MATH3431 Machine Learning and Neural Networks III — Epiphany term 2024

# Exercise sheet

Lecturer: Georgios P. Karagiannis — georgios.karagiannis@durham.ac.uk

**Part** 1. **Elements of convex learning problems**

**Exercise 1.** ($\star$)Let $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(w) = g(< w, x > +y)$ or some $x \in \mathbb{R}^d$, $y \in \mathbb{R}$. Show that: If $g$ is convex function then $f$is convex function.

---

**Exercise 2.** ($\star$)Let functions $g_1$ be $\rho_1$-Lipschitz and $g_2$ be $\rho_2$-Lipschitz. Then, show that, $f$ with $f(x) = g_1(g_2(x))$ is $\rho_1\rho_2$-Lipschitz.

---

**Exercise 3.** ($\star$)Let $f : \mathbb{R}^d \to \mathbb{R}$ with $f(w) = g(\langle w, x\rangle + y)$ $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be a $\beta$-smooth function. Then show that $f$ is a $\left(\beta \|x\|^2\right)$-smooth.

**Hint::** You may use Cauchy-Schwarz inequality $\langle y, x\rangle \leq \|y\| \|x\|$

---

**Exercise 4.** ($\star$)Show that $f : S \to \mathbb{R}$ is $\rho$-Lipschitz over an open convex set $S$ if and only if for all $w \in S$ and $v \in \partial f(w)$ it is $\|v\| \leq \rho$.

**Hint::** You may use Cauchy-Schwarz inequality $\langle y, x\rangle \leq \|y\| \|x\|$

---

**Exercise 5.** ($\star$)Let $g_1(w), ..., g_r(w)$ be $r$ convex functions, and let $f(\cdot) = \max_{\forall j}(g_j(\cdot))$. Show that for some $w$ it is $\nabla g_k(w) \in \partial f(w)$ where $k = \arg\max_j(g_j(w))$ is the index of function $g_j(\cdot)$ presenting the greatest value at $w$.

---

**Exercise 6.** ($\star$)Consider the regression learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with predictor rule $h(x) = \langle w, x\rangle$ labeled by some unknown parameter $w \in \mathcal{W}$, loss function $\ell(w, (x, y)) = (\langle w, x\rangle - y)^2$, feature $x \in \mathcal{X}$, and target $y \in \mathbb{R}$. Let $\mathcal{W} = \mathcal{X} = \left\{\omega \in \mathbb{R}^d : |\omega| \leq \rho\right\}$ for some $\rho > 0$.

(1) Show that the resulting learning problem is Convex-Lipschitz-Bounded learning problem.
(2) Specify the parameters of Lipschitnzess.

---

**Exercise 7.** ($\star$)If $f$is $\lambda$-strongly convex and $u$ is a minimizer of $f$ then for any $w$

$$f(w) - f(u) \geq \frac{\lambda}{2} \|w - u\|^2$$

**Hint::** Use the definition, and set $\alpha \to 0$.

---

The following is given as a homework (Formative assessment 1)

**Exercise 8.** ($\star$) Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and $\beta$-smooth function.

(1) Show that for $v, w \in \mathbb{R}^d$

$$f(v) - f(w) \in \left( \langle \nabla f(w), v - w \rangle, \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} \|v - w\|^2 \right)$$

(2) Show that for $v, w \in \mathbb{R}^d$ such that $v = w - \frac{1}{\beta} \nabla f(w)$, it is

$$\frac{1}{2\beta} \|\nabla f(w)\|^2 \leq f(w) - f(v)$$

(3) Additionally assume that $f(x) > 0$ for all $x \in \mathbb{R}^d$. Show that for $w \in \mathbb{R}^d$,

$$\|\nabla f(w)\| \leq \sqrt{2\beta f(w)}$$

---

The following is given as a homework (Formative assessment 1)

**Exercise 9.** ($\star$)Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\lambda$-strongly convex function. Assume that $w^*$ is a minimizer of $f$ i.e.

