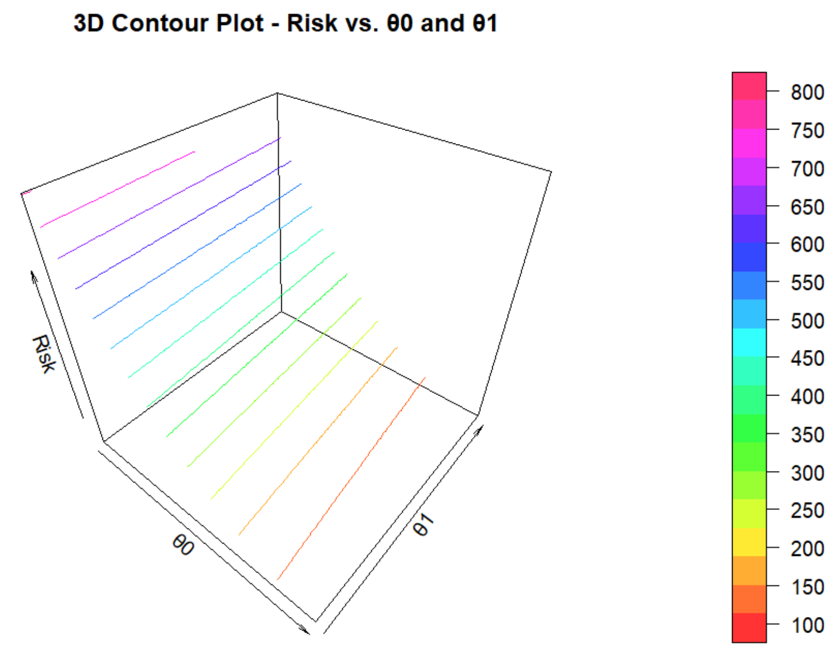
Technical University of Dortmund Summer Semester 2023
Data Science Exercise Sheet 05
Group member name(s): Ahmed Arian Sajid, Sabrina Sultana, Sharmin Ahmad
Group member UID(s): 235061, 235062, 230239

Advanced Statistical Learning

Answer to Exercise 1.a

Visualize the risk as a function from (θ_0, θ_1), e.g. using a contour plot:



Answer to Exercise 1.b

The iterative update rule of Gradient Descent:

$$\theta^{[j+1]} = \theta^{[j]} - \alpha^{[j]}.\nabla_\theta R_{emp}(\theta)$$

The update step of the gradient descent algorithm for our model class with the L2-loss:

$$\theta_1^{[j+1]} = \theta_1^{[j]} + \alpha^{[j]}\frac{1}{n}\sum_{i=1}^{n}(2(y_i - (\theta_0 + \theta_1 x_i)x_i)$$

$$\theta_0^{[j+1]} = \theta_0^{[j]} + \alpha^{[j]}\frac{1}{n}\sum_{i=1}^{n}(2(y_i - (\theta_0 + \theta_1 x_i))$$

where the derivatives come from L2 loss function with respect to θ_1 and θ_0 and α is the learning late.

