# Homework week 11

**Problem Collection**

1 **Question 1-1** The issue of convergence in the stochastic gradient descent method.

**Answer 1-1** *Garrigos, G., Gower, R. M. (2023). Handbook of convergence theorems for (stochastic) gradient methods. arXiv preprint arXiv:2301.11235.*

**Question 1-2** For practical problems, how should one determine the number of hidden layers and the number of neurons in each layer?

**Answer 1-2** A simple answer to this question is that, case by case, the number of hidden layers and neurons is determined by the problem under consideration. For example, if the activation function is given as $\tanh x$ or sigmoid, then both odd and even polynomials can be learned with only one hidden layer, but they require different numbers of neurons.

2 **Question 2-1** How can we find an approximate function that approximates an $L^p$ function using a neural network?

**Answer 2-1** Many related papers can be found on Google Scholar. However, I should emphasize that regardless of whether they use ReLU, sigmoid, or $\tanh x$ as the activation function, these works typically focus on approximating $L^\infty$ functions with some regularity (e.g., Lipschitz continuous functions). What I am curious about is whether there are "analogous" results for the case of arbitrary $2 < p < \infty$, where one seeks approximation in the $L^p$ sense. So far, it seems that no relevant literature has addressed this situation.

3 **Question 3-1** Note that in the paper *Ryck et al., On the approximation of functions by tanh neural networks*, the authors work with function in $W^{k,\infty}$. In theoretical textbook, we all know that using Morrey's estimate, we obtain the embedding $W^{k,\infty}(\Omega) \subset C^{k-1,\gamma}(\Omega)$ for $\Omega \subset \mathbb{R}^n$ and $\gamma \in (0,1)$. That is, every $W^{k,\infty}$ function is $C^{k-1}$ function. Then, by Stone-Weierstrass Theorem, we always can find a neural network approximate function. However, I would like to ask if $W^{k,p}(\Omega)$ with $1 \leq p < n$ (i.e., we don't have Morrey's estimate), how can we construct a neural network approximate function to a $W^{k,p}(\Omega)$ function?

**Answer 3-1** Similar question to **Question 2-1**. Many related papers can be found on Google Scholar, such as $W^{k,\infty}$. I would like to ask if $W^{k,p}(\Omega)$ with $1 \leq p < n$, how can we construct a neural network approximate function to a $W^{k,p}(\Omega)$ function? So far, it seems that no relevant literature has addressed this situation.

4 **Question 5-1** In class, when we formulate the problem of learning $\mathbb{P}(y|x)$, we typically assume that $\mathbb{P}(x|y=0)$ and $\mathbb{P}(x|y=1)$ are Gaussian, while $\mathbb{P}(y=0)$ and $\mathbb{P}(y=1)$ are Bernouli. The teacher also noted that if $\mathbb{P}(x|y=0)$ and $\mathbb{P}(x|y=1)$ are allowed to be a more general distributions, the problem becomes substantially more difficult. My question is whenever this difficulty is primarily computational (numerical implementation), or does it stem from a lack of applicable theoretical tools?

**Answer 5-1** Allowing $\mathbb{P}(x \mid y)$ to be a general distribution makes the classification problem substantially harder mainly for statistical/theoretical reasons. Instead of estimating a small number of Gaussian parameters, one must learn the full class-conditional densities (or equivalently the posterior) in a nonparametric way. In moderate or high dimensions this density-estimation step is ill-posed and suffers from the curse of dimensionality, so good performance typically requires much more data and/or strong structural regularity assumptions to obtain meaningful convergence rates. Because these general models rarely yield closed-form posteriors, practical algorithms then rely on iterative procedures such as EM, numerical integration, MCMC, or variational inference; thus the computational difficulty is largely a consequence of the underlying theoretical bottlenecks.

Relevant literature:

*https://www.jstor.org/stable/24310789*,

*https://arxiv.org/abs/1503.03305*.

5 **Question 8-1** Given a SDEs (Ornstein-Uhlenbeck process)

$$\mathrm{d}Y_t = -kY_t\mathrm{d}t + \sigma\mathrm{d}B_t, \quad Y_0 = y,\ k \in \mathbb{R},\ \sigma > 0. \tag{0.1}$$

How to solve this SDE?

**Answer 8-1** Use Ito formula, we can obtain the solution.

**Problem 2: Toy model/Solvable Model Problem for final project** Motivated by the idea behind PINNs, when we want to approximate function $W^{1,p}$ or $W^{k,p}$ in the regimes $p < d$ or $kp < d$, respectively, the first step is to identify the inequalities satisfied by the relevant exponents, such as Sobolev inequalities or Poincaré inequalities. This is analogous to the PINN setting, where one is given a forcing term and boundary data. Once such additional information is available, we then have a chance to construct reasonable approximations of functions in these spaces.