Homework 1
Statistical Learning for Decision Making 2023

Li Ju
li.ju@it.uu.se

January 31, 2023

1 Problem 1

a). The risk of θ is derived as follows;

$$\begin{aligned}
L(\boldsymbol{\theta}, p) &:= \mathbb{E}[\ell_\theta(\mathbf{z})] \\
&= \int \ell_\theta(\mathbf{z}) p(\mathbf{z}) d\mathbf{z} = \int \|\mathbf{z} - \boldsymbol{\theta}\|^2 p(\mathbf{z}) d\mathbf{z} \\
&= \int (\mathbf{z}^\top \mathbf{z} - 2\boldsymbol{\theta}^\top \mathbf{z} + \boldsymbol{\theta}^\top \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z} \\
&= \int \mathbf{z}^\top \mathbf{z} p(\mathbf{z}) d\mathbf{z} - 2\boldsymbol{\theta}^\top \int \mathbf{z} p(\mathbf{z}) d\mathbf{z} + \boldsymbol{\theta}^\top \boldsymbol{\theta} \int p(\mathbf{z}) d\mathbf{z} \\
&= \mathbb{E}[\mathbf{z}^\top \mathbf{z}] - 2\boldsymbol{\theta}^\top \mathbb{E}[\mathbf{z}] + \boldsymbol{\theta}^\top \boldsymbol{\theta}
\end{aligned}$$

The target parameter is defined as $\boldsymbol{\theta}_\circ(p) := \arg_{\boldsymbol{\theta}} \min L(\boldsymbol{\theta}, p)$, which is given when $\nabla L(\boldsymbol{\theta}) = \mathbf{0}$ as follows:

$$\begin{aligned}
\nabla L(\boldsymbol{\theta}) = 2\boldsymbol{\theta}_\circ - 2\mathbb{E}[\mathbf{z}] &= \mathbf{0} \\
\boldsymbol{\theta}_\circ &= \mathbb{E}[\mathbf{z}]
\end{aligned}$$

b). The contour plot for the risk function $L(\boldsymbol{\theta})$ is shown as Figure 1, with the constant set as 0. The code for plotting the figure is appended in Appendix A.1

c).

$$\begin{aligned}
L(\boldsymbol{\theta}, p) &= \mathbb{E}[\mathbf{z}^\top \mathbf{z}] - 2\boldsymbol{\theta}^\top \mathbb{E}[\mathbf{z}] + \boldsymbol{\theta}^\top \boldsymbol{\theta} \\
&= \mathbb{V}[\mathbf{z}] + \mathbb{E}[\mathbf{z}]^2 - 2\boldsymbol{\theta}^\top \mathbb{E}[\mathbf{z}] + \boldsymbol{\theta}^\top \boldsymbol{\theta}
\end{aligned}$$

For a multivariate t-distribution $t_\nu(\mathbf{z}; \boldsymbol{\mu}, \mathbf{C})$, $\mathbb{E}(\mathbf{z}) = \boldsymbol{\mu}$ for $\nu > 1$, $\mathbb{V}[\mathbf{z}] = \frac{\nu}{\nu-2}\mathbf{C}$ for $\nu > 2$, otherwise the first and second order moment are not defined. For $\nu \in \{1, 2, 3\}$, the risk function is given as follows:

$$L(\boldsymbol{\theta}, p) = \begin{cases} 3\mathbf{C} + \boldsymbol{\mu}^2 - 2\boldsymbol{\theta}^\top \boldsymbol{\mu} + \boldsymbol{\theta}^\top \boldsymbol{\theta}, & \text{if } \nu = 3 \\ \mathbb{V}[\mathbf{z}](\text{undefined}) + \boldsymbol{\mu}^2 - 2\boldsymbol{\theta}^\top \boldsymbol{\mu} + \boldsymbol{\theta}^\top \boldsymbol{\theta}, & \text{if } \nu = 2 \\ \mathbb{V}[\mathbf{z}](\text{undefined}) + \mathbb{E}[\mathbf{z}]^2(\text{undefined}) - 2\boldsymbol{\theta}^\top \mathbb{E}[\mathbf{z}](\text{undefined}) + \boldsymbol{\theta}^\top \boldsymbol{\theta}, & \text{if } \nu = 1 \end{cases}$$

