Exercises
Learning and Intelligent Systems
SS 2017

Series 1, Mar 6, 2017 (Probability and Linear Algebra)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Gunnar Rätch and Prof. Dr. Thomas Hofmann
Web: `http://las.inf.ethz.ch/teaching/lis-s17/`
Email questions to:
Josip Djolonga, `josipd@inf.ethz.ch`

It is not mandatory to submit solutions and sample solutions will be published in two weeks. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise1` containing a PDF (LaTeXor scan) to `josipd@inf.ethz.ch` until Tuesday, Mar 14, 2017.

Problem 1 (Linear Regression and Ridge Regression):

Let $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ be the training data that you are given. As you have to predict a continuous variable, one of the simplest possible models is linear regression, i.e. to predict y as $\mathbf{w}^T\mathbf{x}$ for some parameter vector $\mathbf{w} \in \mathbb{R}^d$.[1] We thus suggest minimizing the following loss

$$\operatorname*{argmin}_{\mathbf{w}} \hat{R}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2. \tag{1}$$

Let us introduce the $n \times d$ matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with the $\mathbf{x}_i$ as rows, and the vector $\mathbf{y} \in \mathbb{R}^n$ consisting of the scalars y_i. Then, (1) can be equivalently re-written as

$$\operatorname*{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

We refer to any $\mathbf{w}^*$ that attains the above minimum as a solution to the problem.

(a) Show that if $\mathbf{X}^T\mathbf{X}$ is invertible, then there is a unique $\mathbf{w}^*$ that can be computed as $\mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

(b) Show for $n < d$ that (1) does not admit a unique solution. Intuitively explain why this is the case.

(c) Consider the case $n \geq d$. Under what assumptions on $\mathbf{X}$ does (1) admit a unique solution $\mathbf{w}^*$? Give an example with $n = 3$ and $d = 2$ where these assumptions do not hold.

The *ridge regression* optimization problem with parameter $\lambda > 0$ is given by

$$\operatorname*{argmin}_{\mathbf{w}} \hat{R}_{\text{Ridge}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \left[\sum_{i=1}^{n} \left(y_i - w^T\mathbf{x}_i\right)^2 + \lambda\mathbf{w}^T\mathbf{w}\right]. \tag{2}$$

(d) Show that $\hat{R}_{\text{Ridge}}(\mathbf{w})$ is convex with regards to $\mathbf{w}$. You can use the fact that a twice differentiable function is convex if and only if its Hessian $\mathbf{H} \in \mathbb{R}^{d \times d}$ satisfies $\mathbf{w}^T\mathbf{H}\mathbf{w} \geq 0$ for all $\mathbf{w} \in \mathbb{R}^d$ (is positive semi-definite).

(e) Derive the closed form solution $\mathbf{w}^*_{\text{Ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda I_d\right)^{-1}\mathbf{X}^T\mathbf{y}$ to (2) where I_d denotes the identity matrix of size $d \times d$.

(f) Show that (2) admits the unique solution $\mathbf{w}^*_{\text{Ridge}}$ for any matrix $\mathbf{X}$. Show that this even holds for the cases in (b) and (c) where (1) does not admit a unique solution $\mathbf{w}^*$.

(g) What is the role of the term $\lambda\mathbf{w}^T\mathbf{w}$ in $\hat{R}_{\text{Ridge}}(\mathbf{w})$? What happens to $\mathbf{w}^*_{\text{Ridge}}$ as $\lambda \to 0$ and $\lambda \to \infty$?

[1] Without loss of generality, we assume that both $\mathbf{x}_i$ and y_i are centered, i.e. they have zero empirical mean. Hence we can neglect the otherwise necessary bias term b.

Solution 1:

(a) Note that

$$\hat{R}(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}.$$

The gradient of this function is equal to

$$\nabla\hat{R}(\mathbf{w}) = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y}.$$

Because $\hat{R}(\mathbf{w})$ is convex (formally proven in (d)), its optima are exactly those points that have a zero gradient, i.e. those $\mathbf{w}^*$ that satisfy $\mathbf{X}^T\mathbf{X}\mathbf{w}^* = \mathbf{X}^T\mathbf{y}$. Under the given assumption, the unique minimizer is indeed equal to $\mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

