture, we will develop upper and lower bounds for nonparametric regression and show that the minimax rate is $\widetilde{\Theta}(n^{-2/3})$. We will also briefly introduce nonparametric density estimation.

1 Nonparametric Regression

Let $\mathcal{F}$ be a class of bounded L-Lipschitz functions in $[0, 1]$.

$$\mathcal{F} = \{f : [0,1] \to [0,1]; |f(x_1) - f(x_2)| \leq L|x_1 - x_2|\}$$

It is worth mentioning that the upper bound (here 1) can be chosen arbitrarily. The same argument can be applied after employing a trivial rescaling. Assume that we observe a dataset $S = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ that is drawn i.i.d from an unknown distribution $P_{XY} \in \mathcal{P}$, where

$$\mathcal{P} = \left\{ P_{XY} \middle| \ 0 < \alpha_0 \leq p(x) \leq \alpha_1 < \infty, \ f(x) \triangleq \mathbb{E}[Y|X = x] \in \mathcal{F}, \ \mathrm{Var}(Y|X = x) \leq \sigma^2 \right\},$$

and $p(x)$ is the marginal density of X. We wish to estimate the nonparametric regression function using the following loss:

$$\ell(P_{XY}, g) \triangleq \int_0^1 [f(x) - g(x)]^2 p(x) dx.$$

Then, the minimax risk is defined as follows:

$$\begin{aligned} R_n^* =& \inf_{\widehat{f}} \sup_{P_{XY} \in \mathcal{P}} \mathbb{E}_{S \sim P_{XY}} \left[\ell(P_{XY}, \widehat{f}) \right] \\ =& \inf_{\widehat{f}} \sup_{P_{XY} \in \mathcal{P}} \mathbb{E}_{S \sim P_{XY}} \left[\int_0^1 (f(x) - \widehat{f}(x))^2 p(x) dx \right] \end{aligned}$$

We want to show that the minimax risk R_n^* is of the order $\Theta(n^{-2/3})$, establishing rate optimality of the Nadaraya-Watson estimator that we will introduce, in two steps:

1. Establish a lower bound with Fano's method
2. Get an upper bound by using the Nadaraya-Watson Estimation

1.1 Lower bound

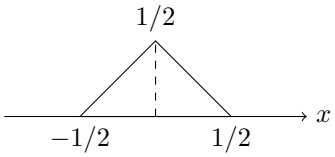
Noticing that $\ell(P_{XY}, g)$ as defined above cannot be written in the form of $\ell = \Phi \circ \rho$, meaning that we cannot utilize the previously established results, we circumvent this problem by constructing a suitable sub-class $\mathcal{P}^{''}$ of $\mathcal{P}$ as follows:

$$\mathcal{P}^{''} = \{P_{XY} \in \mathcal{P};\ p(x) = 1\}\ ^{1}$$

Then $R_n^* \geq \inf_{\widehat{f}} \sup_{P_{XY} \in \mathcal{P}^{''}} \mathbb{E}_{S \sim P_{XY}} \left[\int (f(x) - \widehat{f}(x))^2 dx \right]$, and now we can write $\Phi \circ \rho(f_1, f_2) = \|f_1 - f_2\|_2^2$.

1.1.1 Constructing alternatives

Define $\psi(x) = \begin{cases} x + \frac{1}{2} & \text{if } x \in [-\frac{1}{2}, 0), \\ -x + \frac{1}{2} & \text{if } x \in [0, \frac{1}{2}], \\ 0 & \text{o.w.} \end{cases}$



Note that ψ is 1-Lipschitz, and $\int \psi^2(x)dx = 1/12$. Now let $h > 0$ (we'll specify its value later) and let $m = \frac{1}{h}$, we construct a new function class

$$\mathcal{F}^{'} = \left\{ f_\omega \;\middle|\; f_\omega(\cdot) = \sum_{j=1}^{m} \omega_j \phi_j(\cdot),\ \omega \in \Omega_m \right\}$$

where Ω_m is the Varshamov-Gilbert pruned hypercube of $\{0,1\}^m$, and $\phi_j(x) = Lh \cdot \psi\left(\frac{x-(j-1/2)h}{h}\right)$.

Figure 1: Depiction of ϕ_j and w when $w = \{0, 0, 1, 0, 1\}$.

