Exercises for the course
Machine Learning 1
Winter semester 2021/22

Abteilung 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 12

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:

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

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:

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

(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.

ML1 Sheet 12

1 a) The Hessian of f can be obtained by taking the second partial derivative:

$$H = \frac{\partial}{\partial w}\left(\frac{\partial}{\partial w} f(w)\right) = \frac{\partial}{\partial w}\left(\frac{\partial}{\partial w} \mathbb{E}_{\hat{p}}\left[\frac{1}{2}(1 - y t)^2\right]\right)$$

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

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

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

$$= \frac{\partial}{\partial w w^T} \mathbb{E}_{\hat{p}}\left[\frac{1}{2} - w^T x t + b t + t^2 w^T x x^T w + t^2 w^T x b + t^2 b^2 \frac{1}{2}\right]$$

$$= \mathbb{E}_{\hat{p}}[0 - 0 + 0 + t^2 x x^T + 0 + 0]$$

$$= \underbrace{t^2}_{=1} \mathbb{E}_{\hat{p}}[x x^T] = \mathbb{E}_{\hat{p}}[x x^T]$$

$$= \mathrm{cov}(x) + \mathbb{E}[x]\,\mathbb{E}[x]^T = \sigma^2 \mathbb{1} + \vec{\mu}\vec{\mu}^T$$

b) Show $\frac{\lambda_1}{\lambda_d} = 1 + \frac{\|\mu\|^2}{\sigma^2}$

$$\lambda_1 = \max_{\|v\|=1} v^T H v = \max_{\|v\|=1} v^T (\sigma^2 \mathbb{1} + \vec{\mu}\vec{\mu}^T) v = \max_{\|v\|=1} \sigma^2 + (v^T \mu)^2$$

$$= \sigma^2 + \left\| \frac{\mu^T}{\|\mu\|} \mu \right\|^2 = \sigma^2 + \|\mu\|^2$$

$$\lambda_2 = \max_{\substack{\|v\|=1 \\ v \perp \mu = 0}} v^T H v = \max_{\substack{\|v\|=1 \\ v^T \mu = 0}} v^T (\sigma^2 \mathbb{1} + \mu\mu^T) v = \sigma^2 = \lambda_3 = \dots = \lambda_d$$

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

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

Cauchy-Schwarz: $\langle a, b \rangle^2 \leq \|a\|^2 \|b\|^2$

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

$\leq \sum_{i=1}^{d} h \sum_{j=1}^{h} w_{ij}^2 = h \|W\|_F^2$

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

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

$\leq \sum_{i=1}^{d} \left(\sum_{j=1}^{h} |w_{ij}| \right)^2 = \sum_{i=1}^{d} \|w_{ij}\|_1^2$

$\Rightarrow \left\| \frac{\partial f}{\partial x} \right\| \leq \|W\|_{1,\infty}$

c) $\|W\|_{1,\infty}^2 = \sum_{i=1}^{d} \left(\sum_{j=1}^{h} 1_j \right)^2 \leq \sum_{i=1}^{d} \underbrace{\left(\sum_{j=1}^{h} 1_j^2 \right)}_{h} \left(\sum_{j=1}^{h} w_{ij}^2 \right)$

$= h \|W\|_F^2$