CS 6.5: Theories of Deep Learning
Problem Sheet 2

Prof. Jared Tanner

October 2020

The questions marked with an asterisk should be submitted for marking. The remaining questions entail computational experiments which should be attempted, but need not be submitted; they will be discussed in tutorials.

Networks at initialisation

1. (*) Write a short report summarizing the paper "Exponential Expressivity in Deep Neural Networks Through Transient Chaos" by Poole et al.[1]. Your report should be written in the format and style of a NeurIPS Proceedings, abridged to not exceed 2 pages. Latex style files and an exemplar template are provided on the course page.

2. (a) Given a network defined by $x^l = \phi(h^l)$, $h^l = W^l x^{l-1} + b^l$, it was shown that in the large-width limit, $q^l = \frac{1}{N^l}\sum_i^{N_l}(h_i^l)^2$ converges to a q^* as l grows. Let $\sigma_w = 2$ and $\sigma_b = 0.3$, compute q^* to within 4 decimal places.

   Hint: For the numerical integration, consider using the function `quad(f)` algorithm in the python `scipy.integrate` package where, f is a function

   (b) Next, implement a deep and wide feed-forward fully connected network, with these same σ_w and σ_b. Check whether the empirically observed q^* matches what you calculated.

3. (*) Prove that

$$\chi_1 \equiv \left.\frac{\partial c_{12}^l}{\partial c_{12}^{l-1}}\right|_c = \sigma_w^2 \int \mathcal{D}z[\phi'(\sqrt{q^*}z)]^2. \tag{1}$$

   Hint: You will need to employ the following identity

$$\int \mathcal{D}zF(z)z = \int \mathcal{D}zF'(z) \tag{2}$$

4. Compute the set of points (σ_w, σ_b) such that $\chi_1 = 1$, for q^* varying between 1 and 1000, using the tanh activation function. Plot this curve (i.e. the 'edge of chaos').

5. Choose combinations of σ_w and σ_b which cause χ_1 to be > 1 and < 1 for a tanh network. In each case, pass two points through your feedforward network, and compute their correlation at each layer. What happens in each case? Are either of these behaviours desirable?

Trainability

1. (*) Show that $\chi_1 = \frac{1}{N}\left\langle \mathrm{Tr}((\mathbf{DW})^\top \mathbf{DW})\right\rangle$ in the limit of large N, where $\langle\cdot\rangle$ denotes the expectation, $\mathbf{W} \in \mathbb{R}^{N\times N}$ is a Gaussian random matrix and $\mathbf{D}$ is the diagonal random matrix with entries $D_{ij} = \phi'(h_i^l)\delta_{ij}$ for any $l > 0$, where h^0 was chosen so that $q^0 = q^*$. How does this relate to the singular vales of the matrix $\mathbf{DW}$? How does the matrix $\mathbf{DW}$ relate to back-propagation, and thus what are the implications of different values of χ_1 on trainability from a given initialisation?

[1] http://papers.nips.cc/paper/6322-exponential-expressivity-in-deep-neural-networks-through-transient-chaos.pdf

2. Returning to the network you implemented in Question 1, visualise (e.g. histogram or some other appropriate method) the gradients of the different layers, for different values of χ_1. What do you see?