Answer to Exercise 1.c

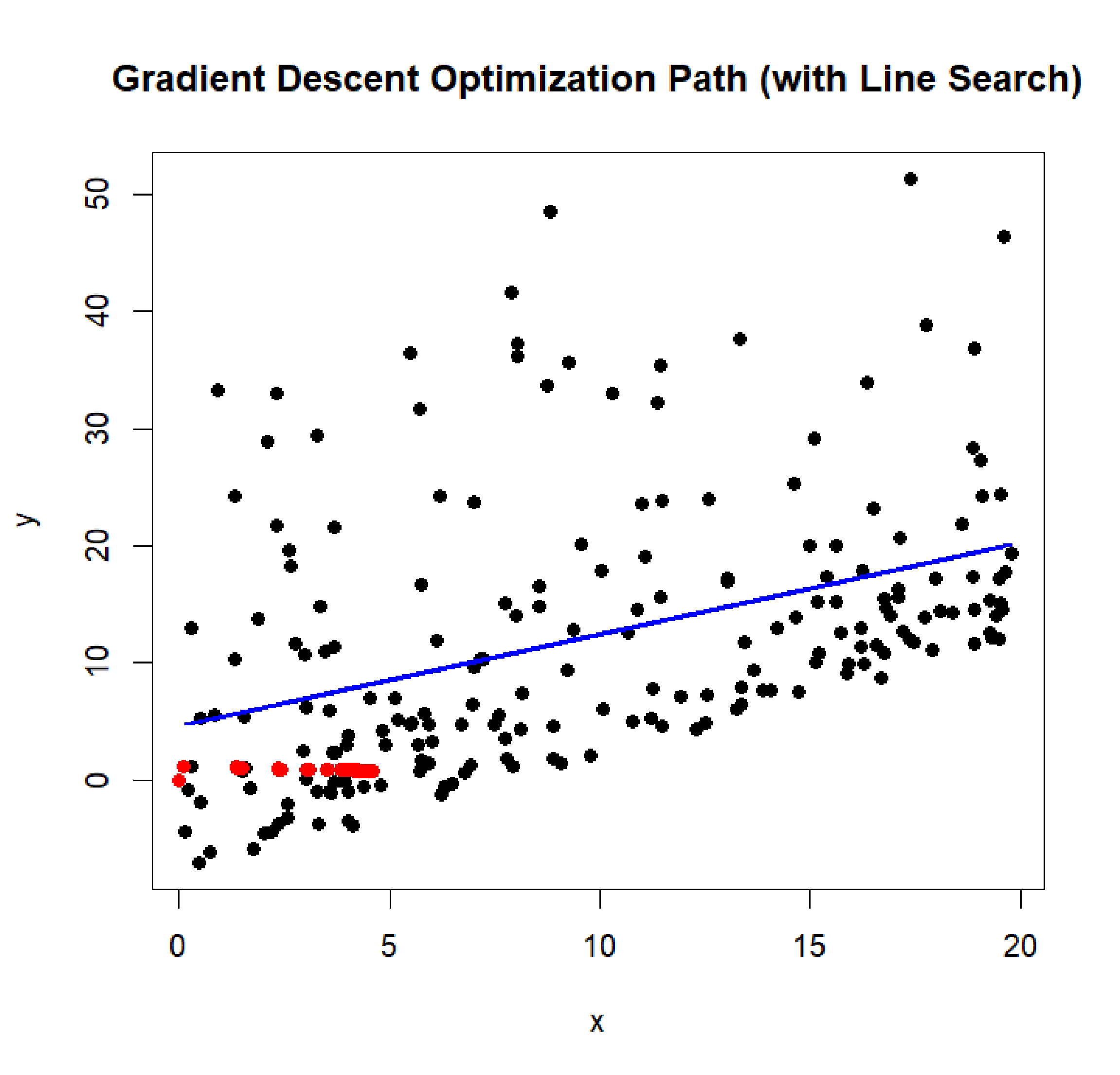
The gradient descent algorithm is implemented in R:

```
data <- read.csv("fitting2.csv")
l2_loss <- function(theta0, theta1) {
n <- nrow(data)
sum((data$y - (theta0 + theta1 * data$x))^2) / n
}
gradient <- function(theta0, theta1) {
n <- nrow(data)
gradient_theta0 <- (-2/n) * sum(data$y - (theta0 + theta1 * data$x))
gradient_theta1 <- (-2/n) * sum((data$y - (theta0 + theta1 * data$x)) * data$x)
c(gradient_theta0, gradient_theta1)
}
line_search <- function(theta0, theta1, gradient, loss, initial_alpha, rho, c) {
alpha <- initial_alpha
grad <- gradient(theta0, theta1)
loss_current <- loss(theta0, theta1)
direction <- -grad
while (loss(theta0 + alpha * direction[1], theta1 + alpha * direction[2]) >
loss_current + c * alpha * sum(grad * direction)) {
alpha <- rho * alpha
}
print(alpha)
}
gradient_descent_line_search <- function(initial_alpha, rho, c, num_iterations) {
theta0 <- 0  # Initial values for theta0
theta1 <- 0  # Initial values for theta1
path <- matrix(c(theta0, theta1), ncol = 2)

for (i in 1:num_iterations) {
grad <- gradient(theta0, theta1)
alpha <- line_search(theta0, theta1, gradient, l2_loss, initial_alpha, rho, c)
theta0 <- theta0 - alpha * grad[1]
theta1 <- theta1 - alpha * grad[2]
path <- rbind(path, c(theta0, theta1))
}
list(theta0 = theta0, theta1 = theta1, path = path)
}
initial_alpha <- 1  # Initial learning rate
rho <- 0.5  # Reduction factor for learning rate
c <- 0.1  # Sufficient decrease parameter
num_iterations <- 1000
result <- gradient_descent_line_search(initial_alpha, rho, c, num_iterations)
optimal_theta0 <- result$theta0
optimal_theta1 <- result$theta1
path <- result$path

# Visualize the optimization path
plot(data$x, data$y, pch = 16, col = "black", xlab = "x", ylab = "y", main
= "Gradient Descent Optimization Path (with Line Search)")
lines(data$x, optimal_theta0 + optimal_theta1 * data$x, col = "blue", lwd = 2)
points(path[, 1], path[, 2], col = "red", pch = 16)
```

We ran the iteration 1000 times and find the optimal values of θ_0 and θ_1) are:

```
> optimal_theta0
[1] 4.614241
> optimal_theta1
[1] 0.7822008
```

Visualizing the optimization path:



Each point in the optimization path is marked with a red dot and the color blue is used to symbolize the fitted line with the optimal values of θ_0 and θ_1.

