CSE 598 TEL - Homework 1

MJT

October 2, 2016

Instructions.

- All rules on the webpage apply.
- You may work in groups of size at most two; put all the NetIDs clearly on the first page, and submit through gradescope.
- Homework is due **Wednesday, October 5, at 11:00am**; no late homework accepted.
- Please consider using the provided LaTeX file as a template (apologies for the weird indentation), or at least something vaguely visually similar, since gradescope tries to automatically locate beginnings and ends of problems.

1. **(Analysis I:** missing step from lecture on proof by Hornik et al. (1989)**.)**

   Provide a proof from the missing step in lecture 4, restated as follows.

   > Let $\sigma : \mathbb{R} \to [0,1]$ be given, and suppose it is *sigmoidal*, meaning continuous, monotone nondecreasing, and satisfying
   >
   > $$\lim_{z \to -\infty} \sigma(z) = 0, \qquad \lim_{z \to +\infty} \sigma(z) = 1.$$
   >
   > Given any any $g \in \mathcal{H}_{\cos} = \left\{ x \mapsto \cos(a^\top x + b) : a \in \mathbb{R}^d, b \in \mathbb{R} \right\}$ and any $\epsilon > 0$, there exists $f \in \operatorname{span}(\mathcal{H}_\sigma)$ with
   >
   > $$\|f - g\|_{\mathrm{u}} = \sup \left\{ |f(x) - g(x)| : x \in [0,1]^d \right\} \leq \epsilon.$$

   Easy mode / hint: feel free to instead prove this approximation result for the norm $\|f - g\|_1 = \int_{[0,1]^d} |f(x) - g(x)| dx$ (which is less finicky and carries more intuition from lecture), and moreover with $d = 1$.

   Solution.

2. **(Analysis II:** a nuisance from the neural net approximation lectures**.)**

Recall that the lectures on approximation of continuous functions by 2- and 3-layer networks did not include a nonlinearity on the final output. This exercise will set the record straight: namely, prove the following, where $\|f - g\|_1 = \int_{[0,1]^d} |f(x) - g(x)|dx$ as in lecture.

> Suppose a function class $\mathcal{F}$ is given so that for any continuous $g : \mathbb{R}^d \to \mathbb{R}$ and any $\tau > 0$, there exists $f \in \mathcal{F}$ with $\|f - g\|_1 \leq \tau$. Prove that for any sigmoidal $\sigma : \mathbb{R} \to [0, 1]$ (as in the previous problem) the function class $\mathcal{F}_\sigma := \{\sigma \circ f : f \in \mathcal{F}\}$ can approximate *appropriately restricted* continuous functions, meaning for any continuous $g : \mathbb{R}^d \to [0, 1]$ and any $\epsilon > 0$, there exists $f \in \mathcal{F}_\sigma$ with $\|f - g\|_1 \leq \epsilon$.

Easy mode / hint: feel free to assume σ is any combination of Lipschitz and bijective (with continuous inverse, even), and that $g : \mathbb{R}^d \to (0, 1)$ rather than $g : \mathbb{R}^d \to [0, 1]$, but please state these assumptions clearly.

Hard mode: *don't* make those assumptions, but be sure to check that your proof doesn't accidentally rely on them.

Solution.

3. **(A negative result** for single node neural networks**.)**

Suppose $\sigma : \mathbb{R} \to [0, 1]$ is sigmoidal as in the previous question. This question will develop a negative result on the representation power of single layer networks (in particular, networks with exactly 1 node). This result makes sense from the perspective of the result presented in class due to Minsky and Papert (1969); they, however, had some motivation in vision tasks, whereas here the task will be even simpler, namely univariate.