(b) Consider the *singular value decomposition* $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ where $\mathbf{U}$ is an unitary $n \times n$ matrix, $\mathbf{V}$ is a unitary $d \times d$ matrix and $\mathbf{\Sigma}$ is a diagonal $n \times d$ matrix with the singular values of $\mathbf{X}$ on the diagonal. We then have

$$\operatorname*{argmin}_{\mathbf{w}} \hat{R}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \left[\mathbf{w}^T\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T\mathbf{w} - 2\mathbf{y}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w}\right]$$

Since $\mathbf{V}$ is unitary, we may rotate $\mathbf{w}$ using $\mathbf{V}$ to $\mathbf{z} = \mathbf{V}^T\mathbf{w}$ and formulate the optimization problem in terms of $\mathbf{z}$, i.e.

$$\operatorname*{argmin}_{\mathbf{z}} \left[\mathbf{z}^T\mathbf{\Sigma}^2\mathbf{z} - 2\mathbf{y}^T\mathbf{U}\mathbf{\Sigma}\mathbf{z}\right] = \operatorname*{argmin}_{\mathbf{z}} \sum_{i=1}^{d} \left[z_i^2\sigma_i^2 - 2(\mathbf{U}^t\mathbf{y})_i z_i \sigma_i\right]$$

where σ_i is the i entry in the diagonal of $\mathbf{\Sigma}$. Note that this problem decomposes into d independent optimization problems of the form

$$z_i = \operatorname*{argmin}_{z} \left[z^2\sigma_i^2 - 2(\mathbf{U}^t\mathbf{y})_i z \sigma_i\right]$$

for $i = 1, 2, \dots, d$. Since each problem is quadratic and thus convex we may obtain the solution by finding the root of the first derivative. For $i = 1, 2, \dots d$ we require that z_i satisfies

$$z_i\sigma_i^2 - (\mathbf{U}^t\mathbf{y})_i\sigma_i = 0.$$

For all $i = 1, 2, \dots d$ such that $\sigma_i \neq 0$, the solution z_i is thus given by

$$z_i = \frac{(\mathbf{U}^t\mathbf{y})_i}{\sigma_i}.$$

For the case $n < d$, however, $\mathbf{X}$ has at most rank n as it is a $n \times d$ matrix and hence at most n of its singular values are nonzero. This means that there is at least one index j such that $\sigma_j = 0$ and hence any $z_j \in \mathbb{R}$ is a solution to the optimization problem. As a result the set of optimal solutions for $\mathbf{z}$ is a linear subspace of at least one dimension. By rotating this subspace using $\mathbf{V}$, i.e. $\mathbf{w} = \mathbf{V}\mathbf{z}$, it is evident that the optimal solution to the optimization problem in terms of $\mathbf{w}$ is also a linear subspace of at least one dimension and that thus no unique solution exists. Furthermore, since $\mathbf{X}$ has at most rank n, $\mathbf{X}^T\mathbf{X}$ is not of full rank. As a result $(\mathbf{X}^T\mathbf{X})^{-1}$ does not exist and $\mathbf{w}^*$ is ill-defined.

The intuition behind these results is that the "linear system" $\mathbf{X}\mathbf{w} \approx \mathbf{y}$ is underdetermined as there are less data points than parameters that we want to estimate.

(c) We showed in (b) that the optimization problem admits a unique solution only if all the singular values of $\mathbf{X}$ are nonzero. For $n \geq d$, this is the case if and only if $\mathbf{X}$ is of full rank, i.e. all the columns of $\mathbf{X}$ are linearly independent. As an example for a matrix not satisfying these assumptions, any matrix with linearly dependent dependent suffices, e.g.

$$\mathbf{X}_{\text{degenerate}} = \begin{pmatrix} 1 & -2 \\ 0 & 0 \\ -2 & 4 \end{pmatrix}.$$

(d) Because convex functions are closed under addition, we will show that each term in the objective is convex, from which the claim will follow. Each data term $(y_i - \mathbf{w}^T\mathbf{x}_i)^2$ has a Hessian $\mathbf{x}_i\mathbf{x}_i^T$, which is positive semi-definite because for any $\mathbf{w} \in \mathbf{R}^d$ we have $\mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} = (\mathbf{x}_i^T\mathbf{w}_i)^2 \geq 0$ (note that $\mathbf{x}_i^T\mathbf{w} = \mathbf{w}^T\mathbf{x}_i$ are scalars). The regularizer $\lambda\mathbf{w}^T\mathbf{w}$ has the identity matrix λI_d as a Hessian, which is also postive semi-definite because for any $\mathbf{w} \in \mathbf{R}^d$ we have $\mathbf{w}^T\lambda I_d\mathbf{w} = \lambda\|\mathbf{w}\|^2 \geq 0$, and this completes the proof.