Since

$$|\phi_j^{'}(x)| = \left| Lh \underbrace{\phi'\left(\frac{x-(j-1/2)h}{h}\right)}_{1} \frac{1}{h} \right| = L,$$

we know that f_ω is L-Lipschitz. We can now define our alternatives:

$$\mathcal{P}^{'} = \left\{ P_{XY} \;\middle|\; p(x) \text{ uniform}, f(x) = \mathbb{E}[Y|X = x] \in \mathcal{F}^{'},\ Y|X = x \sim \mathcal{N}(f(x), \sigma^2) \right\}.$$

As $\mathcal{F}' \subset \mathcal{F}$, we have $\mathcal{P}^{'} \subset \mathcal{P}'' \subset \mathcal{P}$.

1.1.2 Lower bound on separation of alternatives, $\|f_\omega - f_{\omega'}\|$

To better organize our result, we first calculate,

$$\int_{\frac{j-1}{m}}^{\frac{j}{m}} \phi_j^2(x)dx = \int_{\frac{j-1}{m}}^{\frac{j}{m}} L^2h^2 \cdot \psi^2\left(\frac{x-(j-1/2)h}{h}\right) dx = \int_{-1/2}^{1/2} L^2h^3\psi^2(u)du = \frac{L^2h^3}{12}.$$

[1] here we use the uniform density $p(x) = 1$ for convenience, but any fixed density $p(x)$ will still induce a metric.

We then have, $\forall \omega, \omega' \in \Omega_m$,

$$\begin{aligned}
\rho^2(f_\omega, f_{\omega'}) &= \int_0^1 (f_\omega - f_{\omega'})^2 dx \\
&= \sum_{j=1}^m \int_{\frac{j-1}{m}}^{\frac{j}{m}} (\omega_j \phi_j(x) - \omega'_j \phi_j(x))^2 dx \\
&= \sum_{j=1}^m \mathbf{1}\{\omega_j \neq \omega'_j\} \int_{\frac{j-1}{m}}^{\frac{j}{m}} \phi_j^2(x) dx \\
&= \frac{L^2 h^3}{12} \sum_{j=1}^m \mathbf{1}\{\omega_j \neq \omega'_j\} = \frac{L^2 h^3}{12} \cdot H(\omega_j, \omega'_j)
\end{aligned}$$

where $H(\cdot, \cdot)$ is the Hamming distance, and the last equation holds because of the definition of it.

Since $\omega, \omega' \in \Omega_m$, by Varshamov-Gilbert lemma, $H(\omega_j, \omega'_j) \geq \frac{m}{8} = \frac{1}{8h}$. Then we have

$$\min_{\omega_j, \omega'_j} \rho(f_\omega, f_{\omega'}) \geq \sqrt{\frac{L^2 h^3}{12}} \sqrt{H(\omega, \omega')} = \frac{Lh}{\sqrt{96}} \triangleq \delta,$$

where δ is called the separation between hypotheses.

1.1.3 Upper bound KL

Next, we will upper bound the maximum KL divergence between our alternatives. Let $P_\omega, P_{\omega'} \in \mathcal{P}'$. Then,

$$\begin{aligned}
KL(P_\omega, P_{\omega'}) &= \int_{\mathcal{X} \times \mathcal{Y}} p_\omega \log \frac{p_\omega}{p_{\omega'}} \\
&= \int_0^1 \int_{-\infty}^{\infty} p_\omega(x) p_\omega(y|x) \log \frac{p_\omega(x) p_\omega(y|x)}{p_{\omega'}(x) p_{\omega'}(y|x)} dy dx \\
&= \int_0^1 \int_{-\infty}^{\infty} p_\omega(y|x) \log \frac{p_\omega(y|x)}{p_{\omega'}(y|x)} dy dx \qquad (\text{as } p_\omega(x) = p_{\omega'}(x) = 1) \\
&= \int_0^1 KL\left(\mathcal{N}(f_\omega(x), \sigma^2), \mathcal{N}(f_{\omega'}(x), \sigma^2)\right) dx \qquad (\text{as } Y|X = x \sim \mathcal{N}(f(x), \sigma^2)) \\
&= \frac{1}{2\sigma^2} \int_0^1 \left(f_\omega(x) - f_{\omega'}(x)\right)^2 dx \\
&= \frac{1}{2\sigma^2} \rho^2(f_\omega, f_{\omega'}) = \frac{L^2 h^3 \cdot H(\omega, \omega')}{24\sigma^2}.
\end{aligned}$$

