Exercises
Introduction to Machine Learning
FS 2020

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s20`
Email questions to:
Jingwei Tang, jitang@inf.ethz.ch

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

Problem 1 (Sampling):

In this problem, we want to draw samples from a random variable X, knowing its cumulative distribution function (CDF). We will investigate two methods: inverse sampling and rejection sampling.

1. **Inverse Sampling**
   Suppose that X has distribution function F and that F is invertible. Which of the following statements are true?

   (a) If $U \sim$ Unif(0,1), then $F^{-1}(U)$ is always distributed as X.
   (b) If $U \sim \mathcal{N}(0,1)$, then $F^{-1}(U)$ is always distributed as X.
   (c) If $U \sim$ Unif(-1,1), then $F^{-1}(U)$ is always distributed as X.

   Solution:
   Only (a) is *true*. 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.
   Remark: Hence, if the CDF F is invertible, we have just found a way to sample from X. This is called *inverse sampling*.

   Rejection Sampling
   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 is 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. Answering the following questions should help you conclude that the above process generates a random variable of density f.

2. What is the value of $\mathbb{P}(U \leq \frac{f(T)}{\gamma(T)})$?

   (a) $1/\alpha^2$
   (b) $1/\alpha$

(c) $1/(\alpha(1-\alpha))$
(d) $1/(1-\alpha)^2$

Solution:
The correct answer is (b).
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:

$$\begin{aligned}\mathbb{P}\left(U \leq \frac{f(T)}{\gamma(T)}\right) &= \int_{\mathbb{R}} \mathbb{P}\left(U \leq \tfrac{f(T)}{\gamma(T)} \mid T = t\right) g(t)\,dt \\ &= \int_{\mathbb{R}} \mathbb{P}\left(U \leq \tfrac{f(t)}{\gamma(t)}\right) g(t)\,dt \\ &= \int_{\mathbb{R}} \frac{f(t)}{\gamma(t)} g(t)\,dt \\ &= \int_{\mathbb{R}} \frac{1}{\alpha} f(t)\,dt = \frac{1}{\alpha},\end{aligned}$$

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

3. What is the value of $P(T \leq x, U \leq \frac{f(T)}{\gamma(T)})$?

(a) $\frac{1}{\alpha}\int_{-\infty}^{x} f(t)\,dt$
(b) $\frac{1}{\alpha^2}\int_{-\infty}^{x} f(t)\,dt$
(c) $\frac{1}{\alpha(1-\alpha)}\int_{-\infty}^{x} f(t)\,dt$
(d) $\frac{1}{(1-\alpha)^2}\int_{-\infty}^{x} f(t)\,dt$

Solution:
The correct answer is (a).
Computing the following probability for $x \in \mathbb{R}$:

$$\begin{aligned}\mathbb{P}\left(T \leq x, U \leq \frac{f(T)}{\gamma(T)}\right) &= \int_{\mathbb{R}} \mathbb{P}\left(T \leq x, U \leq \tfrac{f(T)}{\gamma(T)} \mid T = t\right) g(t)\,dt \\ &= \int_{-\infty}^{x} \frac{f(t)}{\gamma(t)} g(t)\,dt = \frac{1}{\alpha}\int_{-\infty}^{x} f(t)\,dt.\end{aligned}$$

4. What is the value of $P(T \leq x | U \leq \frac{f(T)}{\gamma(T)})$?

(a) $\frac{1}{\alpha^2}\int_{-\infty}^{x} f(t)\,dt$
(b) $\int_{-\infty}^{x} f(t)\,dt$
(c) $\frac{1}{\alpha^2(1-\alpha)}\int_{-\infty}^{x} f(t)\,dt$
(d) $\frac{1}{(1-\alpha)}\int_{-\infty}^{x} f(t)\,dt$

Solution:
The correct answer is (b). Now, by the definition of conditional probability, we have

$$\mathbb{P}\left(T \leq x \mid U \leq \tfrac{f(T)}{\gamma(T)}\right) = \frac{\mathbb{P}\left(T \leq x, U \leq \frac{f(T)}{\gamma(T)}\right)}{\mathbb{P}\left(U \leq \frac{f(T)}{\gamma(T)}\right)} = \frac{\frac{1}{\alpha}\int_{-\infty}^{x} f(t)\,dt}{1/\alpha} = \int_{-\infty}^{x} f(t)\,dt.$$

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

5. What is the distribution of the number of rejections before X is generated?

   (a) Geometric distribution with parameter $1/\alpha$
   (b) Bernoulli distribution with parameter $1/\alpha$
   (c) Poisson distribution with parameter $1/\alpha$
   (d) Geometric distribution with parameter α
   (e) Bernoulli distribution with parameter α
   (f) Poisson distribution with parameter α

   Solution:
   The correct answer is (a).
   As computed in question 2., 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 with parameter $1/\alpha$.

