Institut for Theoretical Physics
University Leipzig

Prof. Dr. B. Rosenow
M. Kühn

Statistical Mechanics of Deep Learning - Problem set 11

Winter Term 2024/25

The problem set will be discussed in the seminar on **Monday 13.01.2024, 9:15**.

20. On-line learning of soft committee machines

3+3 Points

Consider the order parameter dynamical equations of a soft committee machine with the activation function taken to be the error function, in the symmetric regime (we assume $R_{in} = R$, $Q_{ik} = Q$) and in the small η limit, `https://arxiv.org/abs/2104.14546`

$$\frac{dR}{d\alpha} = \frac{2\eta}{\pi}\,\frac{1}{1+Q}\,\frac{1}{K}\left\{\frac{1+Q-MR^2}{\sqrt{2(1+Q)-R^2}} - \frac{MR}{\sqrt{1+2Q}}\right\} \tag{1}$$

$$\frac{dQ}{d\alpha} = \frac{4\eta}{\pi}\,\frac{1}{1+Q}\,\frac{1}{K}\left\{\frac{MR}{\sqrt{2(1+Q)-R^2}} - \frac{MQ}{\sqrt{1+2Q}}\right\} \tag{2}$$

we take $M = K$, the number of hidden units of the teacher network M is equal to the student network K in the realizable case. The generalization error is given by

$$\begin{aligned}\varepsilon_g = \frac{K}{\pi}\Bigg[&\frac{\pi}{6} + \arcsin\left(\frac{Q}{1+Q}\right) + (K-1)\arcsin\left(\frac{C}{1+Q}\right) - 2\arcsin\left(\frac{R}{\sqrt{2(1+Q)}}\right)\\ &- 2(K-1)\arcsin\left(\frac{S}{\sqrt{2(1+Q)}}\right)\Bigg]\end{aligned} \tag{3}$$

(a) Show that equations (1) and (2) have the following fixed points, which correspond to the symmetric phase solution at the plateu discussed in the lecture

$$R_{pl} = \frac{1}{\sqrt{K(2K-1)}}\ ,\ Q_{pl} = \frac{1}{2K-1}$$

(b) Use the results obtained in (a) to compute the generalization error at the platue

$$\varepsilon_{pl} = \frac{K}{\pi}\left[\frac{\pi}{6} - K\arcsin\left(\frac{1}{2K}\right)\right]$$

Hint : note that on the platue, we have $R_{pl} = S_{pl}$ and $Q_{pl} = C_{pl}$

21. Neural Scaling Laws

3+3+6 Points

Consider online learning of a perceptron with a linear activation function in a student-teacher setup. The setup consits of the teacher vector $\boldsymbol{T} \in \mathbb{R}^N$ and the student vector $\boldsymbol{J}(\alpha) \in \mathbb{R}^N$, where α signifies the ratio between the number of examples already shown and the input dimension N. The examples are drawn from a multivariate Gaussian distribution $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ is diagonal with entries λ_i, $i = 1, \dots, N$, on its diagonal. The student output is given as $\sigma(\boldsymbol{J}, \boldsymbol{\xi}) = \boldsymbol{J} \cdot \boldsymbol{\xi}$, and the teacher output is given as $\tau(\boldsymbol{T}, \boldsymbol{\xi}) = \boldsymbol{T} \cdot \boldsymbol{\xi}$.

(a) The generalization error is given by the mean squared error:

$$\varepsilon_g(\boldsymbol{J}) = \frac{1}{2} \langle (\sigma(\boldsymbol{J}, \boldsymbol{\xi}) - \tau(\boldsymbol{T}, \boldsymbol{\xi}))^2 \rangle_{\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})} \ .$$

Show that when the examples have a diagonal covariance matrix as described before, the generalization error is equal to the following expression:

$$\varepsilon_g(\boldsymbol{J}) = \frac{1}{2} \sum_{i=1}^{N} \lambda_i (J_i - T_i)^2.$$

(b) The teacher vector is given as $T_i \equiv 1$ for $i = 1, \dots, N$, and the student vector is initialized as $J_i(0) = 0$ for $i = 1, \dots, N$. Consider the case of a comparatively small learning rate η in which the dynamics of the student vector are described by the following differential equation:

$$\frac{\partial \langle J_i \rangle}{\partial \alpha} = -\eta \frac{\partial \varepsilon_g}{\partial J_i}.$$

Solve the differential equation for $\langle \boldsymbol{J} \rangle(\alpha)$ with the given initial conditions. Demonstrate that the result for the generalization error, with a general diagonal covariance matrix as described before, is given by:

$$\varepsilon_g(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \lambda_i \exp(-2\eta\alpha\lambda_i).$$

(c) If all eigenvalues of the data covariance matrix are the same ($\lambda_i = \lambda \ \forall i$), it is straightforward to see from the previous equation that the generalization error will decay exponentially with increasing α. However, real-world datasets often exhibit power-law spectra in their covariance matrices. Therefore, assume that the eigenvalues of the given diagonal data covariance are given by:

$$\lambda_i = \frac{\lambda_+}{i^{1+\beta}}, \quad \beta > 0.$$

Here, λ_+ is a normalization factor, and for this task, it is sufficient to assume it to be a positive constant. Show that with this assumption, the generalization error calculated before follows a power-law decrease with increasing α, specifically:

$$\varepsilon_g(\alpha) \propto \alpha^{-\frac{\beta}{1+\beta}}.$$

Hint: Approximate the sum in the generalization error as an integral and make approximations by taking the limit of $N \to \infty$ and by assuming that a significant number of examples have already been shown to the network $\left(\alpha \gg \frac{1}{2\eta\lambda_+}\right)$. You may use the fact that for $0 < s < 1$ and for $x \gg 1$:

$$\begin{aligned} \int_0^x t^{s-1} e^{-t} dt &\approx \int_0^\infty t^{s-1} e^{-t} dt \\ &= \Gamma(s) \end{aligned}$$