CS861: Theoretical Foundations of Machine Learning

Chapter 2: Nonparametric Methods

Kirthevasan Kandasamy

UW-Madison

Introduction

Nonparametric methods: Learn in a function class which cannot be completely characterized by a finite number of scalar parameters.

Examples in regression:

- Nonparametric: the class of all Lipschitz-continuous functions, $\{f; |f(x) - f(x')| \leq L\|x - x'\|_2\}$.
- Parametric: Linear class, $\{f; f(x) = \theta^\top x,\ \text{for some } \theta \in \mathbb{R}^d\}$.

Contents

Ch 2.1: Nonparametric regression

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

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

We observe an i.i.d dataset $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ drawn i.i.d from $P \in \mathcal{P}$ where,

$$\mathcal{P} = \Bigg\{ P; \quad \begin{aligned} & 0 < \alpha_0 \leq p(x) \leq \alpha_1 < \infty, \text{ where } p(x) \text{ is the marginal of } x \\ & \text{the regression function } f(\cdot) = \mathbb{E}[Y|X = \cdot] \in \mathcal{F}, \\ & \mathrm{Var}(Y|X = x) \leq \sigma^2, \text{ for all } x \in [0,1] \end{aligned} \Bigg\}.$$

Nonparametric regression (cont'd)

We wish to estimate the regression function in the following loss. For any $g \in \mathbb{R}^{[0,1]}$,

$$L(g, P) = \int (f(x) - g(x))^2 p(x) \mathrm{d}x, \quad \text{where, } f(\cdot) = \mathbb{E}_P[Y|X = \cdot].$$

The minimax risk is,

$$R_n^\star = \inf_{\widehat{f}} \sup_{P \sim \mathcal{P}} \mathbb{E}_{S \sim P^n}[L(\widehat{f}(S), P)] = \inf_{\widehat{f}} \sup_{P \sim \mathcal{P}} \mathbb{E}_{S \sim P^n} \left[\int (\widehat{f}(S)(x) - f(x))^2 p(x) \mathrm{d}x \right].$$

We will show that the minimax rate is $\Theta(n^{-2/3})$.
- Lower bound via Fano's method.
- Upper bound via the Nadaraya-Watson estimator.

Ch 2.1.1: Lower bounds for nonparametric regression

To apply our lower bound techniques, we cannot write the above loss in the form $\Phi \circ \rho$ where ρ is a metric Hence, let us consider $\mathcal{P}'' = \{P \in \mathcal{P}; p(x) = 1\} \subset \mathcal{P}$. Then, we can lower bound the minimax risk by

$$R_n^\star \geq \inf_{\widehat{f}} \sup_{P \subset \mathcal{P}''} \mathbb{E}_{S \sim P} \left[\underbrace{\int (f(x) - \widehat{f}(x))^2 \mathrm{d}x}_{\Phi \circ \rho(f, \widehat{f})} \right]$$

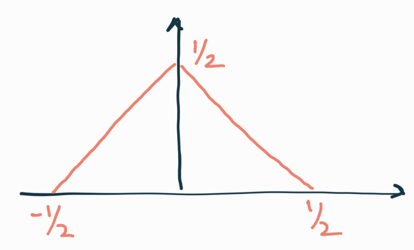
where, $\rho(f, g) = \|f - g\|_2$ and $\Phi(t) = t^2$.

Recall, Local Fano method. Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $N \geq 16$, $\rho(\theta(P_j), \theta(P_k)) \geq \delta$, and $\mathrm{KL}(P_j, P_k) \leq \log(N)/4n$ for all $j \neq k$. Then, $R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)$.

We will apply Fano's method in the following four steps: 1) Constructing alternatives. 2) Lower bound the minimum separation. 3) Upper bound the maximum KL. 4) Obtain final bound.

Step 1: Constructing Alternatives

Let us consider the following function,



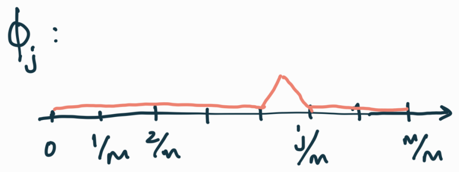
$$\psi(x) = \begin{cases} x + 1/2, & x \in [-1/2, 0], \\ -x + 1/2, & x \in [0, 1/2], \\ 0 & \text{otherwise.} \end{cases}$$

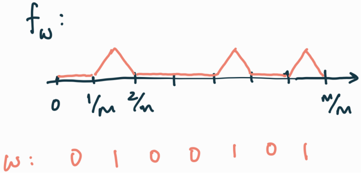
Note that, ψ is 1-Lipschitz, and $\int \psi^2 = 1/12$.

