Exercises
Introduction to Machine Learning
FS 2018

Series 1, Feb 21st, 2019 (Probability, Analysis, Linear Algebra)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s18`
Email questions to:
Mohammad Reza Karimi, mkarimi@ethz.ch

We will publish sample solutions on Friday, Mar 8th.

Problem 1 (Sampling):

Knowing the CDF of a random variable X, enables one to draw samples from that distribution.

(a) Show that if X has distribution function F, and $U \sim \mathrm{Unif}(0,1)$ is a uniform random number in the interval $(0,1)$, then $F^{-1}(U)$ has the same distribution as X.

In situations where the inverse of F is not easy to compute, one can use the following method (known as the *rejection method*) for generating random variables with a density f. Suppose that γ be a function such that $\gamma(x) \geq f(x)$ for all $x \in \mathbb{R}$, and

$$\int_{-\infty}^{\infty} \gamma(x)\,dx = \alpha < \infty.$$

Then, $g(x) = \gamma(x)/\alpha$ is a probability density function. Suppose we generate a random variable X by the following algorithm:

I. Generate a random variable T with density function g.

II. Generate a random variable $U \sim \mathrm{Unif}(0,1)$, independent of T. If $U \leq f(T)/\gamma(T)$ then set $X = T$; if $U > f(T)/\gamma(T)$ then repeat steps I and II.

(b) Show that the generated random variable X has density f.

(c) Show that the number of rejections before X is generated has a Geometric distribution. Give an expression for the parameter of this distribution.

Hints For part (a) note that

$$F^{-1}(u) = \inf\{x \mid F(x) \geq u\},$$

as F is right-continuous. For part (b), you need to evaluate

$$\mathbb{P}(T \leq x \mid U \leq f(T)/\gamma(T)).$$

Solution 1:

(a) Define $Y = F^{-1}(U)$, where $U \sim \mathrm{Unif}(0,1)$ and F is a CDF. Computing the CDF of Y gives

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(F^{-1}(U) \leq y) = \mathbb{P}(U \leq F(y)) = F(y),$$

where we used the fact that $F^{-1}(u) \leq y \iff u \leq F(y)$. Thus, Y has F as its distribution function.

(b) Before continuing reading the solution, take a look at the first box in the solution of problem 2 (b).

First let us compute the probability
$$\mathbb{P}\left(U \le \frac{f(T)}{\gamma(T)}\right).$$

As a reminder, this probability is defined on the joint distribution of U and T. As U and T are independent, the joint probability space is simply the product space defined over $(0,1) \times \mathbb{R}$.

By conditioning on the value of T, and using the fact that T has density g, we get the following:
$$\begin{aligned}
\mathbb{P}\left(U \le \frac{f(T)}{\gamma(T)}\right) &= \int_{\mathbb{R}} \mathbb{P}\left(U \le \tfrac{f(T)}{\gamma(T)} \mid T = t\right) g(t)\,dt \\
&= \int_{\mathbb{R}} \mathbb{P}\left(U \le \tfrac{f(t)}{\gamma(t)}\right) g(t)\,dt \\
&= \int_{\mathbb{R}} \frac{f(t)}{\gamma(t)} g(t)\,dt \\
&= \int_{\mathbb{R}} \frac{1}{\alpha} f(t)\,dt = \frac{1}{\alpha},
\end{aligned}$$

where in the second line, we used the fact that T and U are independent, thus we can remove the conditioning.

Also, we need to compute the following probability for $x \in \mathbb{R}$:
$$\begin{aligned}
\mathbb{P}\left(T \le x, U \le \frac{f(T)}{\gamma(T)}\right) &= \int_{\mathbb{R}} \mathbb{P}\left(T \le x, U \le \tfrac{f(T)}{\gamma(T)} \mid T = t\right) g(t)\,dt \\
&= \int_{-\infty}^{x} \frac{f(t)}{\gamma(t)} g(t)\,dt = \frac{1}{\alpha}\int_{-\infty}^{x} f(t)\,dt.
\end{aligned}$$

Now, by the definition of conditional probability, we have
$$\mathbb{P}\left(T \le x \mid U \le \tfrac{f(T)}{\gamma(T)}\right) = \frac{\mathbb{P}\left(T \le x, U \le \tfrac{f(T)}{\gamma(T)}\right)}{\mathbb{P}\left(U \le \tfrac{f(T)}{\gamma(T)}\right)} = \frac{\frac{1}{\alpha}\int_{-\infty}^{x} f(t)\,dt}{1/\alpha} = \int_{-\infty}^{x} f(t)\,dt.$$

This means that if the choice of T and U resulted in an acceptance, the density of T is f.

