Problem 1

Consider a nonparametric regression problem where we observe the data $\{(x_{i,n},\, y_{i,n})\}_{i=1}^{n}$. Suppose $x_{i,n} = i/n$ (so, $x_{i,n}$ is *deterministic*), and $y_{i,n}$ follows the model

$$y_{i,n} = f\left(x_{i,n}\right) + \epsilon_{i,n},$$

with unknown f, and $\epsilon_{i,n} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
We start by writing the likelihood function of our observations.

(a) Write the likelihood function $L(\{y_{i,n}\}; f, \sigma)$.

Following the 'parametric statistics approach', you start by trying to find a continuous function $\hat{f}$ that maximizes the likelihood.

(b) Does such an approach make sense? If not, why? What is the undesirable property of the resulting estimator?

You then decide to restrict the space of possible estimates $\hat{f}$ to the set of constant functions, i.e. $\hat{f}(x) = c$. However, you know that this is excessively restrictive, and to obtain a nonparametric model you decide to 'localize your estimator', i.e. for a fixed point x_0, you estimate $f(x_0)$ by looking at the likelihoods at the points $\{(x_{i,n}, y_{i,n}) : |x_{i,n} - x_0| \le h\}$.

(c) Write the explicit formula of the maximum likelihood estimator $\hat{f}(x)$ you obtain in this case. Is this a known estimator?

Problem 2

Consider now the density estimation model $x_1, \ldots, x_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ with unknown μ and σ^2. For

$$\hat{\mu} = \frac{1}{n}\sum_{i \le n} x_i$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i \le n} \left(x_i - \hat{\mu}\right)^2$$

(a) Show that for any fixed x_0

$$\left[\phi_{\hat{\mu},\hat{\sigma}^2}\left(x_0\right) - \phi_{\mu,\sigma^2}\left(x_0\right)\right]^2 = O_p\left(\frac{1}{n}\right)$$

where $\phi_{\hat{\mu},\hat{\sigma}^2}\left(x_0\right)$ is our estimated gaussian density (gaussian density with plug-in estimators), and $\phi_{\mu,\sigma^2}\left(x_0\right)$ is the truth. Hint. use the delta method. Also, to simplify things recall that for gaussians, $\hat{\mu}$ and $\hat{\sigma}^2$ are independent.

Problem 3

Suppose $x_i \overset{iid}{\sim} U[-1, 1]$, and

$$y_i = f\left(x_i\right) + \epsilon_i, \qquad i = 1, \ldots, n, \qquad n = 100$$

where $\epsilon_i \sim N(0, 1)$. Consider the following three unknown functions

(a) $f(x) = sin(x * \pi)$.

Problem 3 continued on next page...

(b) $f(x) = 2x + x^3 - 6x^4$.

(c) $f(x) = \frac{1}{1+(5x)^2}$.

The aim of this problem is estimating the optimal bandwidth h of nonparametric estimates of f, using 2-fold CV (i.e. the validation set approach), 5-fold CV and 10-fold CV.

(a) Compare the estimates of the optimal bandwidth h, for (i) Nadaraya-Watson estimation with a "box" kernel, and (ii) Nadaraya-Watson with a "gaussian" kernel using 2-, 5- and 10-fold CV by defining the folds based on the ordering of the observations; this means that in the case of 2-fold CV, or the validation set approach, you use observations $1, \ldots, \lfloor n/2 \rfloor$ for training and the rest for validation – use the same strategy for other folds. Plot the estimated functions for the optimal CV choice of the bandwidth and compare with the unknown function f.

(b) Repeat the above experiment by generating 100 *random* folds for each of 2-, 5- and 10-fold CV estimates. Compare the estimated bandwidths, and their variances, with the 'oracle' bandwidth. The 'oracle' bandwidth is the h that minimizes the misfit from f, computed by assuming f known. Make appropriate plot(s) to compare these estimators. Give a short writeup stating comparisons/conclusions.