Problem 1.

Consider the training objective $J = ||Xw - t||^2$ subject to $||w||^2 \leq C$ for some constant C.

How would the hypothesis class capacity, overfitting/underfittting, and bias/variance vary according to C?

	Larger C	Smaller C
Model capacity (large/small?)	_____	_____
Overfitting/Underfitting?	__fitting	__fitting
Bias variance (how/low?)	__ bias / __ variance	__ bias / __ variance

Note: No proof is needed

Problem 2.

Consider a one-dimensional linear regression model $t^{(m)} \sim N(wx^{(m)}, \sigma_\epsilon^2)$ with a Gaussian prior $w \sim N(0, \sigma_w^2)$. Show that the posterior of w is also a Gaussian distribution, i.e., $w|x^{(1)}, t^{(1)}, \cdots, x^{(M)}, t^{(M)} \sim N(\mu_{post}, \sigma_{post}^2)$. Give the formulas for μ_{post}, σ_{post}^2.

Hint: Work with $P(w|D) \propto P(w)P(D|w)$. Do not handle the normalizing term.

Note: If a prior has the same formula (but typically with different parameters) as the posterior, it is known as a *conjugate prior*. The above conjugacy also applies to multi-dimensional Gaussian, but the formulas for the mean vector and the covariance matrix will be more complicated.

Problem 3.

Give the prior distribution of w for linear regression, such that the max a posteriori estimation is equivalent to l_1-penalized mean square loss.

Note: Such a prior is known as the *Laplace distribution*. Also, getting the normalization factor in the distribution is not required.

END OF W5