$$w^* = \arg\min_w \{f(w)\}$$

Show that for any $w \in \mathbb{R}^d$ it holds

$$f(w) - f(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

**Hint:** Use the definition of $\lambda$-strongly convex function, properly rearrange it, and ...

---

**Exercise 10.** ($\star$)Show that the function $J(x; \lambda) = \lambda \|x\|^2$ is $2\lambda$-strongly convex

---

## Part 2. Stochastic learning

**Exercise 11.** ($\star$) Assume a Bayesian model

$$\begin{cases} z_i | w & \overset{\text{ind}}{\sim} f(z_i | w), \ i = 1, ..., n \\ w & \sim f(w) \end{cases}$$

and consider that our objective is the discovery of MAP estimate $w^*$ i.e.

$$w^* = \arg\min_{\forall w \in \Theta} \left( -\log\left(L_n\left(w\right)\right) - f\left(w\right)\right) = \arg\min_{\forall w \in \Theta} \left( -\sum_{i=1}^{n} \log\left(f\left(z_i|w\right)\right) - \log\left(f\left(w\right)\right)\right)$$

by using SGD with update

$$w^{(t+1)} = w^{(t)} + \eta_t \left( \frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log\left(f\left(z_j|w^{(t)}\right)\right) + \nabla_w \log\left(f\left(w^{(t)}\right)\right)\right)$$

for some randomly selected set $\mathcal{J}^{(t)} \subseteq \{1, ..., n\}^m$ of $m$ integers from 1 to $n$ via simple random sampling (SRS) with replacement. Show that

$$\mathrm{E}_{\mathcal{J}^{(t)} \sim \text{simple-random-sampling}} \left( \frac{n}{m} \sum_{j \in \mathcal{J}^{(t)}} \nabla_w \log\left(f\left(z_j|w^{(t)}\right)\right)\right) = \sum_{i=1}^{n} \nabla_w \log\left(f\left(z_i|w^{(t)}\right)\right)$$

---

**Exercise 12.** ($\star$) Let $\{v_t; t = 1, ..., T\}$ be a sequence of vectors. Consider an algorithm producing $\left\{w^{(t)}; t = 1, 2, 3, ...\right\}$ with

$$w^{(1)} = 0$$
$$w^{(t+1)} = w^{(t)} - \eta v_t$$

Show that

(1) it is

$$\langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \left( -\left\| w^{(t+1)} - w^* \right\|^2 + \left\| w^{(t)} - w^* \right\|^2 \right) + \frac{\eta}{2} \left\| v_t \right\|^2$$

(2) it is

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle = \frac{1}{2\eta} \sum_{t=1}^{T} \left( -\left\| w^{(t+1)} - w^* \right\|^2 + \left\| w^{(t)} - w^* \right\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \left\| v_t \right\|^2$$

(3) (continue ) it is

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\left\| w^* \right\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \left\| v_t \right\|^2$$

---

**Exercise 13.** ($\star$) Let $\{v_t; t = 1, ..., T\}$ be a sequence of vectors. Consider an algorithm producing $\left\{w^{(t)}; t = 1, 2, 3, ...\right\}$ with

$$w^{(1)} = 0$$
$$w^{\left(t+\frac{1}{2}\right)} = w^{(t)} - \eta v_t$$
$$w^{(t+1)} = \arg\min_{w \in \mathcal{H}} \left( \left\| w - w^{\left(t+\frac{1}{2}\right)} \right\| \right)$$

for $t = 1, ..., T$.

**Hint:** You can use the following Lemma

**(Projection Lemma):** Let $\mathcal{H}$ be a closed convex set and let $v$ be the projection of $w$ onto $\mathcal{H}$,i.e.

$$v = \arg\min_{x \in \mathcal{H}} \|x - w\|^2$$

then for every $u \in \mathcal{H}$ it is

$$\|v - u\|^2 \leq \|w - u\|^2$$

Show that

(1) it is

$$\langle w^{(t)} - w^*, v_t \rangle \leq \frac{1}{2\eta} \left( - \left\| w^{(t+1)} - w^* \right\|^2 + \left\| w^{(t)} - w^* \right\|^2 \right) + \frac{\eta}{2} \|v_t\|^2$$

