Exercise Sheet 6

Due: 01.12.25 14h00

Christof Gehrig, Gerrit Gerhartz

General Regulations.

- Please hand in your solutions in groups of two (preferably from the same tutorial group).
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX. For scanned handwritten notes, please ensure they are legible and not blurry.
- For the practical exercises, always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter.
- Please hand in a **single PDF** that includes both the exported notebook and your solutions to the theoretical exercises. Submit the PDF to the Übungsgruppenverwaltung once per group, making sure to include the names of both group members in the submission.
- You can find all the data in the GitHub Repository.

1 Forward- and reverse-mode differentiation

We explore the computation of derivatives on general acyclic computational graphs. Consider the following function:

$$y = \exp\left[\exp(x) + \exp(x)^2\right] + \sin\left[\exp(x) + \exp(x)^2\right]$$

(a) Write down the most efficient computational graph with intermediate functions f_i, where $+, \exp, \cdot^2$ and sin are the primitive computational steps. (1 pt)

(b) Compute the derivative $\frac{\partial y}{\partial x}$ by *forward-mode differentiation*. In other words, compute the derivatives of your intermediate functions starting from the x-node and going *forward* to the y-node. Use the chain rule in each case to make use of the derivatives you already computed. (2 pts)

(c) Compute the derivative $\frac{\partial y}{\partial x}$ by *reverse-mode differentiation*. In other words, compute the derivatives of your intermediate functions starting from the y-node and going *reverse* to the x-node. Use the chain rule in each case to make use of the derivatives you already computed. (2 pts)

2 ADAM optimizer

(a) Write down the formulae for the Adaptive Moment Estimation (ADAM) optimizer, and write a short sentence describing what each line does. (1 pt)

(b) Show that in the very first iteration, the update of the parameters is reduced to the sign of the gradient g. (2 pts)

(c) How could this potentially unwanted behavior be avoided? (1 pt)

(d) Consider an MLP with a set of weights w trained with Adam and L2-regularization. Does it make a difference whether the L2-penalty $\|w\|_2^2$ is included in the loss, or whether weight decay is applied to the weights directly? Can you argue why one may be better than the other? *Hint: AdamW.* (1 pt)

3 Heteroscedastic regression

Use the dataset `x_y.csv` in the data folder. It contains 1,000 data points with 1-dimensional inputs, $\boldsymbol{x}$, and 1-dimensional outputs, y, as shown in the following figure

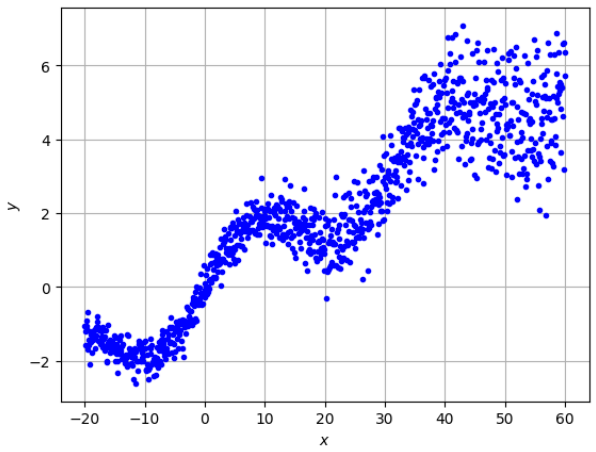


Figure 1: Scatter plot of the dataset contained in `x_y.csv`

(a) Load the dataset, and randomly split it into a training and a test set with ratio (66/33%). (1 pt)

(b) We model the distribution of outputs as

$$p(y|\boldsymbol{x},\boldsymbol{\theta}) = \mathcal{N}(y|f_\mu(\boldsymbol{x}),\sigma^2),$$

where $f_\mu(\boldsymbol{x})$ will be a multilayer perceptron (MLP), and we assume that all measurement errors are identical (i.e., homoscedastic regression), $\sigma(\boldsymbol{x}) = \sigma$. Explain why the mean-squared error is a reasonable log-likelihood function. (1 pt)

(c) Build an MLP with the following architecture: 3 fully-connected linear layers with hidden dimensions 32 and 64 (i.e., the dimensionality of your network should yield $1 \to 32 \to 64 \to 1$). Connect them with the ReLU activation function. Optimize the log-likelihood on the training set. Feel free to use any optimizer (e.g., stochastic gradient descent or Adam). Plot the resulting model on the test set together with the reference data points. Are you underfitting, overfitting? (2 pts)

(d) Play around with the architecture (number of layers / number of neurons / activation functions / regularization / ...) and try to improve the performance of the model. Plot train and validation errors for

- a learning rate that is too big,
- a learning rate that is too small and
- a learning rate scheduler: linear ramp up with cosine annealing. (2 pts)

(e) One standard way to address the prediction of heteroscedastic data for regression is to predict both the mean and the variance of a Normal distribution: $f_\mu(\boldsymbol{x}) = E[y|\boldsymbol{x},\boldsymbol{\theta}]$ and $f_{\sigma^2}(\boldsymbol{x}) = \mathrm{Var}[y|\boldsymbol{x},\boldsymbol{\theta}]$. Derive a log-likelihood function for this heteroscedastic problem. (2 pts)

(f) Modify the MLP to output two values: the mean and the variance. Optimize the model. Plot the mean and variance predicted by your model on the test set together with the reference data points. Are you underfitting, overfitting? (2 pts)

Hint: In case you use PyTorch, you may find the following functions useful to reshape tensors and arrays: `numpy.reshape` and `torch.squeeze`.