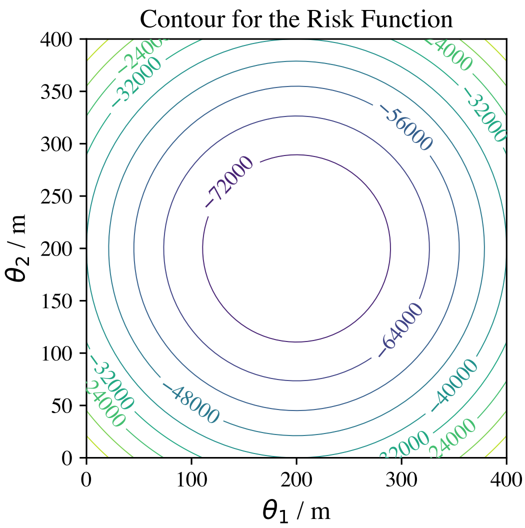

Figure 1: Contour for the Risk function

d). The empirical risk $\hat{L}(\boldsymbol{\theta})$ is given as follows if we approximate $p_{\mathbf{z}}$ with the empirical distribution p_n:

$$\begin{aligned}\hat{L}(\boldsymbol{\theta}) &= \mathbb{E}_n[\ell_\theta(\mathbf{z})] \\ &= \sum_{i=1}^{n} P_n(\mathbf{z}_i)\|\mathbf{z}_i - \boldsymbol{\theta}\|^2 \\ &= \sum_{i=1}^{n} \frac{1}{n}\|\mathbf{z}_i - \boldsymbol{\theta}\|^2\end{aligned}$$

The empirical minimizer is defined as $\hat{\boldsymbol{\theta}}_n := \arg_{\boldsymbol{\theta}} \min \hat{L}(\boldsymbol{\theta})$, which is given when $\nabla\hat{L}(\boldsymbol{\theta}) = \mathbf{0}$ as follows:

$$\begin{aligned}\nabla\hat{L}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}) &= 2\hat{\boldsymbol{\theta}}_n - 2\mathbb{E}_n[\mathbf{z}] = \mathbf{0} \\ \hat{\boldsymbol{\theta}}_n &= \mathbb{E}_n[\mathbf{z}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\end{aligned}$$

e). The empirical risk $\hat{L}(\boldsymbol{\theta})$ can be reformulated as following:

$$\begin{aligned}\hat{L}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \frac{1}{n}\|\mathbf{z}_i - \boldsymbol{\theta}\|^2 \\ &= \sum_{i=1}^{n} \frac{1}{n}\mathbf{z}_i^\top \mathbf{z}_i - 2\boldsymbol{\theta}^\top \sum_{i=1}^{n}\frac{1}{n}\mathbf{z}_i + \boldsymbol{\theta}^\top\boldsymbol{\theta}\end{aligned}$$

With different numbers of samples $\mathbf{z} \sim \mathcal{N}(\mathbf{z}; \left[\begin{smallmatrix}20\\20\end{smallmatrix}\right], \left[\begin{smallmatrix}400 & 50\\50 & 400\end{smallmatrix}\right])$, the contours of the resulting empirical risks are plotted in Figure 2 as follows with codes appended in Appendix A.2.

f). The gradient descent algorithm is implemented to minimizing the empirical Risk function with 1000 samples. A fixed learning rate $\eta = 0.01$ is chosen and 500 steps are applied from an initial point $\hat{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$. The updates of gradient descent are plotted in Figure 3 and the code is deferred in Appendix A.3.