Step 1: Constructing Alternatives (cont'd)

Let $h > 0$ be a parameter to be chosen later. Let $m = 1/h$.
Let $\omega \in \{0,1\}^m$. Let us define f_ω as follows:

$$\phi_j(x) = Lh\psi\left(\frac{x-(j-1/2)h}{h}\right), \qquad f_\omega(x) = \sum_{j=1}^{m} \omega_j \phi_j(x).$$





We will show that each f_ω is L-Lipschitz:
- f_ω is 0 at the boundaries $(1/m, 2/m, \dots)$.
- It is sufficient to check Lipschitzness within each bump ϕ_j.
- The gradient of ϕ_j is $|\phi_j'(t)| = \left| Lh\psi'\left(\frac{x-(j-1/2)h}{h}\right)\frac{1}{h}\right| = L$.

Step 1: Constructing Alternatives (cont'd)

Alternatives: Let Ω_m be the Gilbert-Varshamov-pruned m-hypercube. Define

$$\begin{aligned}\mathcal{F}' &= \{f_\omega; \omega \in \Omega_m\},\\ \mathcal{P}' &= \{P; p(x) \text{ is uniform }, f(\cdot) = \mathbb{E}[Y|X=\cdot] \in \mathcal{F}',\\ &\qquad Y|X=x \sim \mathcal{N}(f(x), \sigma^2)\}.\end{aligned}$$

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

Let P_ω be the distribution in $\mathcal{P}'$ whose regression function is f_ω. Let p_ω be its density.

Step 2: Lower bounding the separation

For any $\omega, \omega' \in \Omega_m$, we have

$$\begin{aligned}
\rho^2(f_\omega, f_{\omega'}) &= \int_0^1 (f_\omega - f_{\omega'})^2 \\
&= \sum_{j=1}^{m} \int_{\frac{j-1}{m}}^{\frac{j}{m}} \left(\omega_j \phi_j - \omega'_j \phi_j\right)^2 \\
&= \sum_{j=1}^{m} \mathbb{1}(\omega_j \neq \omega'_j) \int_{\frac{j-1}{m}}^{\frac{j}{m}} \phi_j^2
\end{aligned}$$

Now, let us consider the integral,

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

Step 2: Lower bounding the separation (cont'd)

Recall, Gilbert-Varshamov bound. In the Gilbert-Varshamov pruned m-hypercube Ω_m of $\{0,1\}^m$, we have, *(i)* $|\Omega_m| \geq 2^{m/8}$, *(ii)* $H(\omega, \omega') \geq m/8$, $\forall\, \omega, \omega' \in \Omega_m$, *(iii)* $\mathbf{0}_m \in \Omega_m$.

Therefore,

$$\rho^2(f_\omega, f_{\omega'}) = H(\omega, \omega') \frac{L^2 h^3}{12}.$$

By the Gilbert-Varshamov bound, $H(\omega, \omega') \geq \frac{m}{8} = \frac{1}{8h}$.

Therefore, the separation can be lower bounded by,

$$\min_{\omega, \omega'} \rho(f_\omega, f_{\omega'}) \geq \frac{Lh}{\sqrt{96}} \triangleq \delta$$

.

Step 3: Upper bounding the KL

Let $p_\omega, p_{\omega'}$ be the joint densities for any $\omega, \omega' \in \Omega_m$, respectively. Then,

