Regularisation

1. Suppose that, in a linear regression task, we have (based on prior knowledge) a diagonal Gaussian prior on the parameter vector w: $w \sim \mathcal{N}(0, \frac{1}{\lambda}I)$ (where I is the (d-dimensional) identity matrix) for some hyperparameter λ.

   The target distribution is a Gaussian with deviance σ and mean $f(x)$ (the target function); as in the slides.

   Derive the error measure (loss function) using Maximum A Posteriori estimation! (Using similar assumptions to when doing just Maximum Likelihood estimation.)

2. Repeat Exercises 1a-c from Lab 4 (last time), but use L^2 regularisation with $\lambda = 0.01$. How did the plots and the errors change for the polynomials of different degrees?

3. (Problem 4.8 in the book) In the augmented error minimization with $\lambda > 0$, assume that E_{in} is differentiable and use gradient descent to minimize E_{aug}:

   $$w(t+1) \leftarrow w(t) - \eta \nabla E_{aug}(w(t)).$$

   Show that the update rule above is the same as

   $$w(t+1) \leftarrow (1 - 2\eta\lambda)w(t) - \eta \nabla E_{in}(w(t)).$$

   Note: This is the origin of the name “weight decay” : $w(t)$ decays before being updated by the gradient of E_{in}.