Answer to Exercise 2

1

Exercise 2: LOSS Function : Quantile loss

Q. Show that the constant $f(x)=c$, that optimizes the quantile loss

$$L(y; f(x)) = \begin{cases} (1-\alpha)(f(x)-y) & \text{if } y<c \\ \alpha(y-f(x)) & \text{if } y \geq c \end{cases}$$

given a specific $\alpha \in (0,1)$, is the α-quantile.

Sol^n: $y = (y^{(1)}, y^{(2)}, \ldots, y^{(n)})$

$$L(y, f(x)) = \begin{cases} (1-\alpha)(f(x)-y) & \text{if } y<c \\ \alpha(y-f(x)) & \text{if } y \geq c \end{cases}$$

$$R_{emp}(f) = \sum_{i=1}^{n} L(y^{(i)}, f(x^{(i)}))$$

set $f(x)=c$, then,

$$\hat{c} = q_\alpha(y) = \begin{cases} \in [y(n\alpha), y(n\alpha+1)] & \text{if } n\alpha \text{ is integer} \\ y([n\alpha+1]) & \text{if otherwise} \end{cases}$$

with e.g. $\hat{c} = 0.5(y_{(n\alpha)} + y_{(n\alpha+1)})$ if $n\alpha$ is integer, minimizes the empirical risk.

We show: (i) $R_{emp}(f)$ is convex

(ii) That q_α minimizes R_{emp} for $n\alpha$ is not an integer.

(iii) That q_α minimizes R_{emp} for $n\alpha$ is an integer.

2

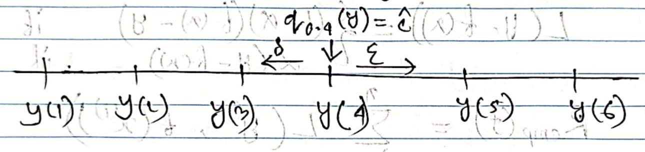
It holds:

(i) Since $L(y, f(x))$ is convex, $R_{emp}(f)$ as sum of convex functions is also convex.

(II) We now assume that $n\alpha$ is no integer, such that $n = 6$ & $\alpha = 0.4$

$$n\alpha = 2.4$$

The idea for proof is to assume that q_α is the minimum and now we calculate the risk for another candidate $q_\alpha + \varepsilon$ or $q_\alpha - \delta$ and show that this risk is higher as for q_α:



without loss of generality set

$$y(\lfloor n\alpha + 1 \rfloor + 1) - y(\lfloor n\alpha + 1 \rfloor) > \varepsilon > 0 :$$

③

$$R_{emp}(\hat{c}+\varepsilon) = (1-\alpha)\sum_{y^{(i)}<\hat{c}+\varepsilon}\left(\hat{c}+\varepsilon-y^{(i)}\right) + \alpha\sum_{y^{(i)}\geq\hat{c}+\varepsilon}\left(y^{(i)}-(\hat{c}+\varepsilon)\right)$$

$$= (1-\alpha)\left[\underbrace{\sum_{y^{(i)}<\hat{c}+\varepsilon}\left(\hat{c}-y^{(i)}\right)}_{=1} + \sum_{y^{(i)}<\hat{c}+\varepsilon}\varepsilon\right] +$$

$$\alpha\left[\underbrace{\sum_{y^{(i)}\geq\hat{c}+\varepsilon}\left(y^{(i)}-\hat{c}\right)}_{=2} - \sum_{y^{(i)}\geq\hat{c}+\varepsilon}\varepsilon\right] \quad \text{--- } ①$$

$$1 = \sum_{y^{(i)}<\hat{c}+\varepsilon}\left(\hat{c}-y^{(i)}\right) = \sum_{y^{(i)}<\hat{c}}\left(\hat{c}-y^{(i)}\right) + \underbrace{\left(\hat{c}-y_{(\lfloor n\alpha+1\rfloor)}\right)}_{=0,\ \hat{c}=q_\alpha=y_{(\lfloor n\alpha+1\rfloor)}}$$