With this in mind, construct an appropriate continuous function $g : \mathbb{R} \to [0, 1]$ and use it to (constructively) prove the following:

> There exists a continuous function $g : \mathbb{R} \to [0, 1]$ and a real $c > 0$ so that
>
> $$\inf_{f \in \mathcal{H}_\sigma} \|f - g\|_1 \geq c$$
>
> where $\mathcal{H}_\sigma := \left\{x \mapsto \sigma(a^\top x + b) : a \in \mathbb{R}, b \in \mathbb{R}\right\}$.

Solution.

4. **(Polynomial approximation** (Weierstrass, 1885)**.)**

The goal of this problem is to prove the following version of the Weierstrass Approximation Theorem:

> For every continuous $g : \mathbb{R}^d \to \mathbb{R}$ and scalar $\epsilon > 0$, there exists a polynomial $f : \mathbb{R}^d \to \mathbb{R}$ so that
> $$\|f - g\|_{\mathrm{u}} = \sup\{|f(x) - g(x)| : x \in [0,1]^d\} \leq \epsilon.$$

The proof will proceed in a few steps. First recall (and don't bother to prove) the following analysis fact from lecture.

Lemma 1. *Given continuous $g : \mathbb{R}^d \to \mathbb{R}$ and scalar $\epsilon > 0$, there exists $\delta > 0$ so that every $x, x' \in [0,1]^d$ with $\|x - x'\|_\infty \leq \delta$ implies $|g(x) - g(x')| \leq \epsilon/2$, and an $M < \infty$ so that $\sup\{|g(x)| : x \in [0,1]^d\} \leq M$.*

In the rest of the proof, fix a continuous g as in the problem statement, let $\delta > 0$ and $M < \infty$ be given according to the preceding lemma. The steps of the proof are as follows.

(a) Let $(X_1, \ldots, X_d)$ denote independent random variables, where X_i has binomial distribution $B(n, x_i)$ corresponding to n flips of an x_i-biased coin. Prove that there exists n such that $n > 1/\delta^3$ and
$$\Pr\left[\exists i \in \{1, \ldots, d\} \centerdot |X_i - nx_i| > n^{2/3}\right] < \frac{\epsilon}{4M}.$$
(See below[1] for a hint.)

(b) Prove
$$\sum_{i_1=0}^{n} \sum_{i_2=0}^{n} \cdots \sum_{i_d=0}^{n} \prod_{j=1}^{d} \binom{n}{i_j} x_j^{i_j} (1 - x_j)^{n - i_j} = 1.$$

(c) Recalling that $\mathbf{e}_i$ denotes the ith standard basis vector, define the polynomial
$$f(x) := \sum_{i_1=0}^{n} \sum_{i_2=0}^{n} \cdots \sum_{i_d=0}^{n} g\left(\sum_{j=1}^{d} i_j \mathbf{e}_j / n\right) \prod_{j=1}^{d} \binom{n}{i_j} x_j^{i_j} (1 - x_j)^{n - i_j},$$
whose form conveniently relates to the $B(n, x_i)$ above. With this and the other parts of this question statement in mind, prove $\|f - g\|_{\mathrm{u}} \leq \epsilon$. (See below[2] for a hint.)

Solution.

[1] **Hint:** what are some useful general-purpose probability inequalities?

[2] **Hint:** split the sum defining f into two parts, one part being handled by the Lemma 1, and the other by the first part of the question... and surely the middle part of the question is useful too...

5. **(Convexity calisthenics.)**

(a) Given any $p \in [1, \infty]$, define $f_p : \mathbb{R}^d \to \mathbb{R}$ as

$$f_p(v) := \begin{cases} \frac{1}{p}\|v\|_p^p = \frac{1}{p}\sum_{i=1}^d |v_i|^p & \text{when } p \in [1, \infty), \\ \iota_{[-1,+1]^d}(v) & \text{when } p = \infty. \end{cases}$$

Prove for any *conjugate exponents* $p, q \geq 1$ (meaning $1/p + 1/q = 1$ with convention $1/\infty = 0$) that $f_p^* = f_q$.
(Note, we're *proving* Hölder's inequality, so don't use it as a step of this proof.)