(e) The gradient of $\hat{R}_{\text{Ridge}}(\mathbf{w})$ with respect to $\mathbf{w}$ is given by

$$\nabla\hat{R}_{\text{Ridge}}(\mathbf{w}) = 2\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + 2\lambda\mathbf{w}.$$

Similar to (a), because $\hat{R}_{\text{Ridge}}(\mathbf{w})$ is convex, we only have to find a point $\mathbf{w}^*_{\text{Ridge}}$ such that

$$\nabla\hat{R}_{\text{Ridge}}(\mathbf{w}^*_{\text{Ridge}}) = 2\mathbf{X}^T(\mathbf{X}\mathbf{w}^*_{\text{Ridge}} - \mathbf{y}) + 2\lambda\mathbf{w}^*_{\text{Ridge}} = 0.$$

This is equivalent to

$$(\mathbf{X}^T\mathbf{X} + \lambda I_d)\mathbf{w}^*_{\text{Ridge}} = \mathbf{X}^T\mathbf{y}$$

which implies the required result

$$\mathbf{w}^*_{\text{Ridge}} = \left(\mathbf{X}^T\mathbf{X} + \lambda I_d\right)^{-1}\mathbf{X}^T\mathbf{y}.$$

(f) Note that $\mathbf{X}^T\mathbf{X}$ is a positive semi-definite matrix since $\forall\mathbf{w} \in \mathbb{R}^d : \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} = \sum_{i=1}^n \left[(\mathbf{X}\mathbf{w})_i\right]^2 \geq 0$, which implies that it has non-negative eigenvalues. But then, $\mathbf{X}^T\mathbf{X} + \lambda I_d$ has eigenvalues bounded from below by $\lambda > 0$, which means that it is invertible and thus the optimum is uniquely defined.

(g) The term $\lambda\mathbf{w}^T\mathbf{w}$ "biases" the solution towards the origin, i.e. there is a quadratic penalty for solutions $\mathbf{w}$ that are far from the origin. The parameter λ determines the extend of this effect: As $\lambda \to 0$, $\hat{R}_{\text{Ridge}}(\mathbf{w})$ converges to $\hat{R}(\mathbf{w})$. As a result the optimal solution $\mathbf{w}^*_{\text{Ridge}}$ approaches the solution of (1). As $\lambda \to \infty$, only the quadratic penalty $\mathbf{w}^T\mathbf{w}$ is relevant and $\mathbf{w}^*_{\text{Ridge}}$ hence approaches the null vector $(0, 0, \ldots, 0)$.

Problem 2 (Normal Random Variables):

Let X be a Normal random variable with mean $\mu \in \mathbb{R}$ and variance $\tau^2 > 0$, i.e. $X \sim \mathcal{N}(\mu, \tau^2)$. Recall that the probability density of X is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\tau} e^{-(x-\mu)^2/2\tau^2}, \quad -\infty < x < \infty.$$

Furthermore, the random variable Y given $X = x$ is normally distributed with mean x and variance σ^2, i.e. $Y|_{X=x} \sim \mathcal{N}(x, \sigma^2)$.

(a) Derive the *marginal distribution* of Y.

(b) Use Bayes' theorem to derive the *conditional distribution* of X given $Y = y$.

Hint: For both tasks derive the density up to a constant factor and use this to identify the distribution.

Solution 2:

As a prelude to both (a) and (b) we consider the joint density function $f_{X,Y}(x,y)$ of X and Y

$$f_{X,Y}(x,y) = f_{Y|X}(y|X=x) f_X(x) = \frac{1}{2\pi\sigma\tau} \exp\left(-\frac{1}{2}\left[\underbrace{\frac{(x-\mu)^2}{\tau^2} + \frac{(y-x)^2}{\sigma^2}}_{\text{(A)}}\right]\right).$$

Using simple algebraic operations, we obtain

$$\begin{aligned}
\text{(A)} &= \frac{(x^2 - 2\mu x + \mu^2)\sigma^2 + (x^2 - 2xy + y^2)\tau^2}{\sigma^2\tau^2} \\
&= \frac{(\sigma^2 + \tau^2)x^2 - 2x(\sigma^2\mu + \tau^2 y) + \sigma^2\mu^2 + \tau^2 y^2}{\sigma^2\tau^2} \\
&= \frac{\left(\sigma^2 + \tau^2\right)\left[x^2 - 2x\left(\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}\right) + \left(\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}\right)^2 - \left(\frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}\right)^2\right] + \sigma^2\mu^2 + \tau^2 y^2}{\sigma^2\tau^2} \\
&= \underbrace{\frac{\left(x - \left(\frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y\right)\right)^2}{\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}_{\text{(B)}} + \underbrace{\frac{\sigma^2\mu^2 + \tau^2 y^2 - \frac{\left(\sigma^2\mu + \tau^2 y\right)^2}{\sigma^2+\tau^2}}{\sigma^2\tau^2}}_{\text{(C)}}.
\end{aligned}$$