(c) As computed in part (b), the probability of acceptance is $1/\alpha$. One can think of it as a coin with bias (probability of heads) $p = 1/\alpha$. Thus, the number of throws (rejections) until the first heads (the first acceptance) has a Geometric distribution.

Problem 2 (Multivariate Normal Distribution):

Recall the following fact about characteristic functions:

Fact 1. For a random vector X in $\mathbb{R}^d$, define its characteristic function φ_X as

$$\varphi_X(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}^\top X)], \quad \text{for all } \mathbf{t} \in \mathbb{R}^d.$$

The characteristic function completely identifies a distribution. For a multivariate Normal distribution $\mathcal{N}(\mu, \Sigma)$, one has

$$\varphi(\mathbf{t}) = \exp(i\mathbf{t}^\top \mu - \tfrac{1}{2}\mathbf{t}^\top \Sigma \mathbf{t}).$$

(a) Let $X = (X_1, \ldots, X_d)$ be a d-dimensional standard Gaussian random vector, that is, $X \sim \mathcal{N}_d(0, I)$. Define $Y = AX + \mu$, where A is a $d \times d$ matrix and $\mu \in \mathbb{R}^d$. What is the distribution of Y? If B is an $r \times d$ matrix, what is the distribution of BY?

(b) Let X be a bivariate Normal random variable (taking on values in $\mathbb{R}^2$) with mean $\mu = (1, 1)$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} 3 & 1 \\ 1 & 2 \end{smallmatrix}\right)$. Find the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$.

(c*) For $Y \sim \mathcal{N}_d(0, I)$, we say that the random variable $V = \|Y\|^2$ has the χ^2 (chi-square) distribution with d degrees of freedom ($V \sim \chi^2(d)$). Assume that $X_1, \ldots, X_n$ are i.i.d. samples from the Normal distribution $\mathcal{N}(\mu, \sigma^2)$. One way to estimate σ^2 from these samples is to look at the sample variance:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

where $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.

Prove that $\frac{(n-1)}{\sigma^2}S^2$ has a chi-square distribution with $n - 1$ degrees of freedom.

Hint: Can you write S^2 as the norm-squared of a vector? Which vector? Take care of the dimensions.

Solution 2:

(a) Let us compute the characteristic function of Y. Define $\mathbf{s} = A^\top \mathbf{t}$. We have

$$\begin{aligned}
\varphi_Y(\mathbf{t}) &= \mathbb{E}[\exp(i\mathbf{t}^\top Y)] \\
&= \mathbb{E}[\exp(i\mathbf{t}^\top AX) \cdot \exp(i\mathbf{t}^\top \mu)] \\
&= \mathbb{E}[\exp(i\mathbf{s}^\top X)] \cdot \exp(i\mathbf{t}^\top \mu) \\
&= \varphi_X(\mathbf{s}) \cdot \exp(i\mathbf{t}^\top \mu) \\
&= \exp(-\tfrac{1}{2}\mathbf{s}^\top \mathbf{s} + i\mathbf{t}^\top \mu) \\
&= \exp(i\mathbf{t}^\top \mu - \tfrac{1}{2}\mathbf{t}^\top AA^\top \mathbf{t}),
\end{aligned}$$

which means that $Y \sim \mathcal{N}(\mu, AA^\top)$. With the same argument as above, one gets $BY \sim \mathcal{N}(B\mu, BAA^\top B^\top)$.

(b) First, take a look at the following facts:

Let A, B be events. The definition of conditional probability $\mathbb{P}(A \mid B)$ assumes that $\mathbb{P}(B) \neq 0$. So one essentially cannot condition on events of zero probability in the usual way. The following is a workaround to this issue.