4 Bonus: ResNet properties

Consider a ResNet with L residual blocks. Each block takes an input x and outputs

$$T(x) = x + F(x),$$

and the full network is the composition of these blocks:

$$f = T_L \circ \cdots \circ T_1.$$

Assume that each block function F is 1-Lipschitz (i.e., $\|F(x) - F(y)\| \leq \|x - y\|$) and that there is no batch normalization or any other operation that changes the Lipschitz constant.

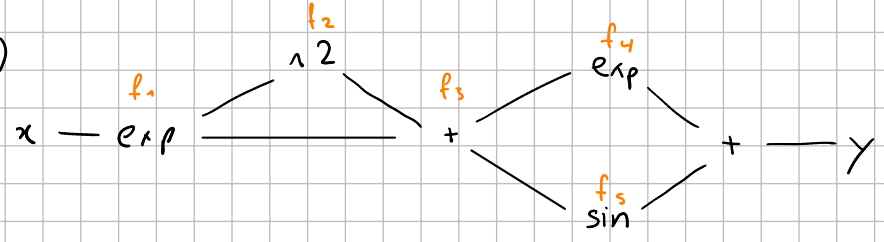
(a) Explain why the mapping $T(x) = x + F(x)$ is still Lipschitz. Derive an upper bound for $\mathrm{Lip}(T)$. (1 bonus pt)

(b) Derive an upper bound on the Lipschitz constant of the full network f as a function of depth L. (2 bonus pts)

(c) Explain why including batch normalization complicates the guarantee of a given Lipschitz constant at test time. (2 bonus pts)

①

a)



b) $f_1 = e^x \Rightarrow f_1' = e^x$

$f_2 = f_1^2 \Rightarrow f_2' = 2 f_1 f_1'$

$f_3 = f_1 + f_2 \Rightarrow f_3' = f_1' + f_2'$

$f_4 = e^{f_3} \Rightarrow f_4' = e^{f_3} f_3'$

$f_5 = \sin(f_3) \Rightarrow f_5' = \cos(f_3) f_3'$

$y = f_4 + f_5 \Rightarrow y' = f_4' + f_5'$

$$\Rightarrow y' = e^{f_3} f_3' + \cos(f_3) f_3' = \exp(f_1 + f_2)(f_1' + f_2') + \cos(f_1 + f_2)(f_1' + f_2')$$
$$= \exp(e^x + f_1^2)(e^x + 2 f_1 f_1') + \cos(e^x + f_1^2)(e^x + 2 f_1 f_1')$$
$$= \exp[e^x + (e^x)^2](e^x + 2e^{2x}) + \cos[e^x + (e^x)^2](e^x + 2e^{2x})$$

c) $\bar{f}_i = \frac{\partial y}{\partial f_i}$

$\Rightarrow \bar{f}_5 = 1, \quad \bar{f}_4 = 1$

$\rightarrow f_4 = \exp(f_3) \Rightarrow \frac{\partial f_4}{\partial f_3} = \exp(f_3) = f_4 \Rightarrow \bar{f}_3^{(f_4)} = \bar{f}_4 \frac{\partial f_4}{\partial f_3} = 1 \cdot \exp(f_3)$

$\Rightarrow f_5 = \sin(f_3) \Rightarrow \frac{\partial f_5}{\partial f_3} = \cos(f_3) \Rightarrow \bar{f}_3^{(f_5)} = \bar{f}_5 \frac{\partial f_5}{\partial f_3} = 1 \cdot \cos(f_3)$

$\Rightarrow f_3 = f_1 + f_2 \Rightarrow \bar{f}_1^{(f_3)} = \bar{f}_3 \frac{\partial f_3}{\partial f_1} = \bar{f}_3 \cdot 1$

$\Rightarrow \bar{f}_2^{(f_3)} = \bar{f}_3 \frac{\partial f_3}{f_2} = \bar{f}_3 \cdot 1$

$\Rightarrow f_2 = f_1^2 \Rightarrow \bar{f}_1^{(f_2)} = \bar{f}_2 \frac{\partial f_2}{\partial f_1} = \bar{f}_2 \, 2 f_1$

$\Rightarrow f_1 = e^x \Rightarrow f_1' = e^x$

$$\Rightarrow y' = \frac{\partial y}{\partial f_1} \frac{\partial f_1}{\partial x} = \bar{f}_1 f_1' = (\bar{f}_1^{(f_2)} + \bar{f}_1^{(f_3)}) e^x$$
$$= (2 f_1 \bar{f}_2 + \bar{f}_3) e^x = (2 f_1 \bar{f}_3 + \bar{f}_3) e^x = (2 f_1 + 1) \bar{f}_3 e^x$$
$$= (2 f_1 + 1)(\bar{f}_3^{(4)} + \bar{f}_3^{(5)}) e^x = (2 f_1 + 1)[e^{f_3} + \cos(f_3)] e^x$$
$$= (2 f_1 + 1)[\exp(f_1 + f_2) + \cos(f_1 + f_2)] e^x$$
$$= (2e^x + 1)\{\exp[e^x + (e^x)^2] + \cos[e^x + (e^x)^2]\} e^x$$