Problem 2 (Multivariate normal distribution):

Recall the following fact about characteristic functions:

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

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

6. 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?

   (a) $\mathcal{N}(\mu, AA^\top)$
   (b) $\mathcal{N}(\mu, A)$
   (c) $\mathcal{N}(\mu, A^\top A)$
   (d) $\mathcal{N}(\mu, A^2)$

   Solution:
   The correct answer is (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 $Y \sim \mathcal{N}(\mu, AA^\top)$.

7. If B is an $r \times d$ matrix, what is the distribution of BY?

   (a) $\mathcal{N}(\mu, BAA^\top B^\top)$
   (b) $\mathcal{N}(B\mu, BAA^\top)$
   (c) $\mathcal{N}(B\mu, BAA^\top B^\top)$
   (d) $\mathcal{N}(\mu, BAA^\top B^\top)$

   Solution:
   The correct answer is (c).
   With the same argument as the previous question, one gets $BY \sim \mathcal{N}(B\mu, BAA^\top B^\top)$.

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

   Solution:
   The correct answer is **2**.

9. What is the **variance** of the conditional distribution of $Y = X_1 + X_2$ given $Z = X_1 - X_2 = 0$?

   Enter your answer up to two decimal places (rounded up or down). **Solution:**
   The correct answer is **6.67**.
   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

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

   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 $\frac{20}{3}$.

For $M \sim \mathcal{N}_d(0, I)$, we say that the random variable $V = \|M\|^2$ 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 $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.

We want to prove that $\frac{(n-1)}{\sigma^2} S^2$ has a chi-square distribution with $n-1$ degrees of freedom. Answer the following questions for a step-by-step guide through the proof.

10. Which of these vectors Y verifies $\|Y\|^2 = (n-1)S^2$?

    (a) $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$
    (b) $(X_1 - \bar{X}/2, \ldots, X_n - \bar{X}/2)$
    (c) $(X_1 - 2\bar{X}, \ldots, X_n - 2\bar{X})$
    (d) $(X_1 + \bar{X}, \ldots, X_n + \bar{X})$

    Solution:
    The correct answer is (a).
    Consider $Y = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$.

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

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

$$(n-1)S^2 = \|Y\|^2$$

11. Recall the following fact about orthogonal projections:

    Let v be a unit vector in $\mathbb{R}^d$. 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 $(v-u)/\|v-u\|$.

    Note that Y is a Gaussian vector, as it is a linear function of X, obtained by the transformation $I-vv^\top$. The transformation 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 $w = (1, 0, \ldots, 0)^T$. 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. What is the mean of Z?

    (a) $(0, 0, ..., 0)^\top$

    (b) Y

    (c) X

    (d) $(1, 0, \ldots, 0)^\top$

    Solution:
    The correct answer is (a).
    $\mu_Z = (I - uu^T)\mu_Y = (I - uu^T)(0, 0, ..., 0)^\top = (0, 0, ..., 0)^\top$

12. What is the **covariance matrix** of Z?

    (a) $\sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$

    (b) $\sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$

    (c) $\sigma^2 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$

    (d) a null matrix

    Solution:
    The correct answer is (a).
    The covariance matrix can be computed using the approach from Question 6:

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

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 $D = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ be the training data that you are given.

We want to predict y from x using linear regression, i.e. we want to predict y as $w^T x$ for some parameter vector $w \in \mathbb{R}^d$. (Without loss of generality, we assume that both x_i and y_i are centered, i.e. they have zero empirical mean. Hence we can neglect the otherwise necessary bias term.) We thus suggest minimizing the following loss

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

Let us introduce the $n \times d$ matrix $X \in \mathbb{R}^{n \times d}$ with the x_i as rows, and the vector $y \in \mathbb{R}^n$ consisting of the scalars y_i. Then, (1) can be equivalently re-written as $\arg\min_w \|Xw - y\|^2$. In this exercise, $\|\cdot\|$ is always the Euclidean norm. We refer to any w^* that attains the above minimum as a solution to the problem.

13. Assuming that $X^T X$ is invertible, the unique solution for w^* is:

    (a) $(X^\top X)^{-1}(X^\top Y)$
    (b) $(X^\top X)^{-1}(X^\top Y)(X^\top X)$
    (c) $X^\top Y$
    (d) $(X^\top X)^{-1} Y$

    Solution:
    The correct answer is (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 $\mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

14. Pick the true statements.

    (a) (1) always admits a unique solution if $n \geq d$
    (b) (1) does not admit a unique solution if $n < d$
    (c) (1) always admits a solution if $n \geq d$ and the columns of X are independent.
    (d) (1) admits a unique solution $\iff$ $X^\top X$ is invertible.
    (e) (1) always admits a unique solution if X is full rank.
    (f) (1) does not admit a solution if $n < d$

Solution:
The true statements are (b), (c), (d) and (e).