Let X, Y be random variables with joint density f and joint CDF F. For $\varepsilon > 0$ and $x, y \in \mathbb{R}$, we compute

$$\begin{aligned}
\mathbb{P}(X \le x \mid Y \in [y, y+\varepsilon]) &= \frac{\mathbb{P}(X \le x, Y \in [y, y+\varepsilon])}{\mathbb{P}(Y \in [y, y+\varepsilon])} \\
&= \frac{F(x, y+\varepsilon) - F(x, y)}{F_Y(y+\varepsilon) - F_Y(y)} \\
&= \frac{[F(x, y+\varepsilon) - F(x, y)]/\varepsilon}{[F_Y(y+\varepsilon) - F_Y(y)]/\varepsilon}.
\end{aligned}$$

Now if $\varepsilon \to 0$, the right hand side has the limit $\frac{\partial_y F(x,y)}{f_Y(y)}$, and the left hand side can be regarded as $\mathbb{P}(X \le x \mid Y = y)$. Taking derivative with respect to x gives the conditional density

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

One can use this density to compute probabilities like $\mathbb{P}(X \in A \mid Y = y) = \iint_A \frac{f(x,y)}{f_Y(y)}\, dxdy$.

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

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

The last equality is due to the fact that the linear map $(x_1, x_2) \mapsto (x_1 + x_2, x_1 - x_2)$ has constant determinant of -2. Thus, by a change of variables formula, the density changes by a constant factor. 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 $\frac{20}{3}$.

In this problem, we used the following trick which prevents a lot of computational headaches. If one is trying to derive the density of a random variable X at x, that is, $f_X(x)$, it is easier to neglect all *multiplicative* terms that does not include x. The reason is simply because $\int_{\mathbb{R}} f_X(x)\, dx = 1$.

Two important examples are single varible Normal random variables and multivariate Gaussian vectors. In the first case, following the trick above, we conclude that if a density function is of the form

$$f(x) \propto \exp(-ax^2 + bx)$$

for $a > 0$ and $b \in \mathbb{R}$, by completing the squares, we obtain

$$-ax^2 + bx = -a(x - \tfrac{b}{2a})^2 + \frac{b^2}{4a},$$

and thus, by removing the terms that does not depend on x, we get

$$f(x) \propto \exp\left(-\frac{(x-\frac{b}{2a})^2}{1/a}\right),$$

meaning that the distribution is a Normal distribution with mean $\frac{b}{2a}$ and variance $1/a$.

The situation for multivariate normal distribution is the same. One needs only to create a proper quadratic form in the exponent to get the familiar multivariate Gaussian density.

APPROACH 2. We define the random variable R as

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

Notice that R is a linear tranformation of a Gaussian vector, and by part (a), it is a Gaussian vector. Thus, we only need to compute its mean and covariance matrix. By linearity of expectation, the mean μ_R of R is

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

The covariance matrix Σ_R of R is also given by part (a):

$$\Sigma_R = A\Sigma A^\top = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 7 & 1 \\ 1 & 3 \end{pmatrix}$$

The conditional density of Y given $Z = 0$ is then given by

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

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

(c*) Before bringing the solution, let us introduce three important linear transformations.

Let v be a unit vector in $\mathbb{R}^d$. The *orthogonal projector* on the direction of v is the linear transformation described by the matrix $vv^\top$ (the outer product of v by itself). The orthogonal projection on the hyperplane defined by v is then $I - vv^\top$. Also the *reflection* about the hyperplane defined by v is $I - 2vv^\top$ (verify these by drawing a picture). Sometimes, the last transformation is called a *Householder Reflector*.

If one is searching for a unitary matrix that maps u to v, one possible way is to consider the Householder reflector about the hyperplane defined by $(v-u)/\|v-u\|$.

Let us now define $X = (X_1, \ldots, X_n)$. The fact that X_i are i.i.d. implies that $X \sim \mathcal{N}(\mu, \sigma^2 I_n)$. Consider the unit vector $v = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$. The projection of X on the direction of v is

$$vv^\top X = \begin{pmatrix} 1/n & \cdots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \cdots & 1/n \end{pmatrix} X = \begin{pmatrix} \bar{X} \\ \vdots \\ \bar{X} \end{pmatrix}.$$

Thus, the projection on the hyperplane defined by v is the vector $Y = (I - vv^\top)X = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$. Note here that $\|Y\|^2 = (n-1)S^2$. Note that Y is a Gaussian vector, as it is a linear function of X. Also notice that the transformation $I - vv^\top$ is of rank $n-1$. Hence, it is better to transform Y in a way that one component becomes zero, while keeping the norm of Y fixed. That is, we need a unitary map that maps v to $w = (1, 0, \ldots, 0)$. Using Householder reflectors, this map is indeed $I - 2uu^\top$, where $u = (v - w)/\|v - w\|$.

