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.

Neural Networks 2

In this homework, we will train neural networks on the Breast Cancer dataset. For this, we will use of the Pytorch library. We will also make use of scikit-learn for the ML baselines. A first part of the homework will analyze the parameters of the network before and after training. A second part of the homework will test some regularization penalties and their effect on the generalization error.

Breast Cancer Dataset

The following code extracts the Breast cancer dataset in a way that is already partitioned into training and test data. The data is normalized such that each dimension has mean 0 and variance 1. To test the robustness of our learning models, we also artificially inject 4% of mislabelings in the training data.

In [1]:

```
import utils

Xtrain,Ttrain,Xtest,Ttest = utils.breast_cancer()

nx = Xtrain.shape[1]
nh = 100
```

Neural Network Classifier

In this homework, we consider the same architecture as the one considered in Exercise 2 of the theoretical part. The class `NeuralNetworkClassifier` implements this network. The function `reg` is a regularizer which we set initially to zero (i.e. no regularizer). Because the dataset is small, the network can be optimized in batch mode, using the Adam optimizer.

In [2]:

```
import numpy,torch,sklearn,sklearn.metrics
from torch import nn,optim

class NeuralNetworkClassifier:

    def __init__(self):

        torch.manual_seed(0)

        self.model = nn.Sequential(nn.Linear(nx,nh),nn.ReLU())
        with torch.no_grad(): list(self.model)[0].weight *= 0.1
        self.s = torch.zeros([100]); self.s[:50] = 1; self.s[50:] = -1
        self.pool  = lambda y: y.matmul(self.s)
        self.loss  = lambda y,t: torch.clamp(1-y*t,min=0).mean()

    def reg(self): return 0

    def fit(self,X,T,nbit=10000):

        X = torch.Tensor(X)
        T = torch.Tensor(T)

        optimizer = optim.Adam(self.model.parameters(),lr=0.01)
        for _ in range(nbit):
            optimizer.zero_grad()
            (self.loss(self.pool(self.model(X)),T)+self.reg()).backward()
            optimizer.step()

    def predict(self,X):
        return self.pool(self.model(torch.Tensor(X)))

    def score(self,X,T):
        Y = numpy.sign(self.predict(X).data.numpy())
        return sklearn.metrics.accuracy_score(T,Y)
```

Neural Network Performance vs. Baselines

We compare the performance of the neural network on the Breast cancer data to two other classifiers: a random forest and a support vector classification model with RBF kernel. We use the scikit-learn implementation of these models, with their default parameters.

In [ ]:

```
from sklearn import ensemble,svm

rfc = ensemble.RandomForestClassifier(random_state=0)
rfc.fit(Xtrain,Ttrain)

svc = svm.SVC()
svc.fit(Xtrain,Ttrain)

nnc = NeuralNetworkClassifier()
nnc.fit(Xtrain,Ttrain)
```

In [4]:

```
def pretty(name,model):
    return '> %10s | Train Acc: %6.3f | Test Acc: %6.3f'%(name,model.score(Xtrain,Ttrain),model.score(Xtest,Ttest
))

print(pretty('RForest',rfc))
print(pretty('SVC',svc))
print(pretty('NN',nnc))
```

```
>    RForest | Train Acc:  1.000 | Test Acc:  0.940
>        SVC | Train Acc:  0.958 | Test Acc:  0.951
>         NN | Train Acc:  1.000 | Test Acc:  0.884
```

The neural network performs not as good as the baselines. Most likely, the neural network has overfitted its decision boundary, in particular, on the mislabeled training examples.

Gradient, and Parameter Norms (25 P)

For the model to generalize better, we assume that the gradient of the decision function should be prevented from becoming too large. Because the gradient can only be evaluated on the current data distribution (and may not generalize outside the data), we resort to the following inequality we have proven in the theoretical section for this class of neural network models:
$$ \text{Grad} \leq \|W\|_\text{Mix} \leq \sqrt{h}\|W\|_\text{Frob} $$

where

- $\|W\|_\text{Frob} = \sqrt{\sum_{i=1}^d \sum_{j=1}^h w_{ij}^2}$
- $\|W\|_\text{Mix} = \sqrt{\sum_{i=1}^d \big(\sum_{j=1}^h | w_{ij}|\big)^2}$
- $\text{Grad} = \textstyle \frac1N \sum_{n=1}^N\|\nabla_{\boldsymbol{x}}f (\boldsymbol{x_n})\|$

and where d is the number of input features, h is the number of neurons in the hidden layer, and W is the matrix of weights in the first layer (*Note that in PyTorch, the matrix of weights is given in transposed form*).

As a first step, we would like to keep track of these quantities during training. The function `Frob(nn)` that computes $\|W\|_\text{Frob}$ is already implemented for you.