(2) it is

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{1}{2\eta} \sum_{t=1}^{T} \left( - \left\| w^{(t+1)} - w^* \right\|^2 + \left\| w^{(t)} - w^* \right\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

(3) (continue ) it is

$$\sum_{t=1}^{T} \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|v_t\|^2$$

**Comment:** Above we show that Lemma **??** from "Handout **??**: Gradient descent" holds even when a projection step is included. Hence, even if a projection step is included after the update step of the recursion of GD algorithm or the SGD algorithm the analysis in Section **??** in "Handout **??**: Gradient descent" holds. Hence, even if a projection step is included after the update step of the recursion of SGD algorithm or the SGD algorithm the analysis in Section **??** in "Handout **??**: Stochastic gradient descent" holds.

---

The following is given as a homework (Formative assessment 2)

**Exercise 14.** ($\star$) [1]Consider the binary classification problem with inputs $x \in \mathcal{X}$ where $\mathcal{X} := \left\{ x \in \mathbb{R}^d : \|x\|_2 \leq L \right\}$ for some given value $L > 0$, target $y \in \mathcal{Y}$ where $\mathcal{Y} := \{-1, +1\}$, and prediction rule $h_w : \mathbb{R}^d \to \{-1, +1\}$ with

(1) $$h_w(x) = \text{sign}\left( w^\top x \right)$$

(2) $$= \text{sign}\left( \sum_{j=1}^{d} w_j x_j \right)$$

[1]We use standard notation

$$\text{sign}(\xi) = \begin{cases} -1, & \text{if } \xi < 0 \\ +1, & \text{if } \xi > 0 \end{cases}$$

$\pm 1$ means either $-1$ or $+1$, $\mathbb{R}_+ := (0, +\infty)$, and $\|x\|_2 := \sqrt{\sum_{\forall j} (x_j)^2}$ for the Euclidean distance.

Let the hypothesis class is

$$\mathcal{H} = \left\{ x \to w^\top x : \forall w \in \mathbb{R}^d \right\} \tag{3}$$

In other words, the hypothesis $h_w \in \mathcal{H}$ is parametrized by $w \in \mathbb{R}^d$, it receives an input vector $x \in \mathcal{X} := \mathbb{R}^d$ and it returns the label $y = \text{sign}\left(w^\top x\right) \in \mathcal{Y} := \{\pm 1\}$ where

$$\text{sign}\left(\xi\right) = \begin{cases} -1, & \text{if}\, \xi < 0 \\ +1, & \text{if}\, \xi > 0 \end{cases}$$

Consider a loss function $\ell : \mathbb{R}^d \to \mathbb{R}_+$ with

$$\ell\left(w, z = (x, y)\right) = \max\left(0, 1 - y w^\top x\right) + \lambda \left\| w \right\|_2^2 \tag{4}$$

for some given value $\lambda > 0$.

Assume there is available a dataset of examples $S_n = \{ z_i = (x_i, y_i)\, ;\, i = 1, ..., n \}$ of size $n$.

Do the following:

(1) Show that the function $f : \mathbb{R} \to \mathbb{R}_+$ with $f\left(x\right) = \max\left(0, 1 - x\right)$ is convex in $\mathbb{R}$; and show that the loss (4) is convex.

**Hint::** You may use Proposition **??** from Handout **??**: Elements of convex learning problems.

(2) Show that the loss $\ell\left(w, z\right)$ for $\lambda = 0$ (4) is $L$-Lipschitz (with respect to $w$) when $x \in \mathcal{X}$ where $\mathcal{X} := \left\{ x \in \mathbb{R}^d : \left\| x \right\|_2 \leq L \right\}$.

**Hint::** You may use the definition of Lipschitz function. Without loss of generality, you can consider any $w_1 \in \mathbb{R}^d$ and $w_2 \in \mathbb{R}^d$ such that $1 - y w_2^\top x \leq 1 - y w_1^\top x$, and then take cases $1 - y w_2^\top x > \text{or} < 0$ and $1 - y w_1^\top x > \text{or} < 0$ to deal with the max.