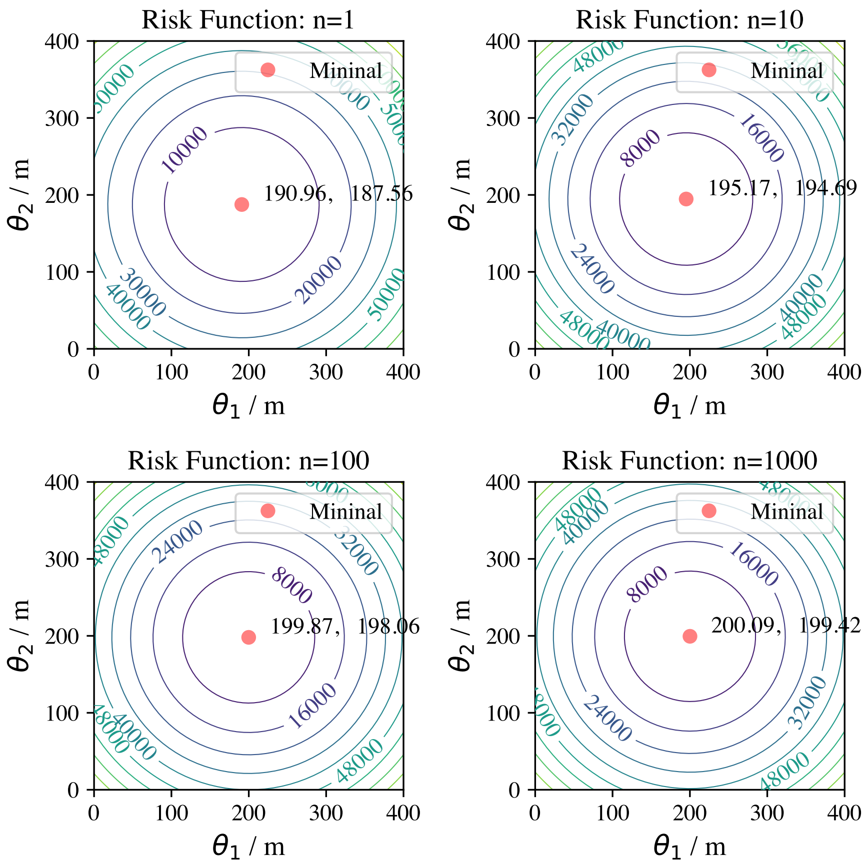

Figure 2: Contour for the empirical Risk function

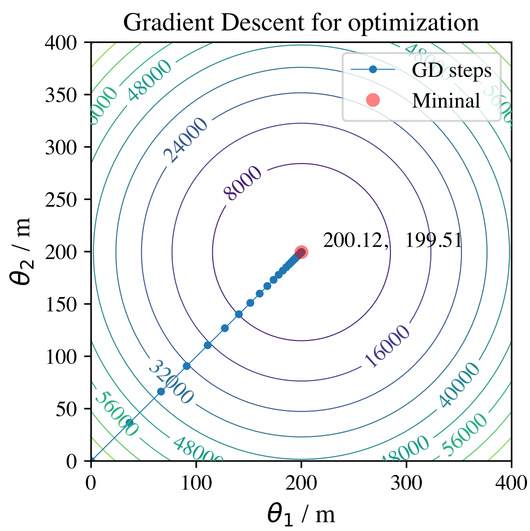

Figure 3: Gradient descent to minimizing the empirical Risk function: $n = 1000$

g). Let $\mathbf{T}_n \coloneqq \sum_{i=1}^{n} \mathbf{z}_i = n\hat{\boldsymbol{\theta}}_n$, it can be derived that

$$
\begin{aligned}
\mathbb{E}[\mathbf{T}_n] &= \mathbb{E}[\sum_{i=1}^{n} \mathbf{z}_i] = \sum_{i=1}^{n} \mathbb{E}[\mathbf{z}_i] = n\boldsymbol{\theta}_\circ \\
\mathbb{V}[\mathbf{T}_n] &= \mathbb{V}[\sum_{i=1}^{n} \mathbf{z}_i] = \sum_{i=1}^{n} \mathbb{V}[\mathbf{z}_i] = n\boldsymbol{\Sigma} \\
\mathbb{E}[\hat{\boldsymbol{\theta}}_n] &= \mathbb{E}[\frac{\mathbf{T}_n}{n}] = \frac{\mathbb{E}[\mathbf{T}_n]}{n} = \boldsymbol{\theta}_\circ \\
\mathbb{V}[\hat{\boldsymbol{\theta}}_n] &= \mathbb{V}[\frac{\mathbf{T}_n}{n}] = \frac{\mathbb{V}[\mathbf{T}_n]}{n^2} = \frac{\boldsymbol{\Sigma}}{n}
\end{aligned}
$$

Then the MSE-matrix of the learned parameter $\hat{\boldsymbol{\theta}}_n$ is derived as follows:

$$
\begin{aligned}
\mathbf{M} &= \mathbb{E}[(\boldsymbol{\theta}_\circ - \hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}_\circ - \hat{\boldsymbol{\theta}}_n)^\top] \\
&= \mathbb{E}[(\mathbb{E}[\hat{\boldsymbol{\theta}}_n] - \hat{\boldsymbol{\theta}}_n)(\mathbb{E}[\hat{\boldsymbol{\theta}}_n] - \hat{\boldsymbol{\theta}}_n)^\top] \\
&= \mathbb{V}[\hat{\boldsymbol{\theta}}_n] = \frac{\boldsymbol{\Sigma}}{n}
\end{aligned}
$$