Then since $\max_{\omega, \omega'} H(\omega, \omega') \leq m = 1/h$,

$$\max_{\omega, \omega'} KL(P_\omega, P_{\omega'}) = \frac{L^2 h^3}{24\sigma^2} \max_{\omega, \omega'} H(\omega, \omega') \leq \frac{L^2 h^2}{24\sigma^2}.$$

1.1.4 Apply local Fano's method

In order to apply Fano's method, we need to satisfy $\max_{\omega, \omega'} KL(P_\omega, P_{\omega'}) \leq \frac{\log |\mathcal{P}'|}{4n}$. Recall that by the Varshamov-Gilbert lemma, $|\mathcal{P}'| \geq 2^{m/8}$. Thus, it is sufficient if we have,

$$\frac{L^2 h^2}{24\sigma^2} \leq \frac{\log(2^{m/8})}{4n} = \frac{m \log 2}{32n} = \frac{\log 2}{32nh}.$$

This suggests that we could choose $h = \left(\frac{3\log 2}{4}\right)^{\frac{1}{3}} \frac{\sigma^{2/3}}{n^{1/3}L^{2/3}}$. Thus the separation between hypotheses $\delta = C_1 \frac{L^{1/3}\sigma^{2/3}}{n^{1/3}}$, where C_1 is some constant. By the local Fano's method, we obtain the following.

$$R_n^* \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) = \frac{1}{8}\delta^2 = C_2 \frac{L^{2/3}\sigma^{4/3}}{n^{2/3}}.$$

Remark: To apply the local Fano's method, it's required that $|\mathcal{P}'| \geq 16$. It's sufficient to have $|\mathcal{P}'| \geq 2^{m/8} \geq 16$, i.e. $m = 1/h \geq 32$, which means the following must hold:

$$h = \left(\frac{3\log 2}{4}\right)^{\frac{1}{3}} \frac{\sigma^{2/3}}{n^{1/3}L^{2/3}} \leq \frac{1}{32} \quad \Longrightarrow \quad n \geq C_3 \frac{\sigma^2}{L^2} \text{ for some constant } C_3.$$

1.2 Upper Bound

To upper-bound the minimax risk, we introduce a specific estimator and establish its risk. Later, we will introduce the class of Nadaraya-Watson estimators and show that our current estimator is a special case of it. Our estimator $\widehat{f}(t)$ is defined as follows, where h is a bandwidth parameter.

$$\widehat{f}(t) = \begin{cases} \text{clip}\left(\frac{1}{N(t)}\sum_{i=1}^n Y_i \mathbf{1}\{X_i \in [t-h, t+h]\}, 0, 1\right) & \text{if } N(t) > 0 \\ 1/2 & \text{if } N(t) = 0 \end{cases}$$

where $N(t) = \sum_{i=1}^n \mathbf{1}\{X_i \in [t-h, t+h]\}$ and $\text{clip}(x, 0, 1)$ means that

$$\text{clip}(x, 0, 1) = \begin{cases} x, & 0 \leqslant x \leqslant 1 \\ 0, & x < 0 \\ 1, & x > 1. \end{cases}$$

By definition,

$$\begin{aligned} R(P_{XY}, \hat{f}) &= \mathbb{E}_S\left[\int (\hat{f}(x) - f(x))^2 p(x) dx\right] \leqslant \alpha_1 \mathbb{E}_S\left[\int (\hat{f}(x) - f(x))^2 dx\right] \\ &= \alpha_1 \int_0^1 \underbrace{\mathbb{E}_S\left[(\hat{f}(t) - f(t))^2\right]}_{\text{err}(t)} dt. \end{aligned}$$

We will next provide a pointwise bound on $\text{err}(t)$ which will translate to an integrated bound. The calculations for the pointwise bound are very similar to an example we saw previously so we will only provide an overview and highlight the differences. Let $G_t = \{N(t) \geqslant \alpha_0 nh\}$ denote the good event that there were a sufficient number of samples in a $2h$ neighborhood of t. We have,

$$\begin{aligned} \mathbb{P}(G_t^c) &= \mathbb{P}\left(\sum_{i=1}^n \mathbf{1}\{X_i \in [t-h, t+h]\} < \alpha_0 nh\right) \\ &= \mathbb{P}\left(\sum_{i=1}^n (\mathbf{1}\{X_i \in [t-h, t+h]\} - \mathbb{P}([t-h, t+h])) < \alpha_0 nh - n\mathbb{P}([t-h, t+h])\right), \end{aligned}$$

where $\mathbb{P}([t-h, t+h]) = \int_{t-h}^{t+h} p(x)dx \geqslant 2\alpha_0 h$. Thus we have $\alpha_0 nh - n\mathbb{P}([t-h, t+h]) \leqslant -\alpha_0 nh$. By Hoeffding's inequality, it follows that $\mathbb{P}(G^c) \leqslant \exp(-2\alpha_0^2 nh^2)$. We can decompose $\text{err}(t)$ as

$$\text{err}(t) = \mathbb{E}_S\left[(\hat{f}(t) - f(t))^2 \mid G_t\right] \cdot P(G_t) + \mathbb{E}_S\left[(\hat{f}(t) - f(t))^2 \mid G_t^c\right] \cdot P(G_t^c)$$

