Homework 2

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.

Adversarial examples

1. Give a brief summary of the main arguments in Ilyas et al. 2019. What are adversarial examples? In what sense could they be "features" rather than bugs? What is the logic behind the experimental setup? How do the experiments disentangle robust and non-robust features?

2. Ilyas et al. argue that their "findings prompt us to view adversarial examples as a fundamentally human phenomenon." What do they mean by this? Do you agree? What might be an explanation for why the "human-selected notion of similarity" is a reasonable one, or at least is the one we evolved to have? Can you come up with ideas for how to define "adversarial robustness" more from first principles, rather than having to make reference to humans at all?

Approximation theory

1. Suppose you are designing a deep neural network and you want to minimize approximation error as much as possible. Answer the following according to the findings in Barron 1993 and the additional results presented in Professor Moitra's lecture. Provide a sentence or two discussions of each:

   (a) How wide should you make your network?

   (b) How deep should you make your network?

   (c) Using your knowledge from last week, do these answers change if you care more about generalization error than approximation error?

   (d) What might be some reasons to restrict the width or depth of your net (feel free to speculate beyond the readings here)?

2. **(optional)** Consider the parity function $\chi : \{0,1\}^n \to \{0,1\}$.

   (a) Construct a deep network with ReLU activations that exactly expresses the parity function with $O(n)$ hidden units.

   (b) Consider a ReLU network with m hidden units. Show that the function computed by the network can be described by a partition of the space of its inputs into at most $f(m)$ parts, where the function is linear on each one. What is your $f(m)$?

3. **(optional)** We say that a function $f : \mathbb{R} \to \mathbb{R}$ is positive semidefinite if for all sequences $x_1, x_2, \cdots, x_k$ we have that $\sum_{i,j} x_i x_j f(x_i - x_j) \geq 0$. Bochner's Theorem gives us an

Homework 2

equivalent characterization that a function is positive semidefinite iff it can be written as

$$f(x) = \int_{\omega} e^{2\pi i \omega^T x} F(\omega) d\omega$$

for some nonnegative function F. Show that

$$C_f \leq O\Big(- f(0)\nabla^2 f(0)\Big)$$

where C_f is the constant in Barron's Theorem. Note that positive semidefinite functions have $\nabla^2 f(0) \leq 0$. *Hint: Use Cauchy-Schwarz*

Submission: Upload a PDF of your response through Canvas by **9/28 at 1pm**.