Denote by $Z = (I - 2uu^\top)Y$. Observe that Z is a Gaussian vector. It is easy to verify that the mean of Z is zero. The covariance matrix can be computed using part (a):

$$\begin{aligned}
\Sigma_Z &= (I - 2uu^\top)(I - vv^\top)(\sigma^2 I)(I - vv^\top)^\top (I - 2uu^\top)^\top \\
&= \sigma^2 (I - 2uu^\top)(I - vv^\top)(I - 2uu^\top) \\
&= \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}
\end{aligned}$$

Thus, Z/σ is a Gaussian random vector, that is supported on a $(n-1)$-dimensional space, with mean 0 and covariance I_{n-1}. That is, it is a standard Gaussian vector in $\mathbb{R}^{n-1}$. Hence, $\frac{1}{\sigma^2}\|Z\|^2$ has chi-square distribution with $(n-1)$ degrees of freedom. But $(n-1)S^2 = \|Y\|^2 = \|Z\|^2$. Thus,

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1).$$

Problem 3 (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

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

In this exercise, $\|\cdot\|$ is always the Euclidean norm. 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

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

(d) Show that $\hat{R}_{\text{ridge}}$ is convex with respect to $\mathbf{w}$. You can use the fact that a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if for any $\mathbf{x} \in \mathbb{R}^d$ its Hessian $D^2 f(\mathbf{x}) \in \mathbb{R}^{d \times d}$ 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) A continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called α-*strongly convex* for some $\alpha > 0$, if for any points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ one has

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\alpha}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

If f is twice differentiable, an equivalent condition is that for any point $\mathbf{x} \in \mathbb{R}^d$, one has

$$D^2 f(\mathbf{x}) \succeq \alpha I,$$

which means $D^2 f(\mathbf{x}) - \alpha I$ is always positive semi-definite. Prove that a strongly convex function admits a unique minimizer in $\mathbb{R}^d$. *Hint: prove that $f(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\| \to \infty$.*

(g) 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 $\mathbf{w}^*$.

(h) What is the role of the term $\lambda\mathbf{w}^T\mathbf{w}$ in $\hat{R}_{\text{ridge}}$? What happens to $\mathbf{w}^*_{\text{ridge}}$ as $\lambda \to 0$ and $\lambda \to \infty$? You do not need to give a complete proof, only an intuitive answer suffice.

[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.

Solution 3:

(a) Note that $\hat{R} : \mathbb{R}^d \to \mathbb{R}$ and

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

The gradient of this function is equal to (see the recap slides; also note that the gradient is a vector in $\mathbb{R}^d$)

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

Because $\hat{R}(\mathbf{w})$ is convex (formally proven in (d)), its optima (if they exist) 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* $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ where $\mathbf{U}$ is an unitary $n \times n$ matrix, $\mathbf{V}$ is a unitary $d \times d$ matrix and Σ is a diagonal $n \times d$ matrix with the singular values of $\mathbf{X}$ on the diagonal. We then have

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

Note that $\mathbf{y}^T\mathbf{U}\Sigma\mathbf{V}^T\mathbf{w} \in \mathbb{R}$ is a number. Thus, we have the equality $\mathbf{y}^T\mathbf{U}\Sigma\mathbf{V}^T\mathbf{w} = \mathbf{w}^T\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y}$. Also, for brevity, we write Σ^2 instead of $\Sigma^T\Sigma \in \mathbb{R}^{d\times d}$.

Since $\mathbf{V}$ is unitary (and hence it is a bijection), 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.

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

where σ_i is the ith entry in the diagonal of Σ. Note that this problem gets decomposed 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 $i = 1, 2, \ldots, d$. Since each problem is quadratic with positive coefficient and thus convex we may obtain the solution by finding the root of the first derivative. For $i = 1, 2, \ldots d$ we require that z_i satisfies

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

For all $i = 1, 2, \ldots d$ 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.