(a) The marginal density of Y is given by

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx = \int_{\mathbb{R}} f_{Y|X}(y|X=x) f_X(x) dx.$$

This is proportional to

$$f_Y(y) \propto \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\left[\underbrace{\frac{\left(x - \left(\frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y\right)\right)^2}{\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}_{\text{(B)}}\right]\right) dx \exp\left(-\frac{1}{2}\left[\underbrace{\frac{\sigma^2\mu^2 + \tau^2 y^2 - \frac{\left(\sigma^2\mu+\tau^2 y\right)^2}{\sigma^2+\tau^2}}{\sigma^2\tau^2}}_{\text{(C)}}\right]\right).$$

Note that (B) matches the functional form of a normal density for the variable x. As a result, the first term integrates to $\sigma\tau\sqrt{2\pi/(\sigma^2+\tau^2)}$ and we thus only need to consider (C) to identify $f_Y(y)$, i.e.

$$
\begin{aligned}
f_Y(y) &\propto \exp\left(-\frac{1}{2}\left[\underbrace{\frac{\sigma^2\mu^2+\tau^2y^2-\frac{(\sigma^2\mu+\tau^2y)^2}{\sigma^2+\tau^2}}{\sigma^2\tau^2}}_{\text{(C)}}\right]\right)\\
&= \exp\left(-\frac{1}{2}\left[\frac{(\sigma^4\mu^2+\sigma^2\tau^2\mu^2+\sigma^2\tau^2y^2+\tau^4y^2)-(\sigma^4\mu^2+2\sigma^2\tau^2\mu y+\tau^4y^2)}{\sigma^2\tau^2(\sigma^2+\tau^2)}\right]\right)\\
&= \exp\left(-\frac{1}{2}\left[\frac{\sigma^2\tau^2\mu^2-2\sigma^2\tau^2\mu y+\sigma^2\tau^2y^2}{\sigma^2\tau^2(\sigma^2+\tau^2)}\right]\right)\\
&= \exp\left(-\frac{1}{2}\left[\frac{(\mu-y)^2}{(\sigma^2+\tau^2)}\right]\right).
\end{aligned}
$$

It can easily be seen that the marginal distribution of Y is the Normal distribution with mean μ and variance $\sigma^2+\tau^2$.

(b) The conditional density of X given $Y=y$ is proportional to the joint density function, i.e.

$$
f_{X|Y}(x|Y=y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}\propto f_{X,Y}(x,y).
$$

Since (C) is independent of x we only need to consider (B) and have

$$
f_{X|Y}(x|Y=y)\propto \exp\left(-\frac{1}{2}\left[\underbrace{\frac{\left(x-\left(\frac{\sigma^2}{\sigma^2+\tau^2}\mu+\frac{\tau^2}{\sigma^2+\tau^2}y\right)\right)^2}{\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}}}_{\text{(B)}}\right]\right).
$$

Similarly to (a), it immediately follows that the conditional distribution of X given $Y=y$ is the Normal distribution with mean $\left(\frac{\sigma^2}{\sigma^2+\tau^2}\mu+\frac{\tau^2}{\sigma^2+\tau^2}y\right)$ and variance $\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$. Note that the mean is a convex combination of μ and the observation y.

Problem 3 (Bivariate Normal Random Variables):

Let X be a bivariate Normal random variable (taking on values in $\mathbb{R}^2$) with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. The density of X is then given by

$$f_X(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)\right).$$

Find the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$.

