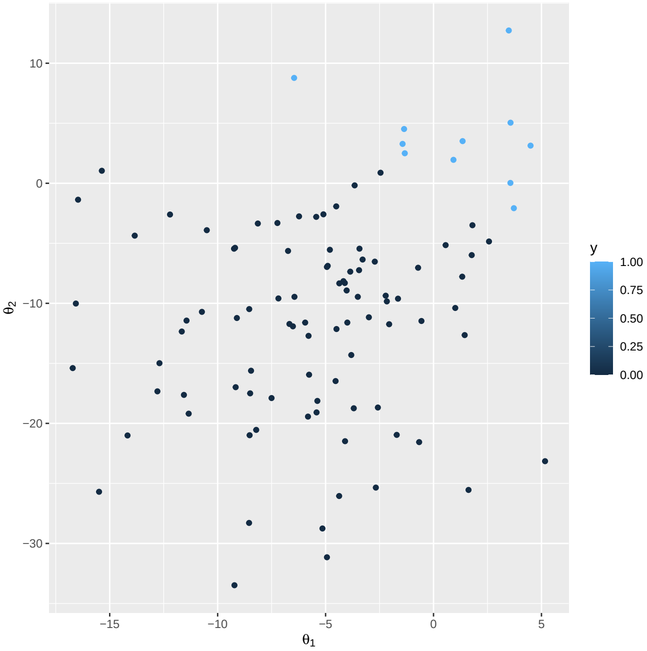
Optimization in machine learning **Exercise sheet 6**
https://slds-lmu.github.io/website_optimization/ **WS 2022/2023**

Multivariate Optimization 1

Exercise 1: Gradient Descent

You are given the following data situation:

```
library(ggplot2)

set.seed(314)
n <- 100
X = cbind(rnorm(n, -5, 5),
  rnorm(n, -10, 10))
X_design = cbind(1, X)

z <- 2*X[,1] + 3*X[,2]
pr <- 1/(1+exp(-z))
y <- as.integer(pr > 0.5)
df <- data.frame(X = X, y = y)

ggplot(df) +
  geom_point(aes(x = X.1, y = X.2, color=y)) +
  xlab(expression(theta[1])) +
  ylab(expression(theta[2]))
```



In the following we want to estimate a logistic regression without intercept via gradient descent[1].

(a) First consider the derivative of $g : \mathbb{R} \to \mathbb{R}, z \mapsto \log(1 + \exp(z)) - z$, i.e.,

$$g'(z) = \underbrace{\frac{\exp(z)}{1+\exp(z)}}_{<1} - 1 < 0 \Rightarrow g \text{ is monotonically decreasing} \Rightarrow g(z) > g(\alpha z) \quad \forall z > 0 \text{ and } \alpha > 0.$$

[1] We chose this algorithm for educational purposes; in practice, we typically use second order algorithms.

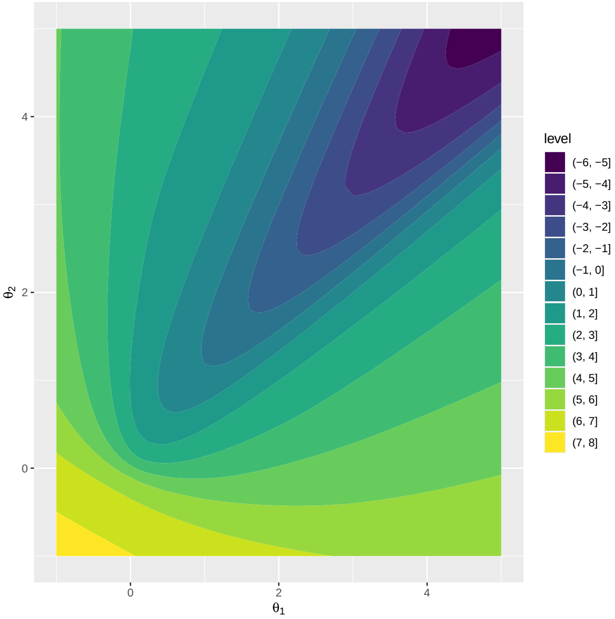
Second consider the derivative of $h : \mathbb{R} \to \mathbb{R}, z \mapsto \log(1 + \exp(-z))$, i.e.,

$$h'(z) = -\underbrace{\frac{\exp(-z)}{1 + \exp(-z)}}_{>0} < 0 \Rightarrow h \text{ is monotonically decreasing} \Rightarrow h(z) > h(\alpha z) \quad \forall z > 0 \text{ and } \alpha > 0.$$