$$
\begin{aligned}
\mathrm{KL}(P_\omega, P_{\omega'}) &= \int_{\mathcal{X}\times\mathcal{Y}} p_\omega \log\left(\frac{p_\omega}{p_{\omega'}}\right) \\
&= \int_0^1 \int_{-\infty}^{\infty} \overbrace{p_\omega(x)}^{=1} p_\omega(y|x) \log\left(\frac{\overbrace{p_\omega(x)}^{=1} p_\omega(y|x)}{\underbrace{p_{\omega'}(x)}_{=1} p_{\omega'}(y|x)}\right) \mathrm{d}y\mathrm{d}x \\
&= \int_0^1 \mathrm{KL}\left(\mathcal{N}(f_\omega(x), \sigma^2), \mathcal{N}(f_{\omega'}(x), \sigma^2)\right) \mathrm{d}x = \frac{1}{2\sigma^2}\int_0^1 \left(f_\omega(x) - f_{\omega'}(x)\right)^2 \mathrm{d}x \\
&= \frac{1}{2\sigma^2}\sum_{j=1}^m \mathbb{1}(\omega_j \neq \omega_j') \int_{\frac{j-1}{m}}^{\frac{j}{m}} \left(\omega_j\phi_j - \omega_j'\phi_j\right)^2 \\
&= \frac{H(\omega,\omega')}{2\sigma^2}\cdot\frac{L^2h^3}{12}.
\end{aligned}
$$

Similar calculations to separaion.

Step 3: Upper bounding the KL (cont'd)

As $\max_{\omega,\omega'} H(\omega,\omega') \leq m = 1/h$, we have

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

Step 4: Obtain final bound

Recall, Local Fano method. Let S be an i.i.d dataset from some distribution $P \in \mathcal{P}$. Let $\{P_1, \dots, P_N\} \subset \mathcal{P}$ such that $N \geq 16$, $\delta \geq \rho(\theta(P_j), \theta(P_k))$, and $\mathrm{KL}(P_j, P_k) \leq \log(N)/4n$ for all $j \neq k$. Then, $R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right)$.

Recall, Gilbert-Varshamov bound. In the Gilbert-Varshamov pruned m-hypercube Ω_m of $\{0,1\}^m$, we have, *(i)* $|\Omega_m| \geq 2^{m/8}$, *(ii)* $H(\omega, \omega') \geq m/8$, $\forall\, \omega, \omega' \in \Omega_m$, *(iii)* $\mathbf{0}_m \in \Omega_m$.

We need $\max_{p_\omega, p_{\omega'}} \mathrm{KL}(P_\omega, P_{\omega'}) \leq \frac{\log(|\mathcal{P}'|)}{4n}$. As $|\mathcal{P}'| \geq 2^{m/8}$, it is sufficient if,

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

Therefore, choose h as follows,

$$h = \left(\frac{3\log(2)}{4}\right)^{1/3} \frac{\sigma^{2/3}}{n^{1/3} L^{2/3}}.$$

Step 4: Obtain final bound (cont'd)

This gives the following δ,

$$\delta = \frac{Lh}{\sqrt{96}} = C_1 \frac{L^{1/3}\sigma^{2/3}}{n^{1/3}}.$$

By the local Fano method, we have obtain the following lower bound on the minimax risk,

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

We require $N \geq 16$, which is satisfied if we have a sufficient number of samples,

$$N \geq 16 \impliedby 2^{m/8} \geq 16 \impliedby \frac{m}{8} = \frac{1}{8h} \geq 4 \impliedby h = C_3 \frac{\sigma^{2/3}}{n^{1/3}L^{2/3}} \leq \frac{1}{32} \impliedby n \geq C_4 \frac{\sigma^2}{L^2}.$$

Ch 2.2: Upper bound

We will define the following estimator $\widehat{f}$.
Let $h > 0$ be a bandwidth parameter whose value we will specify shortly.

$$N(t) = \sum_{i=1}^{n} \mathbb{1}(X_i \in (t-h, t+h)),$$

$$\overline{f}(t) = \begin{cases} \frac{1}{2} & \text{if } N(t) = 0, \\ \frac{1}{N(t)} \sum_{i=1}^{n} Y_i \mathbb{1}\big(X_i \in (t-h, t+h)\big) & \text{if } N(t) > 0. \end{cases}$$

$$\widehat{f}(t) = \operatorname{clip}\left(\overline{f}(t), 0, 1\right).$$

Good events

To bound err_t define $G_t = \{N(t) \geq \alpha_0 nh\}$. to be the "good event" in which a sufficient number of points fall within the $[t-h, t+h]$ interval. We have, by Hoeffding's inequality,

$$
\begin{aligned}
\mathbb{P}(G_t^c) &= \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{1}(X_i \in (t-h, t+h) < \alpha_0 nh\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{n} (\mathbb{1}(X_i \in (t-h, t+h) - P([t-h, t+h])) < -\alpha_0 nh\right) \\
&\qquad \texttt{as } P([t-h, t+h]) = \int_{t-h}^{t+h} p(t)\mathrm{d}t \geq 2\alpha_0 h. \\
&\leq \exp(-2\alpha_0^2 nh^2). \qquad \texttt{Using Hoeffding's} \qquad (1)
\end{aligned}
$$

