Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 9

Exercise 1: Neural Network Optimization (15 + 15 P)

Consider the one-layer neural network

$$y = \boldsymbol{w}^\top \boldsymbol{x} + b$$

applied to data points $\boldsymbol{x} \in \mathbb{R}^d$, and where $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are the parameters of the model. We consider the optimization of the objective:

$$J(\boldsymbol{w}) = \mathbb{E}_{\hat{p}}\Big[\frac{1}{2}(1 - y \cdot t)^2\Big],$$

where the expectation is computed over an empirical approximation $\hat{p}$ of the true joint distribution $p(\boldsymbol{x}, t)$ and $t \in \{-1, 1\}$. The input data follows the distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$ where $\boldsymbol{\mu}$ and σ^2 are the mean and variance.

(a) *Compute* the Hessian of the objective function J at the current location $\boldsymbol{w}$ in the parameter space, and as a function of the parameters $\boldsymbol{\mu}$ and σ of the data.

(b) *Show* that the condition number of the Hessian is given by: $\dfrac{\lambda_1}{\lambda_d} = 1 + \dfrac{\|\boldsymbol{\mu}\|^2}{\sigma^2}$.

Exercise 2: Neural Network Regularization (10 + 10 + 10 P)

For a neural network to generalize from limited data, it is desirable to make it sufficiently invariant to small local variations. This can be done by limiting the gradient norm $\|\partial f / \partial \boldsymbol{x}\|$ for all $\boldsymbol{x}$ in the input domain. As the input domain can be high-dimensional, it is impractical to minimize the gradient norm directly. Instead, we can minimize an upper-bound of it that depends only on the model parameters.

We consider a two-layer neural network with d input neurons, h hidden neurons, and one output neuron. Let W be a weight matrix of size $d \times h$, and $(b_j)_{j=1}^h$ a collection of biases. We denote by $W_{i,:}$ the ith row of the weight matrix and by $W_{:,j}$ its jth column. The neural network computes:

$$\begin{aligned} a_j &= \max(0, W_{:,j}^\top \boldsymbol{x} + b_j) && \text{(layer 1)} \\ f(\boldsymbol{x}) &= \textstyle\sum_j s_j a_j && \text{(layer 2)} \end{aligned}$$

where $s_j \in \{-1, 1\}$ are fixed parameters. The first layer detects patterns of the input data, and the second layer computes a fixed linear combination of these detected patterns.

(a) *Show* that the gradient norm of the network can be upper-bounded as:

$$\Big\|\frac{\partial f}{\partial \boldsymbol{x}}\Big\| \leq \sqrt{h} \cdot \|W\|_F$$

(b) Let $\|W\|_{\text{Mix}} = \sqrt{\sum_i \|W_{i,:}\|_1^2}$ be a ℓ_1/ℓ_2 mixed matrix norm. *Show* that the gradient norm of the network can be upper-bounded by it as:

$$\Big\|\frac{\partial f}{\partial \boldsymbol{x}}\Big\| \leq \|W\|_{\text{Mix}}$$

(c) *Show* that the mixed norm provides a bound that is tighter than the one based on the Frobenius norm, i.e. show that:

$$\|W\|_{\text{Mix}} \leq \sqrt{h} \cdot \|W\|_F$$

.

Exercise 3: Programming (40 P)

Download the programming files on ISIS and follow the instructions.

Exercise 1: Neural Network Optimization (15 + 15 P)

Consider the one-layer neural network

$$y = \boldsymbol{w}^\top \boldsymbol{x} + b$$

applied to data points $\boldsymbol{x} \in \mathbb{R}^d$, and where $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are the parameters of the model. We consider the optimization of the objective:

$$J(\boldsymbol{w}) = \mathbb{E}_{\hat{p}}\Big[\frac{1}{2}(1 - y \cdot t)^2\Big],$$

where the expectation is computed over an empirical approximation $\hat{p}$ of the true joint distribution $p(\boldsymbol{x}, t)$ and $t \in \{-1, 1\}$. The input data follows the distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$ where $\boldsymbol{\mu}$ and σ^2 are the mean and variance.