Solution 3:

We present two approaches for this exercise:

Approach 1. Note that $Z = 0$ implies $X_1 = X_2$. Furthermore by the definition of Y, we have $X_1 = X_2 = Y/2$ given $Z = 0$. Hence the marginal density of Y given $Z = 0$ is proportional to

$$f_{Y|Z}(y|Z = 0) = \frac{f_{Y,Z}(y, 0)}{f_Z(0)} \propto f_{Y,Z}(y, 0) \propto f_X\left[\begin{pmatrix} y/2 \\ y/2 \end{pmatrix}\right].$$

We then have

$$\begin{aligned}
f_X\left[\begin{pmatrix} y/2 \\ y/2 \end{pmatrix}\right] &\propto \exp\left(-\frac{1}{2}\begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}^T \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}\right) \\
&= \exp\left(-\frac{1}{2}\begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}^T \frac{1}{5}\begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} \frac{y}{2} - 1 \\ \frac{y}{2} - 1 \end{pmatrix}\right) \\
&= \exp\left(-\frac{1}{2}\frac{(y-2)^2}{\frac{20}{3}}\right).
\end{aligned}$$

Clearly the conditional distribution of Y given $Z = 0$ is hence Normal with mean 2 and variance $\frac{20}{3}$.

Approach 2. We define the random variable $\mathbf{R}$ as

$$\mathbf{R} = \begin{pmatrix} Y \\ Z \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}}_{=\mathbf{A}} \mathbf{X}.$$

By linearity of expectation, the mean $\mu_{\mathbf{R}}$ of $\mathbf{R}$ is

$$\mathbb{E}[\mathbf{R}] = \mathbf{A}\mathbb{E}[\mathbf{X}] = \mathbf{A}\mu = \begin{pmatrix} 2 \\ 0 \end{pmatrix}.$$

The covariance matrix $\mathbf{\Sigma_R}$ of $\mathbf{R}$ is given by

$$\begin{aligned}
\mathbf{\Sigma_R} &= \mathbb{E}[(\mathbf{R} - \mathbb{E}[\mathbf{R}])(\mathbf{R} - \mathbb{E}[\mathbf{R}])^T] = \mathbb{E}[\mathbf{A}(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T \mathbf{A}^T] \\
&= \mathbf{A}\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]\mathbf{A}^T = \mathbf{A\Sigma A}^T \\
&= \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 4 & 3 \\ 2 & -1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}
\end{aligned}$$

Since $\mathbf{X}$ is multivariate Gaussian and $\mathbf{R}$ is an affine transformation of $\mathbf{X}$, $\mathbf{R}$ is a bivariate Normal random variable with mean $\mu_{\mathbf{R}}$ and covariance matrix $\mathbf{\Sigma_R}$.[2] The conditional density of Y given $Z = 0$ is then given by

$$\begin{aligned} f_{Y|Z}(y|Z=0) &= \frac{f_{Y,Z}(y,0)}{f_Z(0)} \propto f_{Y,Z}(y,0) \\ &\propto \exp\left(-\frac{1}{2}\begin{pmatrix} y-2 \\ 0 \end{pmatrix}^T \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}^{-1} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}\right) \\ &= \exp\left(-\frac{1}{2}\begin{pmatrix} y-2 \\ 0 \end{pmatrix}^T \frac{1}{20}\begin{pmatrix} 3 & -1 \\ -1 & 7 \end{pmatrix} \begin{pmatrix} y-2 \\ 0 \end{pmatrix}\right) \\ &= \exp\left(-\frac{1}{2}\frac{(y-2)^2}{\frac{20}{3}}\right). \end{aligned}$$

Clearly the conditional distribution of Y given $Z = 0$ is hence Normal with mean 2 and variance $\frac{20}{3}$.

[2] This result can be easily derived from the characteristic function of the multivariate Normal distribution. $\mathbf{R}$ is bivariate Normal if and only if for any $\mathbf{t} \in \mathbb{R}^2$

$$\mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{R}}\right] = e^{i\mathbf{t}^T\mu_{\mathbf{R}} - \mathbf{t}^T\mathbf{\Sigma_R}\mathbf{t}/2}.$$

This holds since the corresponding property holds for $\mathbf{X}$ with $\mathbf{s} = \mathbf{t}^T\mathbf{A}$, i.e.

$$\mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{R}}\right] = \mathbb{E}\left[e^{i\mathbf{t}^T\mathbf{A}\mathbf{X}}\right] = \mathbb{E}\left[e^{i\mathbf{s}^T\mathbf{X}}\right] = e^{i\mathbf{s}^T\mu - \mathbf{s}^T\mathbf{\Sigma}\mathbf{s}/2} = e^{i\mathbf{t}^T\mathbf{A}\mu - \mathbf{t}^T\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\mathbf{t}/2} = e^{i\mathbf{t}^T\mu_{\mathbf{R}} - \mathbf{t}^T\mathbf{\Sigma_R}\mathbf{t}/2}.$$