Bounding the Risk

To analyze this estimator, we will first upper bound the risk as follows,

$$
\begin{aligned}
R(P, \widehat{f}) &= \mathbb{E}\left[\int_0^1 (\widehat{f}(t) - f(t))^2 p(t)\mathrm{d}t\right] \\
&\leq \alpha_1 \, \mathbb{E}\left[\int_0^1 (\widehat{f}(t) - f(t))^2 \mathrm{d}t\right] && \texttt{As } p(t) \leq \alpha_1 \texttt{.} \\
&= \alpha_1 \int_0^1 \underbrace{\mathbb{E}\left[(\widehat{f}(t) - f(t))^2\right]}_{\mathrm{err}_t(\widehat{f})} \mathrm{d}t
\end{aligned}
$$

Let us now write,

$$
\begin{aligned}
\mathrm{err}_t(\widehat{f}) &= \mathbb{E}\left[(\widehat{f}(t) - f(t))^2\right] \\
&= \mathbb{E}\left[(\widehat{f}(t) - f(t))^2 \middle| G_t\right] \underbrace{\mathbb{P}(G_t)}_{\leq 1} + \underbrace{\mathbb{E}\left[(\widehat{f}(t) - f(t))^2 \middle| G_t^c\right]}_{\leq 1 \text{ as } f \text{ is bounded in } [0,1]} \underbrace{\mathbb{P}(G_t^c)}_{\leq e^{-2\alpha_0^2 n h^2}}
\end{aligned}
$$

Bounding the error under the good event

To upper bound $\mathbb{E}[(\widehat{f}(t) - f(t))^2 | G_t]$, let us denote $A_i = \mathbb{1}(X_i \in (t-h, t+h))$ and expand $(\widehat{f}(t) - f(t))^2$ as follows:

$$\begin{aligned}
\left(\widehat{f}(t) - f(t)\right)^2 &\le \left(\overline{f}(t) - f(t)\right)^2 \qquad \texttt{As } \widehat{f}(t) = \mathrm{clip}(\overline{f}(t), 0, 1) \\
&= \left(\frac{1}{N(t)} \sum_{i=1}^{n} (A_i Y_i - f(t))\right)^2 \\
&= \Bigg(\underbrace{\frac{1}{N(t)} \sum_{i=1}^{n} A_i (Y_i - f(X_i))}_{v} + \underbrace{\frac{1}{N(t)} \sum_{i=1}^{n} A_i (f(X_i) - f(t))}_{b}\Bigg)^2 \\
&= v^2 + b^2 + 2bv.
\end{aligned}$$

Bounding the variance

Let us decompose the variance v at t as follows:

$$
\begin{aligned}
\mathbb{E}[v^2|G_t] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N(t)}\sum_{i=1}^{n} A_i\,(Y_i - f(X_i))\right)^2 \middle| G_t, X_1, \ldots, X_n\right]\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\frac{1}{N(t)^2}\sum_{i=1}^{n} (A_i Y_i - f(X_i))^2 \middle| G_t, X_1, \ldots, X_n\right]\right] \\
&= \mathbb{E}\left[\frac{1}{N(t)^2}\sum_{i=1}^{n} A_i \operatorname{Var}(Y_i|X_i) \middle| G_t\right] \qquad \text{Note } \mathbb{E}[Y_i|X_i] = f(X_i) \\
&\leq \mathbb{E}\left[\frac{\sigma^2 N(t)}{N(t)^2} \middle| G_t\right] \leq \frac{\sigma^2}{\alpha_0 nh}. \qquad \text{As } N(t) \geq \alpha_0 nh \text{ under } G_t.
\end{aligned}
$$

Above, conditioning on $X_1, \ldots, X_n$ allows us to claim that the cross-terms are 0 in (a).

