Problem 1: **(40 Points)** Repeat the results presented in Lecture 8 "Multistage Neural Networks" Figure 1 with the same target function as in Eq. (2). Note that you don't have to get identical training loss, if you choose to use different optimizers, hyper-parameters, and neural network architecture. The main purpose for repeating this example problem is to practice and demonstrate how the multi-stage (e.g., two-stage) neural networks compare with the single-stage neural network in terms of significantly reducing training loss (or improving training accuracy).

Problem 2: Gaussian process (GP) regression and kernel ridge regression for $y = f(x), \quad x, y \in \mathbb{R}$. (Python package scikit-learn is recommended for this problem, but other packages or languages is also acceptable.)

(1) **(10 Points)** Generate data pairs $\{(x_i, y_i)\}_{i=1}^{1000}$ evenly in the range of $x \in [0, 50]$ from the true function $y = \cos(x)$. Add i.i.d random noises (zero mean and variance of 0.16) to 40 points randomly selected from the first half of all data pairs and only use these 40 points as the training data for this problem. Plot the noisy training data points together with the true function $y = \cos(x)$ in the range of $x \in [0, 50]$.

(2) **(10 Points)** Use zero mean prior and the kernel of the form (the first term is known as periodic kernel, and the second term is known as white kernel):

$$k(x_i, x_j) = \alpha^2 \exp\left(-\frac{2\sin^2(\pi d(x_i, x_j)/p)}{\ell^2}\right) + \sigma^2 \delta_{ij},$$

where $d(x_i, x_j)$ is the Euclidean distance and δ_{ij} denotes the Kronecker delta, to fit a Gaussian process and plot the fitted GP results with the training data and the true function. Report the hyperparameters α, p, ℓ, and σ of the fitted GP (hint: In the scikit-learn package, these hyperparameters are optimized automatically when fitting the GP, and the kernel can be accessed by gaussian_process.kernel_).

(3) **(10 Points)** Try to multiply the optimal σ with a factor of 2 and discuss the effect on GP results.

(4) **(10 Points)** Multiply the optimal ℓ with a factor of 2 and discuss the effect on GP results.

(5) **(10 Points)** Use the periodic kernel (with the same hyperparameters as being optimized in GP model) to fit a kernel ridge regression model. Explain how to choose the penalty coefficient for kernel ridge regression such that the fitted model is similar to the posterior mean of GP model. Report the value of penalty coefficient and plot the fitted model together with the GP regression results.

(6) **(10 Points)** Modify the value of penalty coefficient to 10 and discuss the effect on kernel ridge regression results.