Exercises for the course

Machine Learning 1

Winter semester 2023/24

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.

Solution:

$$H = \frac{\partial}{\partial w}\left(\frac{\partial J}{\partial w}\right) = \frac{\partial}{\partial w}\left(\frac{\partial}{\partial w} E_{\hat{p}}\left[\frac{1}{2}(1-(w^T x+b)\cdot t)^2\right]\right)$$

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

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

$$= 0 + \frac{\partial}{\partial w}\left(\frac{\partial}{\partial w} E_{\hat{p}}\left[\frac{1}{2}(w^T x+b)^2\right]\right)$$

$$= E_{\hat{p}}[x x^T]$$

$$= Cov(x) + E[x]E[x]^T$$

$$= \sigma^2 I + \mu\mu^T$$

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

Solution:

$\because$ λ_1 is the biggest eigenvalue, so

$$\lambda_1 = \max_{\|v\|=1} v^T H v$$

$$= \max_{\|v\|=1} v^T(\sigma^2 I + \mu\mu^T)v$$

$$= \max_{\|v\|=1} \sigma^2 v^T I v + v^T\mu\mu^T v$$

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

日期:　　/

Since if $\nu = \frac{\mu^T}{\|\mu\|}$. which ν is align with μ. then $\nu^T u$ is maximized

$\therefore$ $$\lambda_1 = \sigma^2 + \left(\frac{\mu^T \mu}{\|\mu\|}\right)^2$$

$$\lambda_1 = \sigma^2 + \|\mu\|_2^2$$

And since Hessian H is symmetric, then all other eigen-vectors are orthogonal to $\left(\frac{\mu^T}{\|\mu\|}\right) \longrightarrow V_{remain} \cdot \left(\frac{\mu^T}{\|\mu\|}\right) = 0$

$\therefore$ $$\lambda_2 \cdots \lambda_d = \sigma^2$$

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

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:

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

Solution:

Cauchy-Schwarz inequality is written as

$$\sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i \geqslant \left(\sum_{i=1}^{N} a_i b_i\right)^2$$

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

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

$$\leqslant \sum_{i=1}^{d}\left(h \cdot \sum_{j=1}^{h} W_{ij}^2\right) = h \cdot \|W\|_F^2$$

$$\therefore \quad \left\|\frac{\partial f}{\partial x}\right\| \leqslant \sqrt{h} \cdot \|W\|_F$$

proofed.

(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:

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

Solution:

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

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

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

$$= \|W\|_{Mix}^2$$

$$\therefore \quad \left\|\frac{\partial f}{\partial x}\right\| \leq \|W\|_{Mix}$$

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

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

Solution:

$$\|W\|_{Mix}^2 = \sum_{i=1}^{d}\|W_{i,:}\|_1^2 \qquad\qquad h\cdot\|W\|_F^2 = \sum_{i=1}^{d}\left(h\cdot\sum_{j=1}^{h}W_{ij}^2\right)$$

$$\sum_{i=1}^{d}\|W_{i,:}\|_1^2 = \sum_{i=1}^{d}\left(\sum_{j=1}^{h}|W_{ij}|\right)^2 \leq \sum_{i=1}^{d}\left(\sum_{j=1}^{h}(1)^2\sum_{j=1}^{h}|W_{ij}|^2\right) = h\cdot\|W_F\|^2$$

$$\therefore \quad \sum_{i=1}^{d}\|W_{i,:}\|_1^2 \leq h\cdot\|W_F\|^2$$

$\therefore$ $\|W\|_{Mix}$ is a tighter bound.