Bounding the bias

Next, let us consider the bias b. For any $X_i \in (t-h, t+h)$, we have

$$|f(X_i) - f(t)| \le L|X_i - t| \le Lh.$$

Hence,

$$\begin{aligned} |b| &= \left| \frac{1}{N(t)} \sum_{i=1}^{n} A_i \left(f(X_i) - f(t) \right) \right| \le \frac{1}{N(t)} \sum_{i=1}^{n} A_i \left| f(X_i) - f(t) \right| \\ &\le \frac{1}{N(t)} \sum_{i=1}^{n} A_i Lh = Lh. \end{aligned}$$

Therefore, $\mathbb{E}[b^2 | G_t] \le L^2 h^2$.

Cross-term

Finally, let us consider the cross-term. As $\mathbb{E}[Y_i|X_i] = f(X_i)$, we have,

$$\mathbb{E}[bv|G] = \mathbb{E}\left[b \cdot \mathbb{E}_Y\left[\frac{1}{N(t)}\sum_{i=1}^{n} A_i\left(Y_i - f(X_i)\right)\middle| G_t, X_1, \ldots, X_n\right]\right] = 0.$$

Putting it together

We have shown,

$$R(P,\widehat{f}) \le \alpha_1 \int_0^1 \underbrace{\mathbb{E}\left[(\widehat{f}(t)-f(t))^2\right]}_{\mathrm{err}_t(\widehat{f})} \mathrm{d}t$$

$$\mathrm{err}_t(\widehat{f}) = \mathbb{E}\big[\underbrace{(\widehat{f}(t)-f(t))^2}_{=(b+v)^2}\,\big|\,G_t\big]\underbrace{\mathbb{P}(G_t)}_{\le 1} + \underbrace{\mathbb{E}\left[(\widehat{f}(t)-f(t))^2\Big|\,G_t^c\right]}_{\le 1 \text{ as } f \text{ is bounded in } [0,1]}\underbrace{\mathbb{P}(G_t^c)}_{\le e^{-2\alpha_0 nh^2}}$$

$$|b| \le \frac{1}{N(t)}\sum_{i=1}^{n} A_i L h = Lh,$$

$$\mathbb{E}[v^2|G_t] \le \frac{\sigma^2}{nh},$$

$$\mathbb{E}[bv|G_t] = 0.$$

Putting it together

This gives us,

$$\mathrm{err}_t(\widehat{f}) \le L^2 h^2 + \frac{\sigma^2}{nh} + e^{-2\alpha_0 n h^2}.$$

And hence,

$$R(P, \widehat{f}) \le \alpha_0 \left(L^2 h^2 + \frac{\sigma^2}{nh} + e^{-2\alpha_0 n h^2} \right).$$

Choosing $h = \frac{\sigma^{2/3}}{L^{2/3} n^{1/3}}$, we get

$$R(P, \widehat{f}) \le 2\alpha_1 \frac{\sigma^{4/3} L^{2/3}}{n^{2/3}} + \alpha_1 \exp\left(-2\alpha_0 \frac{\sigma^{4/3} n^{1/3}}{L^{4/3}} \right).$$

This shows us that the minimax rate for this problem is $\Theta(n^{-2/3})$.

The Nadaraya-Watson (Kernel) Estimator

An estimator of the following form,

$$\widehat{f}(t) = \sum_{i=1}^{n} w_i(t) Y_i, \qquad \text{where,} \qquad w_i(t) = \begin{cases} \dfrac{K\left(\frac{t-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{t-X_j}{h}\right)} & \text{if } \sum_{j=1}^{n} K\left(\frac{t-X_j}{h}\right) > 0, \\ 0 & \text{otherwise} \end{cases}$$

Here,

- $K : \mathbb{R} \to \mathbb{R}$ is called a smoothing kernel.
- In our estimator, $K(t) = \mathbb{1}(|t| \leq 1)$. This is sufficient for Lipschtiz smoothness.
- But other kernel choices lead to better rates under stronger guarantees. E.g. In the Hölder class $\mathcal{H}(\beta, L)$ in $\mathbb{R}^d$, where all $(\beta - 1)$ partial derivatives are L-Lipschitz, the minimax rate is $\Theta(n^{\frac{-2\beta}{2\beta+d}})$.