With this we get for $\alpha > 0$
$\mathcal{R}_{\text{emp}}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^n \log(1 + \exp(\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)})) - y^{(i)} \tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)} =$
$\sum_{i=1}^n \mathbb{1}_{y^{(i)}=1}(\log(1 + \exp(|\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}|)) - |\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}|) + \mathbb{1}_{y^{(i)}=0}(\log(1 + \exp(-|\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}|)) >$
$\sum_{i=1}^n \mathbb{1}_{y^{(i)}=1}(\log(1 + \exp(\alpha|\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}|)) - \alpha|\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}|) + \mathbb{1}_{y^{(i)}=0}(\log(1 + \exp(-\alpha|\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)}|)) =$
$\sum_{i=1}^n \log(1 + \exp(\alpha\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)})) - y^{(i)} \alpha\tilde{\boldsymbol{\theta}}^\top \mathbf{x}^{(i)} =$
$\mathcal{R}_{\text{emp}}(\alpha\tilde{\boldsymbol{\theta}})$ since $\tilde{\boldsymbol{\theta}}$ perfectly seperates the data.

(b)

```
lambda = 0

f <- function(theta, lambda) lambda * theta %*% theta +
  sum(-y * X %*% theta + log(1 + exp(X %*% theta)))

x = seq(-1, 5, by=0.1)
xx = expand.grid(X1 = x, X2 = x)

fxx = log(apply(xx, 1, function(t) f(t, lambda)))
df = data.frame(xx = xx, fxx = fxx)

ggplot() +
    geom_contour_filled(data = df, aes(x = xx.X1, y = xx.X2, z = fxx)) +
    xlab(expression(theta[1])) +
    ylab(expression(theta[2]))
```

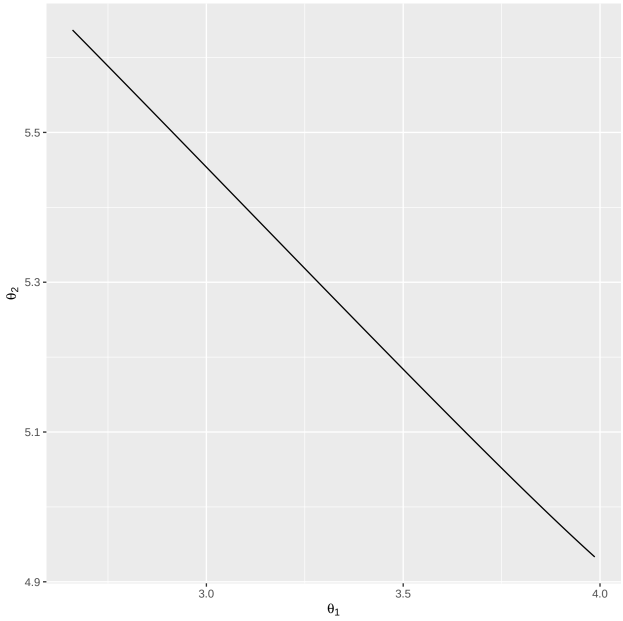


(c) $\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \sum_{i=1}^n \frac{\exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})}{1+\exp(\boldsymbol{\theta}^\top \mathbf{x}^{(i)})} \mathbf{x}^{(i)^\top} - y^{(i)} \mathbf{x}^{(i)^\top}$