(b) Prove f_p is convex for every $t \in [1, \infty]$.

(c) Under the notation (and making use of) the previous part, show for any $u, v \in \mathbb{R}^d$ with $\|u\|_p = 1 = \|v\|_q$ that $|\langle u, v\rangle| \leq 1$.

(d) Now use the previous part to prove the full version of Hölder's inequality (which implies the Cauchy-Schwarz inequality): given any $u, v \in \mathbb{R}^d$ and any conjugate exponents $p, q \geq 1$,

$$|\langle u, v\rangle| \leq \|u\|_p \|v\|_q.$$

(e) **Optional:** prove $(\|\cdot\|_p)_* = \|\cdot\|_q$, where $p, q \geq 1$ are conjugate exponents.

(f) Define $f(x) := x^\top Q x/2$, where $Q \in \mathbb{R}^{d\times d}$ a symmetric positive definite matrix. Derive and rigorously prove an explicit form for f^*. (**Hint:** try to guess the answer first.)

(g) Define $f(x) := x^\top Q x/2$ once again, but now $Q \in \mathbb{R}^{d\times d}$ is merely symmetric positive semi-definite. Now derive and rigorously prove a new expression for f^*. (**Hint**: there are many ways here, but again try to guess the answer, and in times of great need never forget your friend S-V-D.)

Solution.

6. **(Max of random variables; moment generating functions.)**

An important object in the study of random variables is the moment generating function (MGF), $M_X(t)$, defined as $M_X(t) := \mathbb{E}(\exp(tX))$. ($M_X$ will in general fail to be finite for all $t \geq 0$, but in this question it is finite for all $t \geq 0$.)

Given a family $(X_i, \ldots, X_d)$ of i.i.d. random variables drawn according to some distribution, this question will investigate the behavior of the random variable $Z := \|(X_1, \ldots, X_d)\|_\infty = \max_i |X_i|$.

(a) Prove the following inequality, which will be convenient in the remainder of the question: for any $t > 0$,

$$\mathbb{E}(Z) \leq \frac{1}{t} \ln \left(d\mathbb{E} \left(\exp(tX_1) + \exp(-tX_1) \right) \right).$$

Note / hint: consider waiting for the "convexity bootcamp" lectures.

Hint: start your proof from the rearranged left hand $\exp(t\mathbb{E}(Z))$.

(b) **Optional:** Suppose X_1 distributed according to a Gumbel distribution with scale parameter σ, whereby $\mathbb{E}(\exp(sX_1)) = \Gamma(1 - s\sigma)$ for all $s \in \mathbb{R}$, where Γ denotes the gamma function. Prove that

$$\mathbb{E}(Z) \leq 2\sigma \ln(2d\sqrt{\pi}).$$

(See below[3] for a hint.)

(c) Prove that Gaussian distribution is *subgaussian*: in particular, if X_1 is Gaussian with mean 0 and variance σ^2, then $\mathbb{E}(\exp(tX_1)) = \exp(t^2\sigma^2/2)$.

(d) Prove that if X_1 is subgaussian with parameter σ (meaning $\mathbb{E}(\exp(tX_1)) \leq \exp(t^2\sigma^2/2)$, then

$$\mathbb{E}(Z) \leq \sigma\sqrt{2\ln(2d)}.$$

(Together with the preceding part, this implies the bound for X_1 a Gaussian with mean 0 and variance σ^2.)

(e) Was it necessary to assume $(X_1, \ldots, X_d)$ were i.i.d.? Answer this question however you like.

Solution.

[3] The inequality from the first part holds for all t... can you find a particularly nice choice of t?

References

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, july 1989.

Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

Karl Weierstrass. Über die analytische darstellbarkeit sogenannter willkürlicher functionen einer reellen veränderlichen. *Sitzungsberichte der Akademie zu Berlin*, pages 633–639, 789–805, 1885.