Homework 1

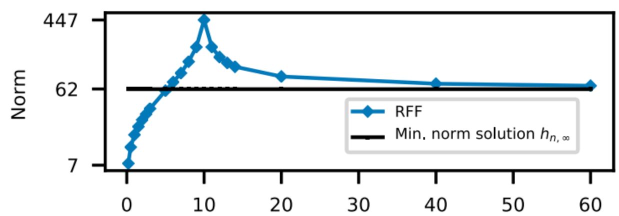
Using the assigned reading listed on the course page, answer the questions below with a short response. Note that we are looking for concise statements that show understanding, not quantity. The total discussion should be less than a page.

Rethinking generalization

1. Double descent refers to a phenomenon in which the generalization error first decreases and then increases and then decreases again.

   (a) How does this differ from the bias-variance trade-off in classical learning theory?

   (b) Belkin et al. use the following figure as a candidate explanation of double descent. How does this figure relate to double descent, and why might we expect the parameter norm to behave this way as a function of the model complexity?



   (**Note**: this is an active area of research, so be aware that this might not be the full explanation, but do your best to articulate Belkin et al.'s take on it, and feel free to provide your own ideas too.)

2. Assume you want to train a neural network in practice. You are given a fixed training dataset of size $|\mathcal{X}| = N$ and want to figure out how big your linear neural network should be. What does the double descent phenomena tell you about the choice of your model size if you want to avoid the critical regime? If you don't know the data size ahead of time, what would be a good strategy for choosing the number of parameters in your model?

3. In the over-parameterized regime, there exist many solutions that can perfectly overfit the data. The solution is measured using excess risk, which tells us how good a solution is to the minimum-norm estimator. Barlett et al. characterize the lower bound for the expected excess risk as:

$$\mathbb{E}\left[R(\hat{\theta})\right] \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right) \qquad \text{(Theorem 4)}$$

   Provide a high-level explanation of what this implies. (**Hint**: Page 2 of Barlett et al.)

Submission: Upload a PDF of your response through Canvas by **9/21 at 1pm**.