By following the calculations from our previous example, we can show

$$\mathbb{E}_S\left[(\hat{f}(t) - f(t))^2\right] \leqslant L^2 h^2 + \frac{\sigma^2}{nh} + e^{-2\alpha_0^2 nh^2}.$$

Therefore,

$$R(P_{XY}, \hat{f}) \leqslant \alpha_1 \int_0^1 \mathbb{E}_S \left[(\hat{f}(t) - f(t))^2 \right] dt \leqslant \alpha_1 \left(L^2 h^2 + \frac{\sigma^2}{nh} + e^{-2\alpha_0^2 nh^2} \right).$$

Now we choose $h = \sigma^{2/3} L^{-2/3} n^{-1/3}$, which implies that

$$R(P_{XY}, \hat{f}) \leqslant 2\alpha_1 \frac{\sigma^{4/3} L^{2/3}}{n^{2/3}} + \alpha_1 \exp\left(-2\alpha_0^2 \frac{\sigma^{4/3} n^{1/3}}{L^{4/3}} \right).$$

This shows that the estimator under consideration achieves the same convergence rate as our lower bound when using appropriately chosen parameters. Therefore the minimax risk is of order $R^* = \Theta(n^{-2/3})$.

Remark: Had we used the multiplication Chernoff bound instead of Hoeffding's inequality, we would have found the following bounds:

$$\mathbb{P}(G^c) \leqslant e^{-\alpha_0 nh/8},$$
$$R(P_{XY}, \hat{f}) \leqslant 2\alpha_1 \frac{\sigma^{4/3} L^{2/3}}{n^{2/3}} + \alpha_1 \exp\left(-\frac{\alpha_0}{4} \frac{\sigma^{2/3} n^{2/3}}{L^{2/3}} \right).$$

For i.i.d Bernoulli random variables with success probability close to 0 or 1, the multiplicative Chernoff bound can provide a tighter bound than Hoeffding's inequality. This does not significantly alter our conclusions in this example but may be significant in other use cases.

1.3 Nadaraya-Watson Estimator

A Nadaraya-Watson Estimator (also known as the kernel estimator) is defined to be

$$\hat{f}(t) = \sum_{i=1}^n y_i w_i(t)$$

$$w_i(t) = \begin{cases} \frac{K((t-X_i)/h)}{\sum_{j=1}^n K((t-X_j)/h)} & \text{if } \sum_{j=1}^n K((t-X_j)/h) > 0 \\ 0 & \text{otherwise} \end{cases},$$

where $K : \mathbb{R} \to \mathbb{R}$ is called a smoothing kernel. For example, in our previous case, the smoothing kernel is $K(u) = \mathbf{1}\{|u| \leqslant 1\}$, which is sufficient for Lipschitz smooth functions. Other kernel choices can lead to better rates under stronger smoothness assumptions. One such class of assumptions is that the regression function falls into a Hölder class in $\mathbb{R}^d$, which is denoted by $\mathcal{H}(\beta, L)$ and defined to be the set of all functions whose $(\beta - 1)$th order partial derivatives are L-Lipschitz. The minimax rate in this class is $\Theta(n^{-2\beta/(2\beta+d)})$. To achieve this rate, we will need to design smarter kernels in the Nadaraya-Watson estimator. The same rates hold for density estimation in the Hölder class.