(3) Construct the set of sub-gradients $\partial f\left(x\right)$ for $x \in \mathbb{R}$ of the function $f : \mathbb{R} \to \mathbb{R}_+$ with $f\left(x\right) = \max\left(0, 1 - x\right)$. Show that the vector $v$ with

$$v = \begin{cases} 2\lambda w, & y w^\top x > 1 \\ 2\lambda w, & y w^\top x = 1 \\ -yx + 2\lambda w, & y w^\top x < 1 \end{cases}$$

is $v \in \partial_w \ell\left(w, z = (x, y)\right)$, aka a sub-gradient of $\ell\left(w, z = (x, y)\right)$ at $w$, for any $w \in \mathbb{R}^d$.

(4) Write down the algorithm of online AdaGrad (Adaptive Stochastic Gradient Descent) with learning rate $\eta_t > 0$, batch size $m$, and termination criterion $t > T_{\max}$ for some $T_{\max} > 0$ in order to discover $w^*$ such as

$$w^* = \arg \min_{\forall w : h_w \in \mathcal{H}} \left( \mathrm{E}_{z \sim g} \left( \ell\left(w, z = (x, y)\right) \right) \right) \tag{5}$$

The formulas in your algorithm shoud be implemented for the above learning problem and tailored to 1, 3, and 4.

(5) Use the R code given below in order to generate the dataset of observed examples $S_n = \{ z_i = (x_i, y_i) \}_{i=1}^n$ that contains $n = 10^6$ examples with inputs $x$ of dimension $d = 2$. Consider $\lambda = 0$. Use a seed $w^{(0)} = (0, 0)^\top$.

(a) By using appropriate values for $m$, $\eta_t$ and $T_{\max}$, code in R the algorithm you designed in part 4, and run it.

(b) Plot the trace plots for each of the dimensions of the generated chain $\left\{w^{(t)}\right\}$ against the iteration $t$.

(c) Report the value of the output $w^*_{\text{adaGrad}}$ (any type) of the algorithm as the solution to (5).

(d) To which cluster $y$ (i.e., $-1$ or 1) $x_{\text{new}} = (1, 0)^\top$ belongs?

```
# R code. Run it before you run anything else
#
data_generating_model <- function(n,w) {
z <- rep( NaN, times=n*3 )
z <- matrix(z, nrow = n, ncol = 3)
z[,1] <- rep(1,times=n)
z[,2] <- runif(n, min = -10, max = 10)
p <- w[1]*z[,1] + w[2]*z[,2] p <- exp(p) / (1+exp(p))
z[,3] <- rbinom(n, size = 1, prob = p)
ind <- (z[,3]==0)
z[ind,3] <- -1
x <- z[,1:2]
y <- z[,3]
return(list(z=z, x=x, y=y))
}
n_obs <- 1000000
w_true <- c(-3,4)
set.seed(2023)
out <- data_generating_model(n = n_obs, w = w_true)
set.seed(0)
z_obs <- out$z #z=(x,y)
x <- out$x
y <- out$y
#z_obs2=z_obs
#z_obs2[z_obs[,3]==-1,3]=0
#w_true <- as.numeric(glm(z_obs2[,3]~ 1+ z_obs2[,2],family = "binomial"
)$coefficients)
```

---

## **Part** 3. **Support Vector Machines**

---

The following is given as a homework (Formative assessment 3)

**Exercise 15.** ($\star\star$) Consider a training data set $\mathcal{D} = \{z_i = (x_i, y_i)\}_{i=1}^m$. Consider the Soft-SVM Algorithm that requires the solution of the following quadratic minimization problem (in a slightly modified but equivalent form to what we have discussed)

**Primal problem:**

$$(w^*, b^*, \xi^*) = \arg\min_{(w,b,\xi)} \left( \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^m \xi_i \right) \tag{6}$$

$$\text{subject to: } y_i \left( \langle w, x_i \rangle + b \right) \geq 1 - \xi_i, \ \forall i = 1, ..., m \tag{7}$$

$$\xi_i \geq 0, \ \forall i = 1, ..., m \tag{8}$$

for some user-specified fixed parameter $C > 0$.