(a) *Compute* the Hessian of the objective function J at the current location $\boldsymbol{w}$ in the parameter space, and as a function of the parameters $\boldsymbol{\mu}$ and σ of the data.

a)

$$J(w) = \mathbb{E}_{\hat{p}}\left[\tfrac{1}{2}(1-yt)^2\right] = \mathbb{E}_{\hat{p}}\left[\tfrac{1}{2}(1-(w^T x+b)\cdot t)^2\right]$$

$$\frac{\partial J}{\partial w_i} = \mathbb{E}_{\hat{p}}\left[\tfrac{1}{2}\cdot 2\cdot(1-(w^T x+b)\cdot t)\cdot(-x_i\cdot t)\right]$$

$$= \mathbb{E}_{\hat{p}}\left[(1-(w^T x+b)\cdot t)\cdot(-x_i\cdot t)\right]$$

$$\frac{\partial J}{\partial w_j}\frac{\partial J}{\partial w_i} = \mathbb{E}_{\hat{p}}\left[(-x_i\cdot t)\cdot(-x_j\cdot t)\right]$$

$$= \mathbb{E}_{\hat{p}}\left[\underbrace{t^2}_{=1\text{, because of } t\in\{-1,1\}} x_i\cdot x_j\right]$$

$$= \mathbb{E}_{\hat{p}}\left[x_i\cdot x_j\right]$$

$$H = \frac{\partial^2 J}{\partial w^2} = \begin{bmatrix} \frac{\partial J}{\partial w_1}\frac{\partial J}{\partial w_1} & \frac{\partial J}{\partial w_1}\frac{\partial J}{\partial w_2} & \cdots & \frac{\partial J}{\partial w_1}\frac{\partial J}{\partial w_d} \\ \frac{\partial J}{\partial w_2}\frac{\partial J}{\partial w_1} & \ddots & & \vdots \\ \vdots & & & \\ \frac{\partial J}{\partial w_d}\frac{\partial J}{\partial w_1} & \cdots & & \frac{\partial J}{\partial w_d}\frac{\partial J}{\partial w_d} \end{bmatrix}$$

$$= \mathbb{E}[x x^T]$$

$$= Cov(x, x^T) + \mathbb{E}[x]\mathbb{E}[x^T]$$

$$= \delta^2 \cdot I + \mu\cdot\mu^T$$

(b) *Show* that the condition number of the Hessian is given by: $\dfrac{\lambda_1}{\lambda_d} = 1 + \dfrac{\|\boldsymbol{\mu}\|^2}{\sigma^2}.$

Assume v is the eigenvalue of λ, then $Hv = \lambda v$

$$\Leftrightarrow \quad v^T H v = \lambda \cdot v^T v = \lambda \cdot \|v\|^2$$

$$\Leftrightarrow \quad \lambda = \frac{v^T H v}{\|v\|^2}$$

$$\lambda_1 = \lambda_{max} = \max \frac{v^T H v}{\|v\|^2}$$

$$= \max \frac{1}{\|v\|^2} \cdot v^T \cdot (\delta^2 I + \mu\mu^T) v$$

$$= \max \frac{1}{\|v\|^2} \delta^2 \cdot v^T v + \frac{1}{\|v\|^2} v^T \mu\mu^T v$$

$$= \max \frac{1}{\|v\|^2} \cdot \delta^2 \cdot \|v\|^2 + \frac{1}{\|v\|^2} \cdot \underbrace{(v^T \mu)}_{(1\times d)} \cdot \underbrace{(v^T \mu)^T}_{(d\times 1)}$$

$$= \max \delta^2 + \frac{1}{\|v\|^2} \cdot \|v^T \cdot \mu\|^2 \qquad = \max \delta^2 + \frac{1}{\|v\|^2} (v^T \mu)^2$$

$$= \max \delta^2 + \frac{\| \|v\| \cdot \|\mu\| \cdot \cos\theta \|^2}{\|v\|^2}$$

$$= \delta^2 + \|\mu\|^2 \qquad (\text{if } |\cos\theta| = 1) \qquad \text{max if } v \text{ alligns with } \mu \Rightarrow \frac{v}{\|v\|} = \frac{\mu}{\|\mu\|}$$

$$\lambda_d = \delta^2 \qquad (\text{if } \|\mu\| = 0) \qquad \text{min if } v \text{ orthogonal to } \mu \Rightarrow v^T \mu = 0$$

$$\frac{\lambda_1}{\lambda_d} = \frac{\delta^2 + \|\mu\|^2}{\delta^2} = 1 + \frac{\|\mu\|^2}{\delta^2}$$

Exercise 2: Neural Network Regularization (10 + 10 + 10 P)

For a neural network to generalize from limited data, it is desirable to make it sufficiently invariant to small local variations. This can be done by limiting the gradient norm $\|\partial f/\partial \boldsymbol{x}\|$ for all $\boldsymbol{x}$ in the input domain. As the input domain can be high-dimensional, it is impractical to minimize the gradient norm directly. Instead, we can minimize an upper-bound of it that depends only on the model parameters.