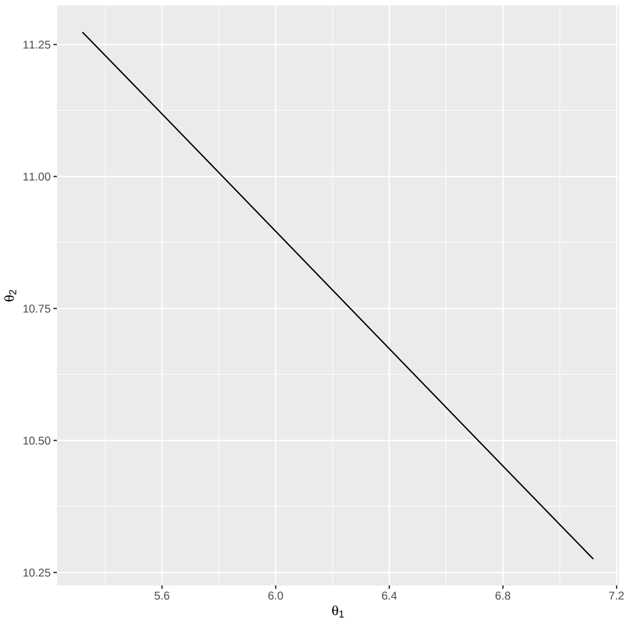
(d)

```
df_t <- function(theta, lambda) lambda * t(theta) -(t(y) %*% X) +
  t(1/(1 + exp(-X %*% theta))) %*% X

gd_step <- function(theta, alpha, lambda) return(theta - alpha * df_t(theta, lambda)[1,])
## Alpha = 0.5
theta = c(0,0)
alpha = 0.01
```

```
thetas = NULL
for(i in 1:500){
  theta = gd_step(theta, alpha, lambda)
  thetas = rbind(thetas, theta)
}

ggplot(as.data.frame(thetas), aes(x=V1, y=V2)) +
  geom_line() +
  xlab(expression(theta[1])) +
  ylab(expression(theta[2]))
```



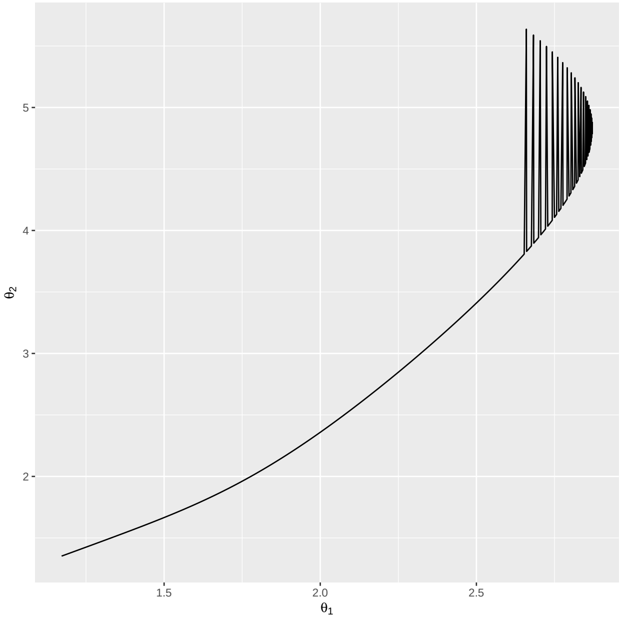
```
## Alpha = 2
theta = c(0,0)
alpha = 0.02

thetas = NULL
for(i in 1:500){
  theta = gd_step(theta, alpha, lambda)
  thetas = rbind(thetas, theta)
}

ggplot(as.data.frame(thetas), aes(x=V1, y=V2)) +
  geom_line() +
  xlab(expression(theta[1])) +
  ylab(expression(theta[2]))
```