2 Density Estimation

We will briefly introduce lower and upper bounds for density estimation. Let $\mathcal{F}$ be the class of bounded Lipschitz functions, i.e.

$$\mathcal{F} = \{f : [0,1] \to [0,B] \ : \ |f(x_1) - f(x_2)| \leqslant L|x_1 - x_2| \ \forall x_1, x_2 \in [0,1]\}.$$

Here, we choose B as the upper bound instead of 1 to allow for nontrivial functions that fulfill the integral constraint imposed by the fact that the functions of interest are densities. The corresponding nonparametric family of alterantives is then defined to be

$$\mathcal{P} = \{P : \text{ The p.d.f. } p \text{ of } P \text{ is in } \mathcal{F}\}.$$

Suppose we observe $S = (X_1, \ldots, X_n)$ drawn i.i.d from some distribution $P \in \mathcal{P}$. We wish to estimate the p.d.f. under the L_2 loss, i.e.

$$\Phi \circ \rho(P_1, P_2) = \int (p_1(t) - p_2(t))^2 dt.$$

By definition, the minimax risk is

$$R_n^* = \inf_{\widehat{p}} \sup_{p \in \mathcal{F}} \mathbb{E}_S \left[||\widehat{p}(S) - p||_2^2 \right].$$

2.1 Lower bound

We will first prove a lower bound via Fano's method.
Step 1: Construct alternatives

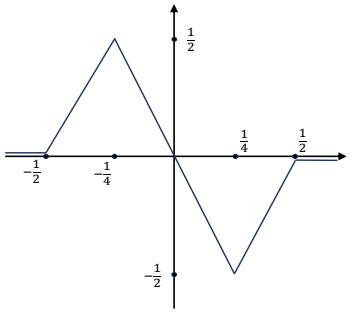


Figure 2: Illustrative figure for the function $\psi(\lambda)$.

Consider the function ψ illustrated in Figure 2. The following facts are straightforward to verify.

- ψ is 1-Lipschitz, meaning that for any two inputs λ_1 and λ_2: $|\psi(\lambda_1) - \psi(\lambda_2)| \leq |\lambda_1 - \lambda_2|$;
- $\int \psi = 0$, which indicates that the area under the curve of the function, over its entire domain, sums up to zero;
- $-\frac{1}{4} \leq \psi \leq \frac{1}{4}$, which gives the range of the function;
- $\int \psi^2 = \frac{1}{48}$, which the squared integral of ψ.

To construct the alternatives, let h be a positive number ($h > 0$) that we will decide later. Let $m = \frac{1}{n}$. The alternative function space $\mathcal{F}^{'}$ is:

$$\mathcal{F}^{'} = \left\{ P_\omega : P_\omega(\cdot) = 1 + \sum_{i=1}^{m} \omega_j \Phi_j(\cdot);\ \omega \in \Omega_m \right\}.$$

This space defines a set of functions P_ω formed by a linear combination of basis function $\phi_j(t)$. The vector ω is an element of the VG pruned hypercube Ω_m. The basis function is defined as:

$$\phi_j(t) = Lh\psi\left(\frac{t - (j - \frac{1}{2}) \cdot h}{h}\right),$$

where L denotes the Lipschitz constant, and h is the bandwidth. Figure 3 provides a visual representation of one such alternative. We can check that P_ω is L-Lipschitz and P_ω is a valid density function because $\int P_\omega = 1$. We need $0 \leq P_\omega \leq B$, this is satisfied if $h \leq \frac{4}{L} \min(B - 1, 1)$.

Step 2: Lower bound the minimum distance $\rho(p_\omega, p_{\omega'})$

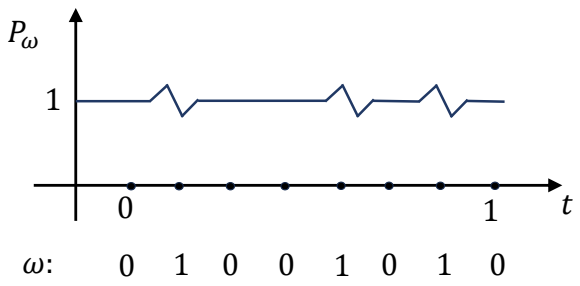


Figure 3: Illustrative figure for the example p_ω.

The objective of this step is to determine a lower bound for the difference between p_ω and $p_{\omega'}$. We can bound the difference from below in the following way: $\forall \omega, \omega' \in \Omega_m$,

$$\rho^2(p_\omega, p_{\omega'}) = \int_0^1 (p_\omega(t) - p_{\omega'}(t))^2 \, dt = \sum_{j=1}^{m} \int_{\frac{j-1}{m}}^{\frac{j}{m}} (1 + \omega_j \phi_j(t) - (1 + \omega_j^{'} \phi_j(t)))^2 \, dt$$
$$= \sum_{j=1}^{m} \mathbb{I}(\omega_j \neq \omega_j^{'}) \int_{\frac{j-1}{m}}^{\frac{j}{m}} \phi_j^2(t) \, dt = \frac{H(\omega, \omega^{'}) L^2 h^3}{48}$$