Ch 2.2: Nonparametric density estimation

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

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

Let $\mathcal{P}$ be all continuous distributions whose pdf is in $\mathcal{F}$.

We observe n samples $S = \{X_1, \ldots, X_n\}$ drawn i.i.d from $P \in \mathcal{P}$ and wish to estimate the density p in the L_2 loss,

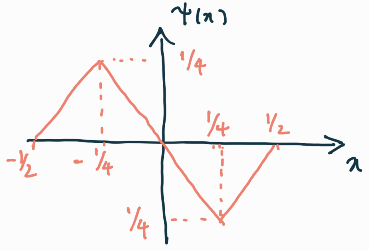
$$\Phi \circ \rho(p_1, p_2) = \|p_1 - p_2\|_2^2 = \int (p_1 - p_2)^2.$$

We will show $R_n^\star \in \Theta(n^{-2/3})$ where,

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

Ch 2.2.1: Lower bounds for density estimation

Step 1: Constructing Alternatives. Let us consider the following function,



$$\psi(x) = \begin{cases} x + 1/2, & \text{if } x \in [-1/2, -1/4], \\ -x, & \text{if } x \in [-1/4, 1/4], \\ x - 1/2, & \text{if } x \in [1/4, 1/2], \\ 0 & \text{otherwise} \end{cases}$$

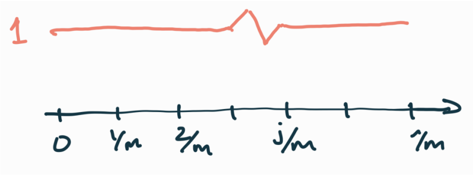
You can check that,

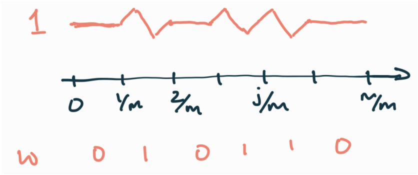
- ψ is 1-Lipschitz.
- $\int \psi = 0$
- $\frac{-1}{4} \leq \psi(t) \leq \frac{1}{4}$.
- $\int \psi^2 = 1/48$.

Step 1: Constructing Alternatives (cont'd)

Let $h > 0$ be a parameter to be chosen later. Let $m = 1/h$.
Let $\omega \in \{0,1\}^m$. Let us define f_ω as follows:

$$\phi_j(x) = Lh\psi\left(\frac{x-(j-1/2)h}{h}\right), \qquad p_\omega(x) = 1 + \sum_{j=1}^{m} \omega_j \phi_j(x).$$





You can verify that

- Each p_ω is L-Lipschitz (similar calculation as before).
- Each p_ω is a valid density ($\int p_\omega = 1$ and $0 \geq p_\omega \geq B$) provided that $h \leq \frac{4}{L}\min(B-1, 1)$.

Alternatives: Let $\mathcal{F}' = \{f_\omega ; \omega \in \Omega_m\}$ where Ω_m is the Gilbert-Varshamov-pruned m-hypercube.

Step 2: Minimum Separation

For any $\omega, \omega' \in \Omega_d$, we have

$$\begin{aligned} \|p_\omega - p_{\omega'}\|_2^2 &= \sum_{j=1}^{m} \mathbb{1}(\omega_j \neq \omega_j') L^2 h^3 \int_{-1/2}^{1/2} \psi^2(u) \mathrm{d}u \\ &= \frac{L^2 h^3}{48} H(\omega, \omega'). \end{aligned}$$

By the Gilbert-Varshamov Lemma, $\min_{\omega \neq \omega'} H(\omega, \omega') \geq m/8 = 1/(8h)$.

