Problem Set 3

Name:

Due Midnight on January 25, 2022

The goal of this problem set is to understand how to compute and use the NNGP. We use (T) to denote theory exercises and (C) to denote coding exercises. We use the flag **Required** to denote problems that are required for all students. We lastly indicate more involved problems with a *****, and these will be equivalent to solving 2 non-$*$ problems.

Students should do all the required problems and at least 3 of the remaining problems (for a total of 4 problems). Note that $*$ problems are worth 2 normal problems, i.e. solving one $*$ problem means submitting a total 3 problems.

Neural Network Gaussian Processes, Dual Activations, and Over-parameterization

The following problems depend on material through Lecture 4.

Problem 1 (T). Let $\xi \in [-1, 1]$. Recall that for a given activation $\phi(x) : \mathbb{R} \to \mathbb{R}$, the dual activation $\check{\phi} : [-1, 1] \to \mathbb{R}$ is defined by:

$$\check{\phi}(\xi) = \mathbb{E}_{(u,v)\sim\mathcal{N}(\mathbf{0},\Lambda)}[\phi(u)\phi(v)] \quad ; \quad \Lambda = \begin{bmatrix} 1 & \xi \\ \xi & 1 \end{bmatrix}$$

Prove that the dual activation of $\phi(x) = e^{-C^2+Cx}$ is $\check{\phi}(\xi) = e^{-C^2+C^2\xi}$.
Hint: Here are two potential approaches:

(1) Directly compute the double integral:

$$\mathbb{E}_{(u,v)\sim\mathcal{N}(\mathbf{0},\Lambda)}[\phi(u)\phi(v)] = \frac{1}{2\pi\sqrt{1-\xi^2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-C^2+Cu} e^{-C^2+Cv} e^{-\frac{u^2+v^2-2uv\xi}{2(1-\xi^2)}} dudv$$

This integral can be evaluated by using the technique from Lecture 1 and Homework 1. Namely, complete the square in the integrand (for u first and then v) and use the fact that the density of the Gaussian has integral 1.

(2) Use the facts that the dual commutes with differentiation, i.e. $(\check{\phi}') = \check{\phi}'$, that $\check{\phi}(1) = 1$, and that the dual of $a\phi(x)$ is $a^2\check{\phi}$ for constant $a \in \mathbb{R}$. Solving the corresponding differential equation in $\check{\phi}$ gives the result.

Problem 2 (T, *). Prove that the dual activation of $\phi(x) = \sin(x)$ is $\check{\phi}(\xi) = e^{-1}\sinh(\xi)$.
Hint: Again, we provide two possible approaches.

(1) Use the fact that $\sin(x) = \frac{1}{2i}(e^{ix} - e^{-ix})$ and use the derivations similar to that of Problem 1. In particular, it may be useful to compute:

$$\mathbb{E}_{(u,v)\sim\mathcal{N}(\mathbf{0},\Lambda)}[\phi(u)\phi(v)] = \frac{1}{2\pi\sqrt{1-\xi^2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{C_1 u} e^{C_2 v} e^{-\frac{u^2+v^2-2uv\xi}{2(1-\xi^2)}} dudv \quad ;$$

for constants C_1, C_2.

(2) Use the fact that $\sin(x) = \frac{1}{2i}(e^{ix} - e^{-ix})$ and then use the closed form for the moment generating function of a Gaussian from Lecture 1.

Problems 3 and 4 depend on the results (e.g. Theorem 1) from Section 3 of Lecture 4.

Problem 3 (T). Prove the following facts about the dual activation, $\check{\phi} : [-1, 1] \to \mathbb{R}$:

1. $(\check{a\phi}) = a^2\check{\phi}$ for any constant $a \in \mathbb{R}$.
2. $\check{\phi}(\xi)$ is convex and non-decreasing for $\xi \in [0, 1]$.

Problem 4 (T). Let $\phi(x) = \sin(x)$ denote the ReLU activation. For $x, \tilde{x} \in \mathcal{S}^{d-1}$ (e.g. $x, \tilde{x}$ on the unit sphere) with $\xi = x^T\tilde{x}$, the dual activation is given by:

$$\check{\phi}(\xi) = e^{-1}\sinh(\xi)$$

Use this to compute the dual activation of $\psi(x) = \cos(x)$.

Problem 5 (T. Required) Complete the proof of Proposition 1 in Lecture 4. Namely, let $w \sim \mathcal{N}(\mathbf{0}, I_d)$, $x, \tilde{x} \in \mathbb{R}^d$. Prove that if $u = w^Tx$, $v = w^T\tilde{x}$, then:

$$\mathbb{E}[(u, v)] = (0, 0) \quad ; \quad \text{Cov}(u, v) = x^T\tilde{x}$$

For Problems 6 and 7, utilize the following setting:

For $d = 200, n = 100$, generate data $X \in \mathbb{R}^{n\times d}$ with entries drawn i.i.d. from a standard normal distribution. Normalize X such that each row of X has norm 1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the function $f(x) = \frac{1}{d}\sum_{i=1}^{d}\sin x_i$, and let $y = f(X) \in \mathbb{R}^n$ where $f(X)_j = f(X_{j,:}^T)$ (i.e. f is applied to each row of X). Lastly, for $n_t = 5000$, generate test samples $X_t \in \mathbb{R}^{n_t\times d}$ where the entries are drawn i.i.d. from a standard normal distribution and test labels $y_t = f(X_t) \in \mathbb{R}^{n_t}$. Normalize X_t such that each row of X_t has norm 1. **Make sure to set the random seed in numpy so that experiments are reproducible.**

Problem 6 (C). Report the test MSE of using kernel regression to fit the data exactly (i.e. via the `solve` function) with the following NNGP kernels of 1 hidden layer networks:

1. The normalized NNGP of a network with ReLU activation: $\check{\phi}(\xi) = \frac{1}{\pi}\left(\xi(\pi - \arccos(\xi)) + \sqrt{1 - \xi^2}\right)$.
2. The normalized NNGP of a network with sine activation: $\check{\phi}(\xi) = \frac{2e}{e^2-1}\sinh(\xi)$.
3. The normalized NNGP of a network with erf activation: $\check{\phi}(\xi) = \frac{1}{\arcsin\left(\frac{2}{3}\right)}\arcsin\left(\frac{2\xi}{3}\right)$.

Compare the above MSEs with the test MSE of solving kernel regression with the Laplace kernel $e^{-L\|x-\tilde{x}\|_2}$ with $L = \frac{1}{200}$.

Problem 7 (C, *). Let $f(x) = \frac{\sqrt{2}}{\sqrt{k}}A\phi(Bx)$ denote a 1 hidden layer neural network with $A \in \mathbb{R}^{1\times k}$, $B \in \mathbb{R}^{k\times d}$, and $\phi : \mathbb{R} \to \mathbb{R}$ be the element-wise ReLU activation (i.e. $\phi(z) = \max(0, z)$). In this problem, we consider the case where the parameters $B_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and are fixed. For $d = 200$ and $k \in [1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024]$, plot k vs. the test MSE when training only the last layer A to fit the data (X, y). On the same plot, present the test MSE of solving kernel regression with the normalized NNGP of the ReLU network given by

$$\check{\phi}(\xi) = \frac{1}{\pi}\left(\xi(\pi - \arccos(\xi)) + \sqrt{1 - \xi^2}\right).$$

How does the test error of the NNGP (corresponding to $k \to \infty$) compare with that of the corresponding finite width models?