h). We know that $\nabla \hat{L}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \hat{L}(\boldsymbol{\theta})}{\partial \theta_{(1)}} \\ \frac{\partial \hat{L}(\boldsymbol{\theta})}{\partial \theta_{(2)}} \end{bmatrix} = 2\begin{bmatrix} \theta_{(1)} - \mathbb{E}_n[z_{(1)}] \\ \theta_{(2)} - \mathbb{E}_n[z_{(2)}] \end{bmatrix} = 2\boldsymbol{\theta} - 2\mathbb{E}_n[\mathbf{z}]$, thus we have the Hessian of $\hat{L}(\boldsymbol{\theta})$ as follows:

$$
\begin{aligned}
\nabla^2 \hat{L}(\boldsymbol{\theta}) &= \begin{bmatrix} \frac{\partial^2 \hat{L}(\boldsymbol{\theta})}{\partial \theta_{(1)}^2} & \frac{\partial^2 \hat{L}(\boldsymbol{\theta})}{\partial \theta_{(1)} \partial \theta_{(2)}} \\ \frac{\partial^2 \hat{L}(\boldsymbol{\theta})}{\partial \theta_{(2)} \partial \theta_{(1)}} & \frac{\partial^2 \hat{L}(\boldsymbol{\theta})}{\partial \theta_{(2)}^2} \end{bmatrix} \\
&= 2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 2\mathbf{I}_2
\end{aligned}
$$

The sensitivity score is defined as $\dot{\ell}_\circ(\mathbf{z}) \coloneqq \partial_{\boldsymbol{\theta}} \ell(\mathbf{z})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\circ}$. With $\ell_{\boldsymbol{\theta}(\mathbf{z})} = \|\mathbf{z} - \boldsymbol{\theta}\|^2$, we have

$$
\begin{aligned}
\dot{\ell}_\circ(\mathbf{z}) &= -2(\mathbf{z} - \boldsymbol{\theta}_\circ) = 2(\boldsymbol{\theta}_\circ - \mathbf{z}) \\
\mathbf{L} &= \mathbb{V}[\dot{\ell}_\circ(\mathbf{z})] = 4\mathbb{V}[\boldsymbol{\theta}_\circ - \mathbf{z}] \\
&= 4\mathbb{V}[\mathbf{z}] = 4\boldsymbol{\Sigma}
\end{aligned}
$$

The as-derived covariance $\frac{\boldsymbol{\Sigma}}{n}$ of $\hat{\boldsymbol{\theta}}_n$ is identical with the MSE-matrix of $\hat{\boldsymbol{\theta}}_n$ in Problem 1g, as $\hat{\boldsymbol{\theta}}_n$ is a unbiased estimation for $\boldsymbol{\theta}_\circ$.

2 Problem 2

a). With the squared-error loss function $\ell_\theta(\mathbf{z}) = (y - \mathbf{x}^\top \boldsymbol{\theta})^2$, the risk function is given by:

$$
\begin{aligned}
L(\boldsymbol{\theta}, p) &= \mathbb{E}[\ell_\theta(\mathbf{z})] = \mathbb{E}[(y - \mathbf{x}^\top \boldsymbol{\theta})^2] \\
&= \mathbb{E}[y^2 - 2y\boldsymbol{\theta}^\top \mathbf{x} + \mathbf{x}^\top \boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbf{x}] \\
&= \mathbb{E}[y^2] - 2\boldsymbol{\theta}^\top \mathbb{E}[y\mathbf{x}] + \boldsymbol{\theta}\boldsymbol{\theta}^\top \mathbb{E}[\mathbf{x}\mathbf{x}^\top]
\end{aligned}
$$

The target parameter $\boldsymbol{\Theta}_\circ \coloneqq \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, p)$, which is given when $\nabla L(\boldsymbol{\Theta}_\circ, p) = \mathbf{0}$ and we have

$$
\begin{aligned}
\nabla L(\boldsymbol{\Theta}_\circ, p) = \mathbf{0} &= -2\mathbb{E}[y\mathbf{x}] + 2\boldsymbol{\Theta}_\circ \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \\
\text{thus } \boldsymbol{\Theta}_\circ &= \{\boldsymbol{\theta} : \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} = \mathbb{E}[\mathbf{x}y]\}
\end{aligned}
$$

b). For any matrix $\mathbb{E}[\phi\phi^\top]^-$ such that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbb{E}[\phi\phi^\top]^-\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$, if $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} = \mathbb{E}[\mathbf{x}y]$, we have

$$\begin{aligned}
\mathbb{E}[\phi\phi^\top]^-\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} &= \mathbb{E}[\phi\phi^\top]^-\mathbb{E}[\mathbf{x}y] \\
\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbb{E}[\phi\phi^\top]^-\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} &= \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbb{E}[\phi\phi^\top]^-\mathbb{E}[\mathbf{x}y] \\
\mathbb{E}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} &= \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbb{E}[\phi\phi^\top]^-\mathbb{E}[\mathbf{x}y]
\end{aligned}$$