We consider a two-layer neural network with d input neurons, h hidden neurons, and one output neuron. Let W be a weight matrix of size $d \times h$, and $(b_j)_{j=1}^h$ a collection of biases. We denote by $W_{i,:}$ the ith row of the weight matrix and by $W_{:,j}$ its jth column. The neural network computes:

$$a_j = \max(0, W_{:,j}^\top \boldsymbol{x} + b_j) \qquad \text{(layer 1)}$$

$$f(\boldsymbol{x}) = \textstyle\sum_j s_j a_j \qquad \text{(layer 2)}$$

where $s_j \in \{-1, 1\}$ are fixed parameters. The first layer detects patterns of the input data, and the second layer computes a fixed linear combination of these detected patterns.

(a) *Show* that the gradient norm of the network can be upper-bounded as:

$$\left\| \frac{\partial f}{\partial \boldsymbol{x}} \right\| \leq \sqrt{h} \cdot \|W\|_F$$

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{h} \frac{\partial f}{\partial a_j} \cdot \frac{\partial a_j}{\partial x_i} = \sum_{j=1}^{h} s_j \cdot 1_{a_j>0} \cdot \frac{\partial (W_{:,j}^\top x + b_j)}{\partial x_i}$$

$$= \sum_{j=1}^{h} s_j \cdot 1_{a_j>0} \cdot \frac{\partial \sum_{i=1}^{d} W_{i,j} \cdot x_i}{\partial x_i}$$

$$= \sum_{j=1}^{h} s_j \cdot 1_{a_j>0} \cdot W_{i,j}$$

$$\left\| \frac{\partial f}{\partial x} \right\| = \sqrt{\sum_{i=1}^{d} \left(\frac{\partial f}{\partial x_i}\right)^2} = \sqrt{\sum_{i=1}^{d} \left(\sum_{j=1}^{h} s_j \cdot 1_{a_j>0} \cdot W_{i,j} \right)^2}$$

$$\leq \sqrt{\sum_{i=1}^{d} \left(\sum_{j=1}^{h} s_j \cdot 1_{a_j>0}^2 \right) \left(\sum_{j=1}^{h} W_{i,j}^2 \right)}$$

$$\leq \sqrt{\sum_{i=1}^{d} \left(\sum_{j=1}^{h} s_j^2 \right) \left(\sum_{j=1}^{h} W_{i,j}^2 \right)}$$

$$\leq \sqrt{\sum_{i=1}^{d} h \cdot \left(\sum_{j=1}^{h} W_{i,j}^2 \right)}$$

$$= \sqrt{h} \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{h} W_{i,j}^2}$$

$$= \sqrt{h} \cdot \|W\|$$

(b) Let $\|W\|_{\text{Mix}} = \sqrt{\sum_i \|W_{i,:}\|_1^2}$ be a ℓ_1/ℓ_2 mixed matrix norm. *Show* that the gradient norm of the network can be upper-bounded by it as:

$$\left\|\frac{\partial f}{\partial \boldsymbol{x}}\right\| \leq \|W\|_{\text{Mix}}$$

$$\left\|\frac{\partial f}{\partial x}\right\| = \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} s_j 1_{a_j>0} \cdot W_{i,j}\right)^2}$$

$$\leq \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} |W_{i,j}|\right)^2}$$

$$= \sqrt{\sum_{i=1}^{d} \|W_{i,:}\|_1^2}$$

$$= \|W\|_{\text{Mix}}$$

$$\|W\|_{\text{Mix}}^2 = \sum_i \|W_{i,:}\|_1^2$$

$$= \sum_i \left(\sum_j |W_{ij}|\right)^2$$

(c) *Show* that the mixed norm provides a bound that is tighter than the one based on the Frobenius norm, i.e. show that:

$$\|W\|_{\text{Mix}} \leq \sqrt{h} \cdot \|W\|_F$$

$$\|W\|_{\text{Mix}} = \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} |W_{i,j}|\right)^2}$$

$$\leq \sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{h} 1^2 \cdot \sum_{j=1}^{h} W_{i,j}^2\right)} \quad \text{Cauchy-Schwarz inequality}$$

$$\leq \sqrt{\sum_{i=1}^{d} h \cdot \sum_{j=1}^{h} W_{i,j}^2}$$

$$= \sqrt{h} \cdot \|W\|_F$$