We use the fact that the rank of a matrix A is equal to the number of nonzero singular values of A.

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 back 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 (for a proof, look at the SVD of $\mathbf{X}^T\mathbf{X}$). 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 the Hessian $\mathbf{x}_i \mathbf{x}_i^T$, which is positive semi-definite because for any $\mathbf{w} \in \mathbb{R}^d$ 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 $\mathbf{w} \in \mathbb{R}^d$ 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) First, let us prove that the function f is *coercive*, i.e., $\lim_{\|\mathbf{x}\| \to \infty} f(\mathbf{x}) = \infty$. In the definition of strong convexity, by putting $\mathbf{x} = 0$ we get

$$f(\mathbf{y}) \geq f(0) + \nabla f(0)^\top \mathbf{y} + \frac{\alpha}{2} \|\mathbf{y}\|^2 \geq f(0) - \|\nabla f(0)\| \cdot \|\mathbf{y}\| + \frac{\alpha}{2} \|y\|^2,$$

where we used the Cauchy-Schwartz inequality: $\nabla f(0)^\top \mathbf{y} \geq -\|\nabla f(0)\| \cdot \|\mathbf{y}\|$. The right-hand side of the equation above is a quadratic function of $\|\mathbf{y}\|$ with a positive coefficient for second degree term. Thus, it goes to infinity as $\|\mathbf{y}\| \to \infty$. Hence, f also goes to infinity.

Next, we prove that f has a global minimum. Denote by $s = \inf_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) < \infty$. Then, there exists a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ such that $f(\mathbf{x}_n) \to s$. We claim that this sequence is bounded: otherwise, there was a subsequence that $\|\mathbf{x}_{n_i}\| \to \infty$. But as f is coercive, $f(\mathbf{x}_{n_i}) \to \infty$, contradicting $f(\mathbf{x}_{n_i}) \to s < \infty$. Hence, the sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots$ is inside some bounded set. By compactness, we obtain that there exists a convergent subsequence. As f is continuous, the f value of this subsequence converges as well, meaning that the infimum is attained. That is, $\exists \mathbf{x}_\infty : f(\mathbf{x}_\infty) = s = \inf f(\mathbf{x})$.

Finally, we prove uniqueness. If $\mathbf{x}$ and $\mathbf{y}$ were two distinct global minima for f, then, by strong convexity, we have

$$f\left(\frac{\mathbf{x} + \mathbf{y}}{2}\right) < \frac{1}{2}(f(\mathbf{x}) + f(\mathbf{y})) = \min f,$$

a contradiction.

(g) Note that $\mathbf{X}^T \mathbf{X}$ is a positive semi-definite matrix, since $\forall \mathbf{w} \in \mathbb{R}^d : \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} = \|\mathbf{X}\mathbf{w}\|^2 \geq 0$, 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.

Note. Since $\mathbf{X}^T\mathbf{X}$ is symmetric, all of its eigenvalues are real, and it is clear that μ is an eigenvalue of $\mathbf{X}^T\mathbf{X}$ if and only if $\mu + \lambda$ is an eigenvalue of $\mathbf{X}^T\mathbf{X} + \lambda I$. Also note that if a linear function is injective, then its kernel is $\{\mathbf{0}\}$, meaning that it does not have a zero eigenvalue. The converse is also true.

Another way to state the result is that $\mathbf{X}^T\mathbf{X} + \lambda I_d \succeq \lambda I_d$, which means that $\hat{R}_{\text{ridge}}$ is λ-strongly convex. Thus, the claim follows.

(h) 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 $(0, 0, \ldots, 0)$.

One can also show this interesting property (however, the proof is involved): Assume $n < d$ (as the situation discussed in (b)). Then $\mathbf{w}^*$ for linear regression is not unique. Denote by $\mathbf{w}^*_\lambda$ the *unique* solution to the Ridge regression problem for $\lambda > 0$. Then the limit $\lim_{\lambda \to 0} \mathbf{w}^*_\lambda$ exists, and the limit point falls inside the space of solutions to linear regression problem. One can further show that this solution is the one with the minimum norm.