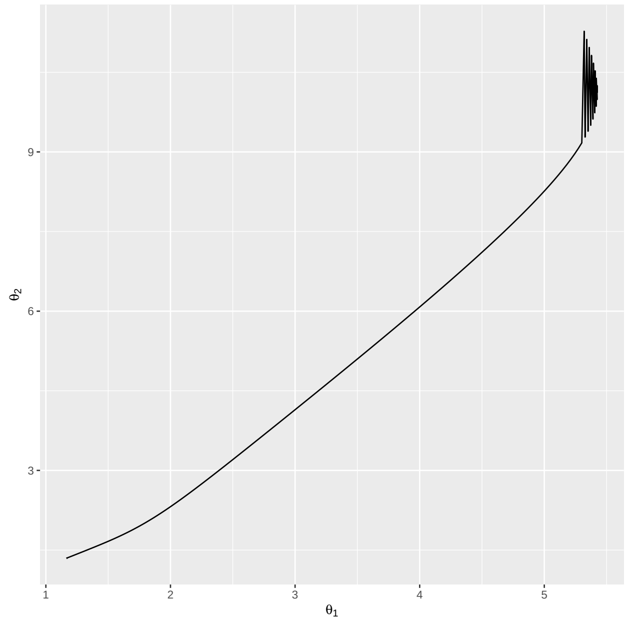
Gradient descent will in theory not converge since $\mathcal{R}_{\text{emp}}$ has no minimum (a))

(e)

```
## Lambda = 0.5, alpha = 0.5
theta = c(0,0)
alpha = 0.01

thetas = NULL
for(i in 1:500){
  theta = gd_step(theta, alpha, 0.5)
  thetas = rbind(thetas, theta)
}

ggplot(as.data.frame(thetas), aes(x=V1, y=V2)) +
  geom_line() +
  xlab(expression(theta[1])) +
  ylab(expression(theta[2]))
```

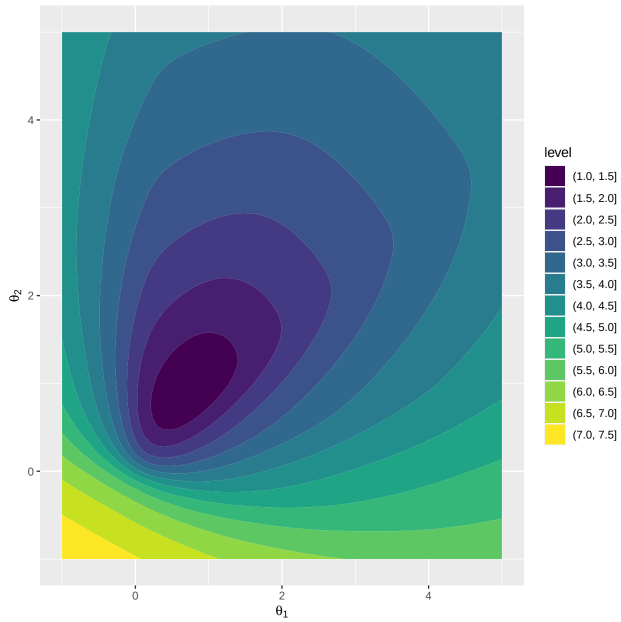


```
theta = c(0,0)
alpha = 0.02

thetas = NULL
for(i in 1:500){
  theta = gd_step(theta, alpha, 0.5)
  thetas = rbind(thetas, theta)
}
## Lambda = 0.5, alpha = 0.02
theta = c(0,0)
alpha = 0.02

thetas = NULL
for(i in 1:500){
  theta = gd_step(theta, alpha, 0.5)
  thetas = rbind(thetas, theta)
}

ggplot(as.data.frame(thetas), aes(x=V1, y=V2)) +
  geom_line() +
  xlab(expression(theta[1])) +
  ylab(expression(theta[2]))
```