(1) Specify the Lagrangian function $L$ associated to the above primal quadratic minimization problem, where $\{\alpha_i\}$ are the Lagrange coefficients wrt (7), and $\{\beta_i\}$ are the Lagrange coefficients wrt (8). Write down any possible restrictions on the Lagrange coefficients.

(2) Compute the dual Lagrangian function denoted as $\tilde{L}$ as a function of the Lagrange coefficients and the data points $\mathcal{D}$.

(3) Apply the Karush–Kuhn–Tucker (KKT) conditions to the above problem, and write them down.

(4) Derive and write down the dual Lagrangian quadratic maximization problem, along with the inequality and equality constraints, where you seek to find $\{\alpha_i\}$.

(5) Justify why the $i$-th point $x_i$ lies on the margin boundary when $\alpha_i \in (0, C)$ ( beware it is $\alpha_i \neq C$), and why the $i$-th point $x_i$ lies inside the margin when $\alpha_i = C$.

(6) Given optimal values $\{\alpha_i^*\}$ for Lagrangian coefficients $\{\alpha_i\}$ as they are derived by solving the dual Lagrangian maximization problem in part 4, derive the optimal values $w^*$ and $b^*$ for the parameters $w$ and $b$ as function of the support vectors. Regarding parameter $b$ it should be in the derived in the form

$$b^* = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left( y_i - \sum_{j \in \mathcal{S}} \alpha_j^* y_j \langle x_j, x_i \rangle \right)$$

where you determine the sets $\mathcal{M}$ and $\mathcal{S}$.

(7) Report the halfspace predictive rule $h_{w,b}(x)$ of the above problem as a function of $\alpha^*$ and $b^*$.

---

**Exercise 16.** ($\star\star\star$) *[This is the Relevance Vector Machine. The Exercise is taken from "Exercise Sheet: Bayesian Statistics" of the module "Bayesian Statistics III/IV (MATH3361/4071)" taught in "Michaelmas term 2021". The supplementary material in the box was mainly provided for the students who had not been introduced to the SVM ideas or the Kernel trick -so it can be skipped. Also, the supplementary material in the box is presented with a statistical (geostatistical modeling) motivation. The exercise requires basic knowledge of Bayesian statistical inference and in particular the use of Bayes theorem for the computation of the posterior as well as basis probability density*

*calculus. However, the exercise is a useful example of extending the SVM ideas to the Bayesian learning setting.]*

Regarding the statistical model: Long story, short (supplementary material)

Consider that we are interested in recovering the mapping

$$x \overset{\eta}{\longmapsto} \eta(x)$$

in the sense that $y \in \mathbb{R}$ is the response (output quantity) that depends on $x = (x_1, ..., x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$ which is the independent variable (input quantity) in a procedure; E.g.:,

- $y$: precipitation in log scale
- $x = (\text{longitude}, \text{latitude})$: geographical coordinates.

Consider a set of observed data $\{(y_i, x_i)\}_{i=1}^n$, which may be contaminated by additive noise of unknown variance; i.e.

$$y_i = \eta(x_i) + \epsilon_i,$$

where $\epsilon_i \overset{\text{IID}}{\sim} \text{N}\left(0, \sigma^2\right)$ and $\sigma^2 > 0$ is unknown. We wish to recover $\eta(x)$ by using the Tikhonov regularization on the functional space $\mathcal{H}$ such that

$$\eta = \arg\min_{\forall \tilde{\eta} \in \mathcal{H}} \left\{ \sum_{i=1}^{n} L\left(y_i - \tilde{\eta}\left(x_i\right)\right) + \lambda \left\|\tilde{\eta}\right\|_{\mathcal{H}}^2 \right\} \tag{9}$$

By assuming that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS), the solution to the above Ridge regularizes loss minimization problem is such that

$$\eta\left(x\right) = \beta_0 + \sum_{j=1}^{n} k(x, x_j)\beta_j = k(x)^\top \beta$$

where $k(x) = \left(1, k\left(x, x_1\right), ..., k\left(x, x_n\right)\right)^\top$, $k\left(x, x_j\right)$ is the reproducing kernel (such as $k_\phi\left(x, x_j\right) = \exp\left(-\phi \left\|x - x_j\right\|^2\right)$ for some known parameter $\phi > 0$), and $\beta \in \mathbb{R}^{n+1}$ is an unknown vector.