Problem 4 (Normal Random Variables):

Let X be a Normal random variable with mean $\mu \in \mathbb{R}$ and variance $\tau^2 > 0$, i.e. $X \sim \mathcal{N}(\mu, \tau^2)$. 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, i.e. compute the density $f_Y(y)$.

(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 4:

Before starting calculations, it is good to mention that one can easily compute the following integral for $a > 0$ by creating complete squares:

$$\begin{aligned}\int_{\mathbb{R}} e^{-(ax^2+2bx+c)} dx &= \int_{\mathbb{R}} \exp\left(-a\left[\left(x+\frac{b}{a}\right)^2 - \frac{b^2-ac}{a^2}\right]\right) dx \\ &= \exp\left(\frac{b^2-ac}{a}\right) \cdot \int_{\mathbb{R}} \exp\left(-\frac{1}{2}\frac{\left(x+\frac{b}{a}\right)^2}{1/2a}\right) dx \\ &= \exp\left(\frac{b^2-ac}{a}\right) \sqrt{\pi/a}\end{aligned}$$

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

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

For brevity, let us define

$$\begin{aligned}a &:= \frac{\sigma^2+\tau^2}{2\sigma^2\tau^2}, \\ b &:= -\frac{\sigma^2\mu+\tau^2 y}{2\sigma^2\tau^2}, \\ c &:= \frac{\sigma^2\mu^2+\tau^2 y^2}{2\sigma^2\tau^2}.\end{aligned}$$

Using simple algebraic operations, we obtain that $\text{(A)} = ax^2 + 2bx + c$.

(a) The marginal density of Y is given by

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

Using the formula discussed at the beginning of the solution, we can compute this integral by just putting in the values of a, b and c:

$$\begin{aligned}
f_Y(y) &= \int_{\mathbb{R}} f_{X,Y}(x,y)dx \\
&= \int_{\mathbb{R}} \frac{1}{2\pi\sigma\tau} e^{-(ax^2+2bx+c)}dx \\
&= \frac{1}{2\pi\sigma\tau} \exp\left(\frac{b^2-ac}{a}\right)\sqrt{\pi/a} \\
&\propto \exp\left(\frac{b^2-ac}{a}\right) \quad (a \text{ does not depend on } y)
\end{aligned}$$

Now we try to write $(b^2 - ac)/a$ as a complete square:

$$\begin{aligned}
\frac{b^2-ac}{a} &= \frac{1}{a}\left\{\left(\frac{\sigma^2\mu+\tau^2 y}{2\sigma^2\tau^2}\right)^2 - \frac{(\sigma^2+\tau^2)(\sigma^2\mu^2+\tau^2y^2)}{(2\sigma^2\tau^2)^2}\right\} \\
&= -\frac{1}{a}\cdot\frac{1}{(2\sigma^2\tau^2)^2}\cdot(\sigma^2\tau^2y^2 - 2\tau^2\sigma^2\mu y + \sigma^2\tau^2\mu^2) \\
&= -\frac{1}{a}\cdot\frac{\sigma^2\tau^2}{(2\sigma^2\tau^2)^2}\cdot((y-\mu)^2+\cdots) \\
&= -\frac{1}{2}\frac{1}{(\sigma^2+\tau^2)}\cdot((y-\mu)^2+\cdots)
\end{aligned}$$

Putting everything together yields

$$f_Y(y) \propto \exp\left[-\frac{1}{2}\frac{(y-\mu)^2}{(\sigma^2+\tau^2)}\right],$$

meaning that Y has a Gaussian 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.

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

By the discussion at the beginning of the solution, $f_{X,Y}(x,y) \propto \exp(-(ax^2+2bx+c))$. Since c does not depend on x (and y is considered as fixed/given), we can say :

$$f_{X|Y}(x|y) \propto \exp\left(-\tfrac{1}{2}\frac{\left(x+\frac{b}{a}\right)^2}{1/2a}\right)$$

So the mean would be $-b/a$ and the variance will be $1/2a$. Concretely:

$$\text{mean} = -\frac{b}{a} = \frac{\sigma^2\mu+\tau^2y}{\sigma^2+\tau^2} = \frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y$$

Note that the mean is a convex combination of μ and the observation y. Also

$$\text{variance} = \frac{1}{2a} = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$