(a) (1) does not a unique solution if $n \geq d$ when the rows of X are dependent.

(b) The intuition behind this statement 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. We now mathematically formulate this intuition.
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.

(c) The columns of X being independent implies that the column rank of X is d. Since the row rank of a matrix is equal to the column rank, there are d independent equations to determine the d parameters in w The "linear system" $\mathbf{X}\mathbf{w} \approx \mathbf{y}$ is now completely determined.

(d) The truth of this statement can be inferred from Question 13.

(e) X, being a full rank matrix, implies that the "linear system" $\mathbf{X}\mathbf{w} \approx \mathbf{y}$ is now completely determined.

(f) Since the "linear system" $\mathbf{X}\mathbf{w} \approx \mathbf{y}$ is underdetermined, the system admits infinite solutions.

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

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

15. $\hat{R}_{ridge}$ is convex with respect to w if and only if for any $w \in \mathbb{R}^d$ its Hessian $D^2 \hat{R}_{ridge}(w) \in \mathbb{R}^{d \times d}$ is:

(a) invertible

(b) positive definite

(c) positive semi-definite

(d) its columns are linearly independent

Solution:
The correct answer is (c).
Because convex functions are closed under addition, we will show that each term in the objective is convex, from which the answer 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.

16. The solution ${w^*}_{ridge}$ to $\hat{R}_{ridge}$ is:

(a) $(X^\top X)^{-1}(X^\top Y)$

(b) $(X^\top X + \lambda I)^{-1}(X^\top Y)$

(c) $(X^\top X - \lambda I)^{-1}(X^\top Y)$

(d) $(X^\top X)^{-1}(X^\top Y + \lambda I)$

Solution:
The correct answer is (b).

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

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

17. Pick the true statements about ridge regression.

(a) (2) always admits a unique solution, even when (1) does not.

(b) $\|w^*{}_{ridge}\| \to 0$ as $\lambda \to \infty$

(c) If (1) has a solution, then $w^*{}_{ridge}$ converges to a solution of (1) as $\lambda \to 0$

(d) (2) admits a unique solution if and only if $n \geq d$ and the columns of X are independent.

Solution:
The true statements are (a), (b) and (c).

(a) 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 $\mathbf{X}^T\mathbf{X} + \lambda I$.

(b) 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 (**??**). 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 $(0, 0, \ldots, 0)$.

(c) Refer to (b).

(d) Since (2) always admits a unique solution, even when (1) does not, this statement is false.

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. $X \sim \mathcal{N}(\mu, \tau^2)$. Let Y be a random variable such that Y given $X = x$ is normally distributed with mean x and variance σ^2, i.e. $Y|_{X=x} \sim \mathcal{N}(x, \sigma^2)$.

18. The mean of $f_Y(\ y)$, the marginal distribution of Y, is:

(a) μ

(b) $\mu/2$

(c) $\mu + x$

(d) $\mu - x$

Solution:
The correct answer is (a).

19. The variance of $f_Y(\ y)$, the marginal distribution of Y, is:

(a) σ^2

(b) $\sigma^2 + \tau^2$

(c) $\sigma^2 - \tau^2$

(d) $(\sigma + \tau)^2$

Solution:
The correct answer is (b).

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

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

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

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

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

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

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 $\sigma^2 + \tau^2$.

20. Using Bayes' theorem, find the mean of $f_{X|Y}(x|y)$, the conditional distribution of X given $Y = y$.

    (a) $\frac{\mu+y}{\sigma^2+\tau^2}$
    (b) $\frac{\sigma^2}{\sigma^2+\tau^2}\mu + \frac{\tau^2}{\sigma^2+\tau^2}y$
    (c) $\frac{\sigma^2}{\sigma^2+\tau^2}\mu - \frac{\tau^2}{\sigma^2+\tau^2}y$
    (d) $\frac{\mu-y}{\sigma^2+\tau^2}$

    Solution:
    The correct answer is (b).

21. Find the variance of $f_{X|Y}(x|y)$, the conditional distribution of X given $Y = y$.

    (a) $\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$
    (b) $\frac{\sigma^2}{\sigma^2+\tau^2}$
    (c) $\frac{\tau^2}{\sigma^2+\tau^2}$
    (d) $\frac{1}{\sigma^2+\tau^2}$

    Solution:
    The correct answer is (a).
    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:

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

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

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