$$= \sum_{y^{(i)}<\hat{c}}\left(\hat{c}-y^{(i)}\right)$$

$$2 = \sum_{y^{(i)}\geq\hat{c}+\varepsilon}\left(y^{(i)}-\hat{c}\right) = \sum_{y^{(i)}\geq\hat{c}}\left(y^{(i)}-\hat{c}\right) - \underbrace{\left(y_{(\lfloor n\alpha+1\rfloor)}-\hat{c}\right)}_{=0,\ \hat{c}=q_\alpha=y_{(\lfloor n\alpha+1\rfloor)}}$$

$$= \sum_{y^{(i)}\geq\hat{c}}\left(y^{(i)}-\hat{c}\right)$$

from ①.

$$R_{emp}(\hat{c}+\varepsilon) = (1-\alpha)\left[\underbrace{\sum_{y^{(i)}<\hat{c}}\left(\hat{c}-y^{(i)}\right)}_{=(1)} + \sum_{y^{(i)}<\hat{c}+\varepsilon}\varepsilon\right] + \alpha\left[\underbrace{\sum_{y^{(i)}\geq\hat{c}}\left(y^{(i)}-\hat{c}\right)}_{=2} + \sum_{y^{(i)}\geq\hat{c}+\varepsilon}\varepsilon\right]$$

$$= \underbrace{(1-\alpha)\sum_{y^{(i)}<\hat{c}}\left(\hat{c}-y^{(i)}\right) + \alpha\sum_{y^{(i)}\geq\hat{c}}\left(y^{(i)}-\hat{c}\right)}_{R_{emp}(\hat{c})} +$$

$$\underbrace{(1-\alpha)\sum_{y^{(i)}<\hat{c}+\varepsilon}\varepsilon - \alpha\sum_{y^{(i)}\geq\hat{c}+\varepsilon}\varepsilon}_{\Delta R}$$

④

$$\Delta R = (1-\alpha)\,\varepsilon \sum_{y^{(i)} < \hat{c}+\varepsilon} 1 - \alpha\varepsilon \sum_{y^{(i)} \geq \hat{c}+\varepsilon} 1$$

$$= \varepsilon\left(\sum_{y^{(i)} < \hat{c}+\varepsilon} 1 - \alpha \sum_{y^{(i)} < \hat{c}-\varepsilon} 1 - \alpha \sum_{y^{(i)} \geq \hat{c}+\varepsilon} 1\right)$$

$$= \varepsilon\left(\sum_{y^{(i)} < \hat{c}+\varepsilon} 1 - n\alpha\right)$$

$$= \varepsilon\left(\lfloor n\alpha + 1 \rfloor - n\alpha\right) > 0$$

$$\Rightarrow R_{emp}(\hat{c}+\varepsilon) = R_{emp}(\hat{c}) + \underbrace{\Delta R}_{>0} > R_{emp}(\hat{c})$$

proceed analogously for δ with $y(\lfloor n\alpha+1\rfloor) - y(\lfloor n\alpha+1\rfloor)-1) > \delta > 0$

and $R_{emp}(\hat{c}-\hat{\delta}) > \Delta R + R_{emp}(\hat{c})$.

(III) We now assume that $n\alpha$ is an integer.

Hence, we have $\hat{c} = q_\alpha \in \left[y_{(n\alpha)}, y_{(n\alpha+1)}\right]$. Here, we compare two cases:

※ $\varepsilon \leq y_{(n\alpha+1)} - \hat{c}$ and $\delta \leq \hat{c} - y_{(n\alpha)}$:

Here, we can calculate ΔR very easily since $n\alpha$ y values are left of $\hat{c}$ and $n - n\alpha$ values right of $\hat{c}$.

therefore:

$$\Delta R = n\alpha(1-\alpha)\,\varepsilon - (n-n\alpha)\,\alpha\varepsilon$$

$$= n\alpha\varepsilon - n\alpha^2\varepsilon - n\alpha\varepsilon + n\alpha^2\varepsilon$$

$$= 0$$

Proceed analogously for δ.

This means that all points in $\left[y_{(n\alpha)}, y_{(n\alpha+1)}\right]$ give the minimal empirical risk.

5

* $\varepsilon > y_{(n\alpha+1)} - \hat{c}$ and $\delta > \hat{c} - y_{(n\alpha)}$:

same as for (II),

With (II) and (III) we have, shown that the α-quantile gives the local minimum since other values $\boxed{q_\alpha - \delta, q_\alpha + \varepsilon}$ have a higher empirical risk. This optima gives us also the global minimum, since our objective function is a convex function (i).