(2)

a) $\vec{m}_t = \beta_1 \vec{m}_{t-1} + (1-\beta_1)\vec{g}_t$ (1)

$s_t = \beta_2 s_{t-1} + (1-\beta_2)\vec{g}_t^2$ (2)

$\vec{\vartheta}_t = \vec{\vartheta}_{t-1} + \eta \frac{1}{\sqrt{s_t + \varepsilon}} \vec{m}_t$ (3)

(1) The space of variables is augmented by using momentum $\vec{m}_t$ which is updated directly by the gradient → suppresses oscillations.

(2) This term normalises the step size in (3) by taking into account the more recent gradients.

(3) The next best parameter is chose by combining (1) and (2) with the learning rate η.

b) $\vec{m}_1 = \beta_1 \vec{m}_0 + (1-\beta_1)\vec{g}_1$

$s_1 = \beta_2 s_0 + (1-\beta_2)\vec{g}_1^2$

$\vec{\vartheta}_1 = \vec{\vartheta}_0 + \eta \frac{\beta_1 \vec{m}_0 + (1-\beta_1)\vec{g}_1}{\sqrt{\beta_2 s_0 + (1-\beta_2)\vec{g}_1^2 + \varepsilon}}$

Assume $\vec{m}_0 = 0$, $\vec{s}_0 = 0$. ← Can this be assumed ?

$\Delta\vec{\vartheta} = \eta \frac{1-\beta_1}{\sqrt{(1-\beta_2) + \varepsilon \vec{g}_1^2}} \frac{\vec{g}_1}{|\vec{g}_1|} \overset{\varepsilon \text{ small}}{\approx} \underbrace{\eta \frac{1-\beta_1}{\sqrt{1-\beta_2}}}_{\text{const.}} \operatorname{sgn}(\vec{g}_1)$

c) This could be avoided by...

- choosing new initial values for $\vec{m}_0$, $\vec{s}_0$.
  - ↳ definining $\vec{m}_t$ so that $\vec{m}_1$ does not depend on any $\vec{m}_0$ but instead gets its value from other relevant variables.
- having a larger ε.

After some research, I found the bias correction term:

Corrected Adam equations (component-wise / vector form)

$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\,\mathbf{g}_t$	(1) 1st moment (momentum) — running average of gradients
$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\,\mathbf{g}_t^2$	(2) 2nd moment — running average of *squared* gradients (elementwise)
$\hat{\mathbf{m}}_t = \frac{\mathbf{m}_t}{1-\beta_1^t}$ $\quad \hat{\mathbf{v}}_t = \frac{\mathbf{v}_t}{1-\beta_2^t}$	(3) bias-correction to undo zero initialization bias
$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \varepsilon}$	(4) parameter update (note the minus sign)

d) If the weight decay was added to the loss it would get rescaled by $\frac{1}{\sqrt{s_t+\varepsilon}}$. This would result in inconsistent regularisation:

$$\vec{\vartheta}_t = \vec{\vartheta}_{t-1} + \eta \frac{1}{\sqrt{s_t+\varepsilon}} \frac{\partial}{\partial \vartheta} \ell(\vartheta) = \vec{\vartheta}_{t-1} - \eta \frac{1}{\sqrt{s_t+\varepsilon}} \left(\vec{m}_t + 2\lambda \|\vartheta\|_2^2\right)$$

The effect could be removed by applying the weight decay directly to the weights $\vec{\vartheta}_t$.

(3)

a) $p(y|x,\vartheta) = \mathcal{N}(y | f_\mu(x), \sigma^2)$

↑ evaluated at y

$$\Rightarrow \log\left(p(y|x,\vartheta)\right) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y-f_\mu(x))^2}{2\sigma^2}$$

$$= -\left[\frac{1}{2}\log(2\pi\sigma^2) + \frac{(y-f_\mu(x))^2}{2\sigma^2}\right]$$

Here we can identify $(y-f_\mu(x))^2$ as the MSE.

e) LL is derived in a), where the predicted $f_\mu(x)$, σ^2 can be inserted.

(4)

a) $\|T(x)-T(y)\| = \|F(x)+x-F(y)-y\| \overset{\Delta \text{ ineq.}}{\leq} \|F(x)-F(y)\| + \|x-y\| \leq 2\|x-y\|$

$\Rightarrow$ 2-LIPSCHITZ

b) For every further concatenated $T(x)$, a new factor 2 can be extracted using the triangle inequality.

e.g. $\|T_2(T_1(x)) - T_2(T_1(y))\| \leq 2\|T_1(x) - T_1(y)\| \leq 4\|x-y\|$

$\Rightarrow$ T_L is 2^L-LIPSCHITZ.

c)