Therefore, we have

$$\min_{p_\omega, p'_\omega \in \mathcal{H}'} \|p_\omega - p_{\omega'}\|_2 \geq \sqrt{\frac{L^2 h^3}{48}} \sqrt{\min_{\omega, \omega'} H(\omega, \omega)} = \frac{Lh}{8\sqrt{6}} \triangleq \delta\text{separation} \tag{2}$$

Step 3: Upper bounding the maximum KL

To upper bound the KL divergence,

$$\mathrm{KL}(p_\omega, p_{\omega'}) = \int p_\omega \log\left(\frac{p_\omega}{p_{\omega'}}\right) = \sum_{j=1}^{m} \underbrace{\int_{(j-1)/m}^{j/m} (1+\omega_j\phi_j)\log\left(\frac{1+\omega_j\phi_j}{1+\omega_j'\phi_j}\right)}_{A(\omega_j,\omega_j')} \mathbb{1}(\omega_j \neq \omega_j')$$

Next, you can show (after some algebra), that when $\omega \neq \omega_j'$,

$$A(\omega_j, \omega_j') \leq L^2 h^3 \int_{-1/2}^{1/2} \psi^2 = \frac{L^2 h^3}{48}.$$

Hint. Use inequalities $\log(1+x) \geq x - x^2$ for $x \geq -0.68$ and $\log(1+x) \leq x$.

Therefore, $\mathrm{KL}(p_\omega, p_{\omega'}) \leq \frac{L^2h^3}{48} H(\omega, \omega')$. As, $\max_{\omega, \neq \omega'} H(\omega,\omega') \leq m = 1/h$, we have

$$\max_{p_\omega, p_{\omega'} \in \mathcal{H}'} \mathrm{KL}(p_\omega, p_{\omega'}) \leq \frac{L^2 h^4}{48}.$$

Step 4: Final Bound

We require the maximum KL to be smaller than $\frac{\log(|\mathcal{F}'|)}{4n}$. But we know $|\mathcal{F}'| \geq 2^{m/8}$. So it is sufficient if

$$\frac{L^2 h^4}{48} < \frac{\log(2^{m/8})}{4n} = \frac{\log(2)}{32nh}.$$

So choose,

$$h = \left(\frac{3\log(2)}{2} \frac{1}{nL^2} \right)^{1/3}.$$

By the local Fano method, we obtain the following lower bound on the minimax risk,

$$R_n^\star \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) = C_2 L^2 h^2 = C_3 \frac{L^{2/3}}{n^{2/3}}.$$

Requirements

Requirements:

- $h \leq \frac{4}{L}\min(B-1,1)$, for it to be a valid density.
- $\frac{1}{h} = m \geq 8$ for Gilbert-Varshamov.
- $2^{m/8} \geq 16$, which is satisfied if $h \leq 1/32$.
- $h \leq 2.72/L$, an inequality we used.

These conditions are satisfied if n is sufficiently large.

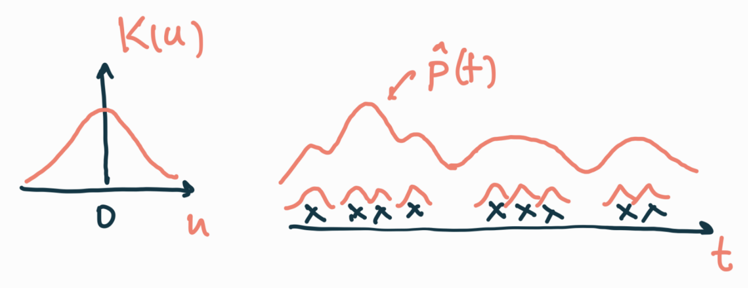
Ch 2.2.2: Upper bound via Kernel Density Estimation

A kernel density estimator $\widehat{p}$ with kernel K and bandwidth parameter h (to be chosen shortly) has the following form,

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

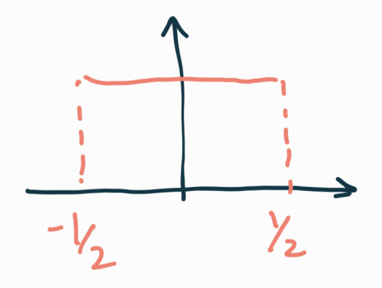
Here, $K : \mathbb{R}_+ \to \mathbb{R}_+$ is a smoothing kernel with the following properties,

$$\int K(u)\mathrm{d}u = 1, \qquad K(u) = K(-u).$$



Boxcar Kernel

For our problem, we will use $K(u) = \mathbb{1}(|u| \leq 1/2)$.



- This kernel is sufficient for Lipschitz functions.
- In HW2, you will analyse KDE for densities whose first derivative is Lipschitz. Then, we need to use a different kernel.