Tasks:

- Implement the function `Mix(nn)` that receives the neural network as input and returns $\|W\|_\text{Mix}$.
- Implement the function `Grad(nn,X)` that receives the neural network and some dataset as input, and computes the averaged gradient norm (Grad).

In [5]:

```
def Frob(nn):
    W = list(nn.model)[0].weight
    return (W**2).sum()**.5

def Mix(nn):
    # ---------------------------------
    # TODO: Replace by your code
    # ---------------------------------
    import solution; return solution.Mix(nn)
    # ---------------------------------

def Grad(nn,X):
    # ---------------------------------
    # TODO: Replace by your code
    # ---------------------------------
    import solution; return solution.Grad(nn,X)
    # ---------------------------------
```

The following code measures these three quantities before and after training the model.

In [6]:

```
def fullpretty(name,nn):
    return pretty(name,nn) + ' | Grad: %7.3f | WMix: %7.3f | sqrt(h)*WFrob: %7.3f'%(Grad(nn,Xtest),Mix(nn),nh**.5
*Frob(nn))

nnr = NeuralNetworkClassifier()
print(fullpretty('Before',nnr))
nnr.fit(Xtrain,Ttrain)
print(fullpretty('After',nnr))
```

```
>     Before | Train Acc:  0.391 | Test Acc:  0.372 | Grad:   0.389 | WMix:   4.966 | sqrt(h)*WFrob:
5.751
>      After | Train Acc:  1.000 | Test Acc:  0.884 | Grad:   7.297 | WMix:  40.103 | sqrt(h)*WFrob:
56.739
```

We observe that the inequality $\text{Grad} \leq \|W\|_\text{Mix} \leq \sqrt{h} \|W\|_\text{Frob}$ we have proven also holds empirically. We also observe that these quantities tend to increase as training proceeds. This is a typical behavior, as the network starts rather simple and becomes complex as more and more variations in the training data are being captured.

Norm Penalties (15 P)

We consider the new objective $J^\text{Frob}(\theta) = \text{MSE}(\theta) + \lambda \cdot (\sqrt{h} \|W\|_\text{Frob})^2$, where the first term is the original mean square error objective and where the second term is the added penalty. We hardcode the penalty coeffecient to $\lambda = 0.005$. In principle, for maximum performance and fair comparison between the methods, several of them should be tried (also for other models), and selected based on some validation set. Here, for simplicity, we omit this step.

A downside of the Frobenius norm is that it is not a very tight upper bound of the gradient, that is, penalizing it is does not penalize specifically high gradient. Instead, other useful properties of the model could be negatively affected by it. Therefore, we also experiment with the mixed-norm regularizer $\textstyle \lambda \cdot \|W\|_\text{Mix}^2$, which is a tighter bound of the gradient, and where we also hardcode the penalty coefficient to $\lambda = 0.025$.

Task:

- Create two new classifiers by reimplementing the regularization function with the Frobenius norm regularizer and Mixed norm regularizer respectively. You may for this task call the norm functions implemented in the question above, but this time you also need to ensure that these functions can be differentiated w.r.t. the weight parameters.

The code below implements and train neural networks with the new regularizers, and compares the performance with the previous models.

In [7]:

```
import solution

class FrobClassifier(NeuralNetworkClassifier):

    def reg(self):
        # ---------------------------------
        # TODO: Replace by your code
        # ---------------------------------
        import solution; return solution.FrobReg(self)
        # ---------------------------------

class MixClassifier(NeuralNetworkClassifier):

    def reg(self):
        # ---------------------------------
        # TODO: Replace by your code
        # ---------------------------------
        import solution; return solution.MixReg(self)
        # ---------------------------------
```

In [8]:

```
nnfrob = FrobClassifier()
nnfrob.fit(Xtrain,Ttrain)

nnmix = MixClassifier()
nnmix.fit(Xtrain,Ttrain)
```

In [9]:

```
print(pretty('RForest',rfc))
print(pretty('SVC',svc))
print(fullpretty('NN',nnc))
print(fullpretty('NN+Frob',nnfrob))
print(fullpretty('NN+Mix',nnmix))
```

```
>    RForest | Train Acc:  1.000 | Test Acc:  0.940
>        SVC | Train Acc:  0.958 | Test Acc:  0.951
>         NN | Train Acc:  1.000 | Test Acc:  0.884 | Grad:   7.297 | WMix:  40.103 | sqrt(h)*WFrob:
56.739
>    NN+Frob | Train Acc:  0.961 | Test Acc:  0.954 | Grad:   0.735 | WMix:   1.767 | sqrt(h)*WFrob:
2.689
>     NN+Mix | Train Acc:  0.951 | Test Acc:  0.961 | Grad:   0.745 | WMix:   1.637 | sqrt(h)*WFrob:
4.138
```

We observe that the regularized neural networks now performs on par with the baselines. It is interesting to observe that the mixed norm penalty more selectively reduced the gradient, and has let the Frobenius norm take higher values.

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

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