Thus, $\boldsymbol{\theta} = \mathbb{E}[\phi\phi^\top]^-\mathbb{E}[\mathbf{x}y]$ is in the set of $\boldsymbol{\Theta}_\circ$.

If $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ has full rank, the (generalized) inverse matrix $\mathbb{E}[\phi\phi^\top]^- = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]^{-1}$ is uniquely determined by $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ and $\boldsymbol{\Theta}_\circ = \{\mathbb{E}[\mathbf{x}\mathbf{x}^\top]^{-1}\mathbb{E}[\mathbf{x}y]\}$.

c). Suppose $\mathbf{x} = [\begin{smallmatrix}1\\x\end{smallmatrix}]$, the target parameter $\boldsymbol{\theta}_\circ$is derived as follows:

$$\begin{aligned}
\boldsymbol{\theta}_\circ &= \mathbb{E}[\mathbf{x}\mathbf{x}^\top]^{-1}\mathbb{E}[\mathbf{x}y] \\
&= \mathbb{E}[\begin{bmatrix}1 & x \\ x & x^2\end{bmatrix}]^{-1}\mathbb{E}[\begin{bmatrix}y \\ xy\end{bmatrix}] = \begin{bmatrix}1 & \mathbb{E}[x] \\ \mathbb{E}[x] & \mathbb{E}[x^2]\end{bmatrix}^{-1}\begin{bmatrix}\mathbb{E}[y] \\ \mathbb{E}[xy]\end{bmatrix} \\
&= \frac{1}{\mathbb{E}[x^2] - \mathbb{E}[x]^2}\begin{bmatrix}\mathbb{E}[x^2] & -\mathbb{E}[x] \\ -\mathbb{E}[x] & 1\end{bmatrix}\begin{bmatrix}\mathbb{E}[y] \\ \mathbb{E}[xy]\end{bmatrix} \\
&= \frac{1}{\mathbb{E}[x^2] - \mathbb{E}[x]^2}\begin{bmatrix}\mathbb{E}[x^2]\mathbb{E}[y] - \mathbb{E}[x]\mathbb{E}[xy] \\ -\mathbb{E}[x]\mathbb{E}[y] + \mathbb{E}[xy]\end{bmatrix} \\
&= \begin{bmatrix}(\mathbb{E}[x^2] - \mathbb{E}[x]^2)\mathbb{E}[y] - \mathbb{E}[x](\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y])/(\mathbb{E}[x^2] - \mathbb{E}[x]^2) \\ \frac{\mathbb{V}[x,y]}{\mathbb{V}[x]}\end{bmatrix} \\
&= \begin{bmatrix}\mathbb{E}[y] - \frac{\mathbb{V}[x,y]}{\mathbb{V}[x]}\mathbb{E}[x] \\ \frac{\mathbb{V}[x,y]}{\mathbb{V}[x]}\end{bmatrix}
\end{aligned}$$

d). The term $\frac{\mathbb{V}[x,y]}{\mathbb{V}[x]}$, the standardized coefficient (or Beta coefficient), is generally interpreted as the 'average treatment' effect of x on y. Greater $|\frac{\mathbb{V}[x,y]}{\mathbb{V}[x]}|$ indicates that x has greater effect on the variable y, and the sign of $\frac{\mathbb{V}[x,y]}{\mathbb{V}[x]}$ indicates the effect is positive or negative. When x is independent from y, $\frac{\mathbb{V}[x,y]}{\mathbb{V}[x]} = 0$ shows that x has no effect on y.

