CS4070 – PART 2
EXERCISES RELATED TO LECTURE 2

(1) Proof that for a vector $Z \in \mathbb{R}^n$ we have

$$\mathrm{E}\left[\|Z\|^2\right] = \mathrm{tr}\left(\mathrm{Cov}\, Z\right) + \mathrm{E}\left[Z\right]^T \mathrm{E}\left[Z\right]$$

Hint: use $\mathrm{E}\left[Z_i^2\right] = \mathrm{Var}(Z_i) + (\mathrm{E}\left[Z_i\right])^2$.

(2) Consider the model

$$Y_i = r(x_i) + \sigma \epsilon_i$$

with $\{\epsilon_i\}_i$ a sequence of independent $N(0,1)$-distributed random variables and r defined by

$$r(x) = \sum_{j=1}^{M} \theta_j \varphi_j(x)$$

The φ_j functions are *basis functions*. To include an intercept it is customary to take $\varphi_1(x) = 1$. Examples include

- $\varphi_j(x) = x^j$ (polynomials);
- spline functions;
- Gaussian basis functions $\varphi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$;
- sigmoidal basis functions $\varphi_j(x) = \psi\left(\frac{x-\mu_j}{s}\right)$ with $\psi(x) = (1+e^{-x})^{-1}$;
- Fourier basis;
- Wavelet basis.

These are just a few examples, I don't expect you to know for all of these how exactly these are defined. In this example you will experiment a bit with sigmoidal basis functions, to get a feeling for overfitting.

(a) Write the model in matrix vector notation $y = X\theta + \epsilon$, what is the design matrix X?

(b) Generate data as follows: sample x-values from the uniform distribution on $[-2, 2]$. Take $r(x) = \sin(\exp(x))$ and draw $y_i \mid x_i \sim N(r(x_i), (0.4)^2)$. Take sample size $n = 50$. So this gives the data $\mathcal{D} = \{(x_i, y_i),\, 1 \le i \le n\}$.

(c) Take μ_j to be L equidistant points on $[-2, 2]$ and set $s = 0.1$. This specifies the sigmoidal basisfunctions. Write a function that takes as input the vector of x-values and returns the design-matrix (don't forget the "intercept basis function").

(d) Compute the maximum likelihood estimator and make a plot of $\mathcal{D}$ together with the fitted curve. Take $L = 10$.

(e) Repeat for higher values of L. At some point you should notice bad behaviour!

Date: November 16, 2023.

(f) One way to deal with this problem is penalisation. Read Section 6 of Chapter 1 in the book by RG. The idea is that if we allow L to be large, not too many coefficients in the estimator for θ can be very large. The loglikelihood is given by (ignoring that we don't know σ for a moment)

$$\ell(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y - X\theta\|^2.$$

Now we define a new "regularised" estimator by maximising this expression under the restriction that $\|\theta\|^2 \leq C$. Using the method of Lagrange multipliers, it can be shown that this is equivalent to minimising

$$\theta \mapsto \|y - X\theta\|^2 + \lambda\|\theta\|^2,$$

where $\lambda > 0$ is related to C. Verify that the resulting estimator is given by $(X^T X + \lambda I)^{-1} X^T y$.

(g) Explain that the inverse in this expression always exists, even if X has more columns than rows. *Hint: show that the smallest eigenvalue of $X^T X + \lambda I$ is strictly positive.*

(h) Implement this estimator as well, take L large, and empirically find a value of λ that does a reasonable (visually) bias-variance trade-off.

A good value for λ can be obtained by cross-validation for example. An alternative is a Bayesian approach where λ gets assigned a prior distribution. It is good to experiment a bit and see if you understand what happens if you take λ either very close to zero or very large.