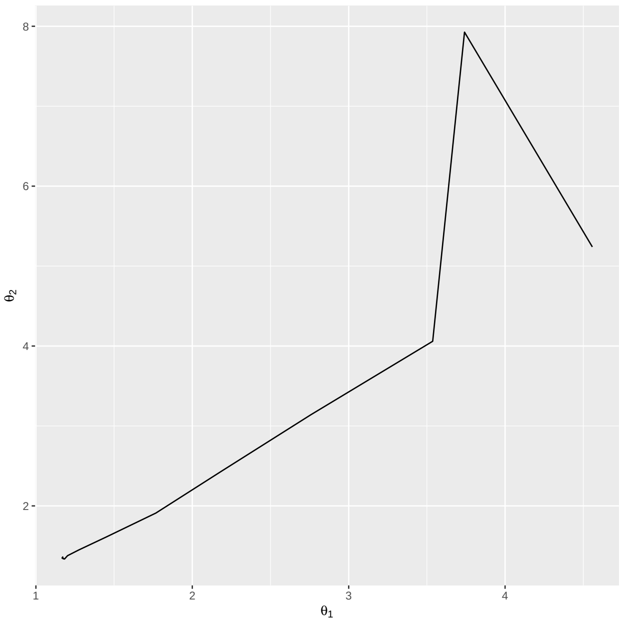
(f)

```
lambda = 1

fxx_reg = log(apply(xx, 1, function(t) f(t, lambda)))
df_reg = data.frame(xx = xx, fxx = fxx_reg)

ggplot() +
    geom_contour_filled(data = df_reg, aes(x = xx.X1, y = xx.X2, z = fxx)) +
    xlab(expression(theta[1])) +
    ylab(expression(theta[2]))
```

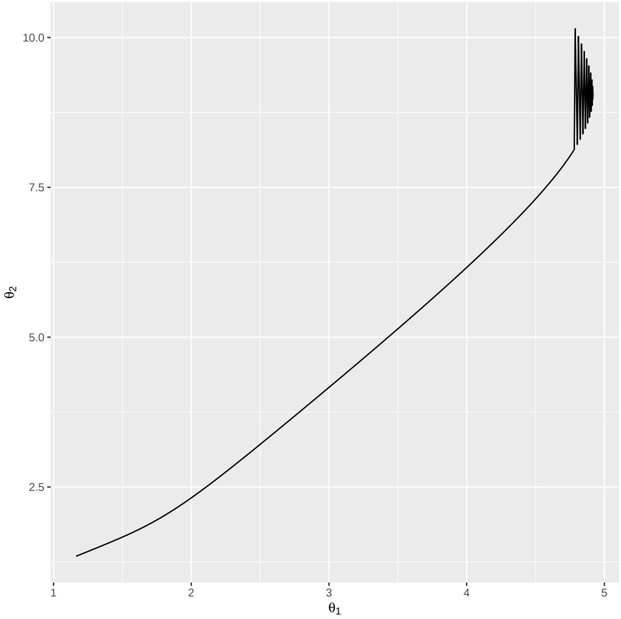


(g)

```r
gd_backtracking_step <- function(theta, alpha, gamma, tau, lambda){
    ftheta = f(theta, lambda)
    for(i in 1:1000){
      theta_prop = theta - alpha * gamma * df_t(theta, lambda)[1,]
      if(f(theta_prop, lambda) < ftheta){
        return(theta_prop)
      }else{
        alpha = tau * alpha
      }
    }
    return(theta)
}

## Lambda = 0.5, alpha = 0.5

theta = c(0,0)
alpha = 0.5

thetas = NULL
for(i in 1:500){
  theta = gd_backtracking_step(theta, alpha, 0.9, 0.5, 0.5)
  thetas = rbind(thetas, theta)
}

ggplot(as.data.frame(thetas), aes(x=V1, y=V2)) +
  geom_line() +
  xlab(expression(theta[1])) +
  ylab(expression(theta[2]))
```



```r
## Lambda = 0.5, alpha = 0.02

theta = c(0,0)
alpha = 0.02

thetas = NULL
for(i in 1:500){
  theta = gd_backtracking_step(theta, alpha, 0.9, 0.5, 0.5)
  thetas = rbind(thetas, theta)
}

ggplot(as.data.frame(thetas), aes(x=V1, y=V2)) +
  geom_line() +
  xlab(expression(theta[1])) +
  ylab(expression(theta[2]))
```


θ_2
θ_1