A Problem 1

A.1 Code for 1.b

```
import numpy as np
import matplotlib.pyplot as plt


def risk(theta1, theta2):
    return theta1**2 + theta2**2 - 400*(theta1 + theta2)


minv, maxv = 0, 400
eval_nums = 400

theta1, theta2 = np.meshgrid(np.linspace(minv, maxv, eval_nums),
                             np.linspace(minv, maxv, eval_nums))

values = risk(theta1, theta2)

fig, ax = plt.subplots(figsize=(4, 4))
ax.grid(False)
CS = ax.contour(theta1, theta2, values, levels=10, linewidths=0.5)
ax.set_title("Contour for the Risk Function")
ax.set_xlabel(r"$\theta_1$ / m")
ax.set_ylabel(r"$\theta_2$ / m")
ax.clabel(CS)
fig.tight_layout()
fig.savefig("hw1_1b.pdf", dpi=500)
```

A.2 Code for 1.e

```
import numpy as np
import matplotlib.pyplot as plt


mean = [200, 200]
cov = [[400, 50], [50, 400]]
ns = [1, 10, 100, 1000]

fig, axs = plt.subplots(nrows=2, ncols=2, figsize=(6, 6))

for idx, n in enumerate(ns):
    row, col = idx // 2, idx % 2
    ax = axs[row][col]
    z = np.random.multivariate_normal(mean, cov, n)

    def risk(z, theta1, theta2):
        # shape of z: num * 2
        t1 = np.sum(z * z) / z.shape[0]
        avg = np.sum(z, 0)/z.shape[0]
        t2 = -2 * ((theta1 * avg[0]) + (theta2 * avg[1]))
        t3 = theta1 ** 2 + theta2 ** 2
        return t1 + t2 + t3
```

```
    minv, maxv = 0, 400
    eval_nums = 400

    theta1, theta2 = np.meshgrid(np.linspace(minv, maxv, eval_nums),
                                 np.linspace(minv, maxv, eval_nums))

    values = risk(z, theta1, theta2)

    ax.grid(False)
    CS = ax.contour(theta1, theta2, values, levels=10, linewidths=0.5)
    opt_point = np.sum(z, 0)/z.shape[0]
    ax.annotate(f'{opt_point[0]:9.2f},{opt_point[1]:9.2f}', xy=opt_point+5)
    ax.plot(opt_point[0], opt_point[1], 'ro', alpha=0.5, label="Mininal")
    ax.set_title(f"Risk Function: n={n}")
    ax.set_xlabel(r"$\theta_1$ / m")
    ax.set_ylabel(r"$\theta_2$ / m")
    ax.clabel(CS)
    ax.legend()
fig.tight_layout()
fig.savefig("hw1_1e.pdf", dpi=500)
```

A.3 Code for 1.f

```
import numpy as np
import matplotlib.pyplot as plt


mean = [200, 200]
cov = [[400, 50], [50, 400]]
n = 1000

z = np.random.multivariate_normal(mean, cov, n)


def risk(z, theta1, theta2):
    t1 = np.sum(z * z) / z.shape[0]
    avg = np.sum(z, 0)/z.shape[0]
    t2 = -2 * ((theta1 * avg[0]) + (theta2 * avg[1]))
    t3 = theta1 ** 2 + theta2 ** 2
    return t1 + t2 + t3


def grad(z, theta):
    avg = np.sum(z, 0)/z.shape[0]
    return 2*(theta - avg)


lr = 0.01
steps = 500

theta = np.array([0, 0])
thetas = []
for i in range(steps):
    if not i % 10:
        thetas.append(theta)
```

```python
    gradient = grad(z, theta)
    theta = theta - lr * gradient

thetas = np.array(thetas)
print(thetas)
# plot GD steps
minv, maxv = 0, 400
eval_nums = 400

theta1, theta2 = np.meshgrid(np.linspace(minv, maxv, eval_nums),
                             np.linspace(minv, maxv, eval_nums))

values = risk(z, theta1, theta2)

fig, ax = plt.subplots(figsize=(4, 4))
ax.grid(False)
CS = ax.contour(theta1, theta2, values, levels=10, linewidths=0.5)
ax.plot(thetas[:-1, 0], thetas[:-1, 1], '-o',
        markersize=3, linewidth=0.5, label="GD steps")
opt_point = np.sum(z, 0)/z.shape[0]
ax.annotate(f'{opt_point[0]:9.2f},{opt_point[1]:9.2f}', xy=opt_point+5)
ax.plot(opt_point[0], opt_point[1], 'ro', alpha=0.5, label="Mininal")
ax.set_title("Gradient Descent for optimization")
ax.set_xlabel(r"$\theta_1$ / m")
ax.set_ylabel(r"$\theta_2$ / m")
ax.clabel(CS)
ax.legend()
fig.tight_layout()
fig.savefig("hw1_1f.pdf", dpi=500)
```