Consider the following Bayesian model[2]

$$\begin{cases} y|\beta, \sigma^2 & \sim \text{N}\left(K\beta, I\sigma^2\right) \\ \beta|\lambda & \sim \text{N}\left(0, D^{-1}\right), \quad D = (\lambda_0, \lambda_1, ..., \lambda_n) \\ \lambda_i & \overset{\text{iid}}{\sim} \text{d}\Pi\left(\lambda_i\right) \propto \lambda_i^{a-1} \exp\left(-b\lambda_i\right) \text{d}\lambda_i, \ \ \forall i = 1, ..., n \\ \sigma^2 & \sim \text{d}\Pi\left(\sigma^2\right) \propto \left(\sigma^2\right)^{c-1} \exp\left(-\frac{1}{\sigma^2} d\right) \text{d}\sigma^2 \\ \beta, \sigma^2 & \text{a priori independent} \end{cases}$$

[2]Dixit, A., & Roy, V. (2021). Posterior impropriety of some sparse Bayesian learning models. Statistics & Probability Letters, 171, 109039.

where $K$ is a known matrix with size $n \times (n+1)$ such that

$$K = \begin{bmatrix} 1 & k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}.$$

The quantities $a > 0$, $b > 0$, $c > 0$, $d > 0$, and $\phi > 0$ are considered as fixed.

(1) When $b = 0$, show that a necessary condition for a valid posterior inference is $a \in (-1/2, 0)$ for any choice of prior for $\tau$ (i.e. any choice of $(c, d)$).

(2) Let $P = K\left(K^\top K\right)^{-1} K^\top$. Show that (2a) and (2b) are sufficient conditions for the Bayesian model to lead to a valid posterior inference
   (a) if $a > 0$ and $b > 0$ , or
   (b) if $y^\top (I - P)\, y + 2d > 0$ and $c > -\frac{n}{2}$

(3) Does the the improper Uniform prior on the joint $\log(\lambda_i)$ and $\log\left(\sigma^2\right)$, i.e. $\pi\left(\log(\lambda_i), \log\left(\sigma^2\right)\right) \propto 1$, lead to a valid inference?

(4) Does the Jeffreys' prior $\pi(\lambda_i) \propto 1/\lambda_i$ lead to a valid inference?

**Hint-1::**

$$(y - K\beta)^\top (y - K\beta) + (\beta - \mu)^\top V^{-1} (\beta - \mu) = (\beta - \mu^*)^\top (V^*)^{-1} (\beta - \mu^*) + S^*;$$

$$S^* = \mu^\top V^{-1} \mu - (\mu^*)^\top (V^*)^{-1} (\mu^*) + y^\top y; \qquad V^* = \left(V^{-1} + K^\top K\right)^{-1}; \qquad \mu^* = V^* \left(V^{-1}\mu + K^\top y\right)$$

**Hint-2::** Sherman-Morrison-Woodbury formula:

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U \left(C^{-1} + V A^{-1} U\right)^{-1} V A^{-1}$$

**Hint-3::**

$$-\frac{y^\top y}{2\sigma^2} \leq -\frac{y^\top \left(I\sigma^2 + K D^{-1} K^\top\right)^{-1} y}{2} \leq -\frac{1}{2\sigma^2} y^\top (I - P)\, y$$

where $P = K\left(K^\top K\right)^{-1} K$.

**Hint-4::** It is given that $\int_{(0,\infty)} \frac{t^{-(a+1)}}{(\xi+t)^{1/2}} \mathrm{d}t < \infty$ if and only if $a \in (-1/2, 0)$.

---

**Exercise 17.** ($\star$) Students are encouraged to practice on the Exercises 6.1-6.19 from the textbook

- Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.

available from

- https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

The solutions are available from

- https://blackboard.durham.ac.uk/ultra/courses/_44662_1/outline/create/document?id=_1396738_1

---