By the Varshamov-Gilbert lemma, we have $H(\omega, \omega^{'}) \geq \frac{m}{8} = \frac{1}{8h}$, allowing us to find the following.

$$\min_{\omega, \omega^{'}} ||p_\omega - p_{\omega'}|| = \sqrt{\frac{L^2 h^3}{48}} \cdot \min_{\omega, \omega^{'}} \sqrt{H(\omega, \omega^{'})} \geq \frac{Lh}{8\sqrt{6}} \triangleq \delta.$$

Step 3: Upper bound the maximum KL divergence
In this step, the goal is to determine an upper bound for the Kullback-Leibler (KL) divergence between two functions, p_ω and $p_{\omega'}$. Based on the definition of KL divergence and expanding KL for p_ω and $p_{\omega'}$:

$$KL(p_\omega, p_{\omega'}) = \int_0^1 p_\omega(t) \log \left(\frac{p_\omega(t)}{p_{\omega'}(t)} \right) \, dt$$
$$= \sum_{j=1}^{m} \mathbf{1}(\omega_j \neq \omega_j^{'}) \int_{\frac{j-1}{m}}^{\frac{j}{m}} (1 + \omega_j \phi_j(t)) \log \left(\frac{1 + \omega_j \phi_j(t)}{1 + \omega_j^{'} \phi_j(t)} \right) \, dt$$

After some algebra (you will do this in the homework), we have the following upper bound:

$$\forall \omega, \omega^{'}, KL(p_\omega, p_{\omega'}) \leq H(\omega, \omega^{'}) \frac{L^2 h^3}{48} \quad \Rightarrow \quad \max_{\omega, \omega^{'}} KL(p_\omega, p_{\omega'}) \leq \frac{L^2 h^2}{48}$$

because $H(\omega, \omega') \leq m = \frac{1}{h}$ A formal proof of this statement is left as an exercise in a homework assignment.

Step 4: Apply local Fano
Applying local Fano's inequality in this step, we derive conditions and constraints for the estimation problem. We want $\max_{\omega, \omega'} KL(p_\omega, p_{\omega'}) \leq \frac{\log(\omega)}{4n}$. This serves to upper bound the maximum KL divergence between any two functions in the set by a term that diminishes with increasing sample size n. It is sufficient if we have $\frac{L^2 h^2}{48} \leq \frac{\log(2^{\frac{m}{8}})}{4n} = \frac{\log(2)}{32nh}$. Thus, choose $h = C \frac{1}{n^{1/3} L^{2/3}}$, which determines the choice of h as a function of n and L. Then, we have the following:

$$R_n^* \geq \frac{1}{2} \Phi(\frac{Lh}{2 \times 8\sqrt{6}}) = C \frac{L^{2/3}}{n^{2/3}}.$$

This offers a lower bound on the risk, R_n^*, quantifying the estimation error. The following requirements ensure the validity and applicability of the above derivations.

- $m = \frac{1}{h} \geq 8$. This is necessary for the Varshamov-Gilbert lemma to be applicable.
- Cardinality of $\mathcal{F}'$: $|\mathcal{F}'| \geq 16 \Leftarrow 2^{\frac{m}{8}} \geq 16 \Leftarrow h \leq \frac{1}{32} \Rightarrow$ satisfied if $n \geq \frac{c}{L^2}$. This ensures a sufficient number of observations given the Lipschitz constant L.
- KL bounding condition: $h \leq \frac{2.72}{L}$. This imposes that the KL divergence between functions remains bounded and ties the bandwidth h to the Lipschitz constant L.

2.2 Upper Bound via Kernel Density Estimation

Kernel Density Estimation (KDE) is a nonparametric technique to estimate the probability density function of a continuous random variable. The Kernel Density Estimator (KDE) has the following form:

$$\hat{p}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{t - x_i}{h}\right),$$

where $\hat{p}(t)$ is the estimated density at point t, n is the number of data points, and x_i are the observed data points. h is a bandwidth parameter chosen by the researcher that plays a critical role in KDE. K is a (smoothing) kernel with the following properties:

1. **Normalization:**
$$\int K(u)du = 1,$$
This ensures the result will integrate to 1 over its entire domain, maintaining the fundamental property of a probability density function.

2. **Symmetry:**
$$K(u) = K(-u),$$
This property ensures that the kernel is symmetric around zero.