Bias–Variance Decomposition

Let us decompose the risk as follows,

$$
\begin{aligned}
&\mathbb{E}\left[\|p-\widehat{p}\|_2^2\right] \\
&\quad = \mathbb{E}\left[\int (p-\widehat{p})^2\right] \\
&\quad = \mathbb{E}\left[\int (p-\mathbb{E}\widehat{p})^2 + \int (\mathbb{E}\widehat{p}-\widehat{p})^2 + 2\int (p-\mathbb{E}\widehat{p})(\mathbb{E}\widehat{p}-\widehat{p})\right] \\
&\quad = \int \underbrace{(p(t)-\mathbb{E}\widehat{p}(t))^2}_{\mathrm{bias}^2(t)}\mathrm{d}t + \int \underbrace{\mathbb{E}\left[(\widehat{p}(t)-\mathbb{E}\widehat{p}(t))^2\right]}_{\mathrm{var}(t)}\mathrm{d}t + \\
&\qquad\qquad 2\int (p(t)-\widehat{p}(t))\underbrace{\mathbb{E}\left[\widehat{p}(t)-\mathbb{E}\widehat{p}(t)\right]}_{=0}\mathrm{d}t
\end{aligned}
$$

Bounding the Bias

Let us first bound the bias,

$$
\begin{aligned}
\mathrm{bias}(t) = \mathbb{E}[\widehat{p}(t)] - p(t) &= \mathbb{E}_{X\sim P}\left[\frac{1}{h}K\left(\frac{X-t}{h}\right)\right] - p(t) \\
&= \int \frac{1}{h}K\left(\frac{x-t}{h}\right)p(x)\mathrm{d}x \; - p(t)\int K(u)\mathrm{d}u \\
&= \int K(u)(p(t+uh)-p(t))\mathrm{d}u.
\end{aligned}
$$

Therefore, noting that $\int |u|K(u)\mathrm{d}u = 1/4$, we have

$$
|\mathrm{bias}(t)| \leq \int K(u)|p(t+uh)-p(t)|\mathrm{d}u \leq \int K(u)Lh|u|\mathrm{d}u = \frac{1}{4}Lh.
$$

Bounding the Variance

Let us next bound the variance.

$$\begin{aligned}
\mathrm{var}(t) = \mathbb{V}\mathrm{ar}\left(\widehat{p}(t)\right) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{t-X_i}{h}\right)\right) \\
&= \frac{1}{n}\mathbb{V}\mathrm{ar}_{X\sim P}\left(\frac{1}{h}K\left(\frac{t-X}{h}\right)\right) && \texttt{as the terms are i.i.d} \\
&\leq \frac{1}{n}\mathbb{E}\left[\frac{1}{h^2}K^2\left(\frac{X-t}{h}\right)\right] \\
&= \frac{1}{nh^2}\int K^2\left(\frac{x-t}{h}\right)p(x)\mathrm{d}x \\
&= \frac{1}{nh}\int K^2(u)p(x+uh)\mathrm{d}u \\
&\leq \frac{B}{nh}\int K^2(u)\mathrm{d}u = \frac{B}{nh}. && \texttt{as } \int K^2 = 1
\end{aligned}$$

Putting It All Together

Putting these bounds together we get,

$$\begin{aligned}\mathbb{E}\left[\|p-\widehat{p}\|_2^2\right] &= \int \mathrm{bias}^2(t) + \int \mathrm{var}(t) \\ &\leq \int_0^1 \frac{1}{4}L^2h^2 + \int_0^1 \frac{B}{nh} \\ &\leq \frac{1}{4}L^2h^2 + \frac{B}{nh}\end{aligned}$$

Choosing $h = n^{-1/3}L^{-1/3}$, we get

$$\mathbb{E}\left[\|p-\widehat{p}\|_2^2\right] \in \mathcal{O}\left(\frac{L^{2/3}}{n^{2/3}}\right).$$

Boundary correction

- The bounds for $\mathrm{bias}(t)$ and $\mathrm{var}(t)$ are valid only within the interval $[h, 1-h]$.
- In the analysis above, we may ignore this issue (*i.e.*, assume the worst-case error) and still obtain the correct convergence rates.
- For higher-order smoothness assumptions, however, this boundary error may no longer be negligible.
- A common remedy is to mirror the data at the boundaries, which is effective when the underlying density has zero derivative at the boundary.