# Neural Networks

1. **Multiclass Classification Error Function (Bishop 5.5)**

   Consider a $K$-class supervised classification scenario with training data $\{\boldsymbol{x}_n, \boldsymbol{t}_n\}_{n=1}^{N}$, where the $\boldsymbol{t}_n$ are 1-hot binary vectors, with $t_{nk}=1$ iff $\boldsymbol{x}_n$'s true class is $k$. Assume we model this problem using a neural network with $K$ output-units, where the interpretation of the $k$'th output unit, denoted $y_k(\boldsymbol{x}_n, \boldsymbol{w})$, is $y_k(\boldsymbol{x}_n, \boldsymbol{w}) = p(t_{nk}=1 \mid \boldsymbol{x}_n)$.
   Show that maximizing the (conditional) likelihood of such a model is equivalent to minimizing the cross-entropy loss function given by

   $$E(\boldsymbol{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{x}, \boldsymbol{w}) \qquad \text{(Bishop 5.24)}$$

2. **Activation Gradients (Bishop 5.6)**

Consider a single-output network solving a binary classification problem, and trained with the following cross-entropy error function

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N} \left[ t_n \ln y(\boldsymbol{x}_n, \boldsymbol{w}) + (1 - t_n) \ln(1 - y(\boldsymbol{x}_n, \boldsymbol{w})) \right], \qquad \text{(Compare Bishop 5.21)}$$

where $y(\boldsymbol{x}_n, \boldsymbol{w}) = \sigma(a_n) = \frac{1}{1+\exp(-a_n)}$ for an activation $a_n$. Show that the derivative of the error function above wrt a particular $a_n$ satisfies

$$\frac{\partial E}{\partial a_n} = y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n \qquad \text{(Compare Bishop 5.18)}$$

3. **Positive Definiteness of Hessian (Bishop 5.10)**

Consider a hessian matrix $\boldsymbol{H} \in \mathbb{R}^D$ with eigenvector equation

$$\boldsymbol{H}\boldsymbol{u}_d = \lambda_d \boldsymbol{u}_d, \qquad \text{(Bishop 5.33)}$$

where the eigenvectors $\boldsymbol{u}_d$ form an orthonormal basis of $\mathbb{R}^D$. As Bishop points out, since any vector $\boldsymbol{v} \in \mathbb{R}^D$ can be written as

$$\boldsymbol{v} = \sum_{d=1}^{D} c_d \boldsymbol{u}_d, \qquad \text{(Bishop 5.38)}$$

the orthonormality of the eigenvectors implies that

$$\boldsymbol{v}^\top \boldsymbol{H}\boldsymbol{v} = \sum_{d=1}^{D} c_d^2 \lambda_d. \qquad \text{(Bishop 5.39)}$$

(a) As a first step, show how to derive (Bishop 5.39) from Bishop (5.38).

(b) Now, recall that a matrix is positive definite if $\boldsymbol{v}^\top \boldsymbol{H}\boldsymbol{v} > 0$ for all nonzero $\boldsymbol{v}$. Show that $\boldsymbol{H}$ is positive definite if and only if all of its eigenvalues are positive.

4. **(Bishop 5.2)**

The outer product approximation to the Hessian matrix for a neural network using a sum-of-squares error function is given by

$$\boldsymbol{H} = \sum_{n=1}^{N} \boldsymbol{b}_n \boldsymbol{b}_n^{\mathsf{T}}$$
$$\boldsymbol{b}_n = \nabla y_n = \nabla a_n$$

This is the approximation to the Hessian matrix of $E = \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2$. Derive an expression for an approximation to the Hessian for the case of multiple outputs. Consider the multivariate generalization of the energy function:

$$E = \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{y}_n - \boldsymbol{t}_n)^{\mathsf{T}}(\boldsymbol{y}_n - \boldsymbol{t}_n)$$

5. **Likelihood Function of Conditional (Bishop 5.2)**

Show that maximizing the likelihood function under the conditional distribution given by

$$p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}\boldsymbol{I})$$

for a multioutput neural network is equivalent to minimizing the sum of squares error function given by

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N} ||\boldsymbol{y}(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{t}_n||^2$$

6. **Degrees of Freedom of the Hessian (Bishop 5.13)**

Determine that the number of independent elements (degrees of freedom) in the quadratic error function, given by the Taylor Expansion of the error function around the point $\hat{\boldsymbol{w}}$

$$E(\boldsymbol{w}) = E(\hat{\boldsymbol{w}}) + (\boldsymbol{w} - \hat{\boldsymbol{w}})^{\mathsf{T}}\boldsymbol{b} + \frac{1}{2}(\boldsymbol{w} - \hat{\boldsymbol{w}})^{\mathsf{T}}\boldsymbol{H}(\boldsymbol{w} - \hat{\boldsymbol{w}})$$

is given by $\frac{W(W+3)}{2}$. Hint: What properties of the Hessian matrix restrict its number of independent elements?