Machine Learning	Prof. Dr. Mario Fritz
Sheet 5 (Solution)	***Tutorial 2022.06.02***

Exercise 1: Lagrange Dual Problem

Consider the optimization problem with variable $x \in \mathbb{R}$ given in Equation 1.

i) State the dual problem and verify that it is a concave maximization problem.

ii) Find the dual optimal value and dual optimal solution λ^*.

iii) Does strong duality hold?

$$\begin{array}{ll} \text{minimize} & x^2 + 1 \\ \text{subject to} & (x-2)(x-4) \leq 0 \end{array} \tag{1}$$

Solution:

i) Lagrange dual function:

$$\begin{aligned} g(\lambda) &= \inf_{x \in \mathbb{R}} L(x, \lambda) \\ &= \inf_{x \in \mathbb{R}} x^2 + 1 + \lambda(x-2)(x-4) \\ &= \inf_{x \in \mathbb{R}} (1+\lambda)x^2 - 6\lambda x + 8\lambda + 1 && \text{(Rearrange)} \\ &= \begin{cases} (1+\lambda)\left(\frac{3\lambda}{1+\lambda}\right)^2 - 6\lambda\left(\frac{3\lambda}{1+\lambda}\right) + 8\lambda + 1 & \text{if } \lambda > -1 \\ -\infty & \text{if } \lambda \leq -1 \end{cases} && \text{(Minimization of quadratic function)} \\ &= \begin{cases} \frac{-9\lambda^2}{1+\lambda} + 8\lambda + 1 & \text{if } \lambda > -1 \\ -\infty & \text{if } \lambda \leq -1 \end{cases} && \text{(Rearrange)} \end{aligned}$$

Dual problem:

$$\begin{array}{ll} \text{maximize} & \dfrac{-9\lambda^2}{1+\lambda} + 8\lambda + 1 \\ \text{subject to} & \lambda \geq 0 \end{array}$$

The dual problem is equivalent to maximizing a concave function (as $g''(\lambda) < 0$ in the domain $\mathbb{R}^+$) over convex set $\mathbb{R}^+$, thus it is a concave maximization problem.

ii) Taking the first derivative:

$$\begin{aligned} g'(\lambda) &= \frac{-18\lambda(1+\lambda) - (-9\lambda^2)}{(1+\lambda)^2} + 8 \\ &= \frac{-9\lambda^2 - 18\lambda}{(1+\lambda)^2} + 8 \\ &= \frac{-\lambda^2 - 2\lambda + 8}{(1+\lambda)^2} \end{aligned}$$

Solving for $g'(\lambda) \stackrel{!}{=} 0$ leads to candidate solutions $\lambda = 2$ or -4.

As $\lambda \geq 0$, we rule out the negative candidate and obtain the dual optimal solution $\lambda^* = 2$.

And thus, the dual optimal value $d^* = g(\lambda^*) = 5$

iii) To investigate the duality, we need to solve the primal problem.

We first solve the inequality constraint in the primal problem.

$$(x-2)(x-4) \leq 0 \implies 2 \leq x \leq 4$$

Also, we know the objective is a quadratic function that monotonically increasing for $x \geq 0$.

Thus the optimal solution is $x^* = 2$, and $p^* = (x^*)^2 + 1 = 5$.

Hence the strong duality holds as $d^* = p^*$.

Exercise 2: Inequality constraint

i) Express the dual problem of the primal problem given in Equation 2 with $c \neq 0$ in terms of the conjugate f^*.

ii) Explain why the dual problem you give is convex. Note that we do not assume f is convex.

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & f(x) \leq 0 \end{array} \tag{2}$$

Solution:

i) The Fenchel conjugate of function f is defined as:

$$f^*(v) := \sup_{x \in \mathbb{R}^n} \left(v^\top x - f(x) \right)$$

Lagrange dual function:

$$\begin{aligned} g(\lambda) &= \inf_{x \in \mathbb{R}^n} L(x, \lambda) \\ &= \inf_{x \in \mathbb{R}^n} c^\top x + \lambda f(x) \\ &= -\sup_{x \in \mathbb{R}^n} \left(-c^\top x - \lambda f(x) \right) \\ &= \begin{cases} -\lambda \sup_{x \in \mathbb{R}^n} \left(-\frac{1}{\lambda} c^\top x - f(x) \right) & \text{if } \lambda \neq 0 \\ -\infty & \text{if } \lambda = 0 \end{cases} \\ &= \begin{cases} -\lambda f^*(-\frac{c}{\lambda}) & \text{if } \lambda \neq 0 \\ -\infty & \text{if } \lambda = 0 \end{cases} \end{aligned}$$

Dual problem:

$$\begin{array}{ll} \text{maximize} & -\lambda f^*(-\dfrac{c}{\lambda}) \\ \text{subject to} & \lambda \geq 0 \end{array}$$

ii) $f^*(v)$ is a convex function of v, as it is pointwise supremum of linear functions.

$\lambda f^*(\frac{-c}{\lambda})$ is the perspective of f^*, thus it preserves the convexity. Thus $-\lambda f^*(\frac{-c}{\lambda})$ is concave.

The dual problem is to maximize a concave function (with convex constraints), and is thus a convex problem.

Exercise 3: KKT conditions

i) Derive the KKT conditions for the problem given in problem 3 with variable $\mathbf{X} \in \mathbf{S}^n$ ($n \times n$ symmetric matrix) and domain $\mathbf{S}^n_{++}$ ($n \times n$ symmetric positive-definite matrix). $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{s} \in \mathbb{R}^n$ are given with $\mathbf{s}^\top \mathbf{y} = 1$.

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(\mathbf{X}) - \log\det \mathbf{X} \\ \text{subject to} & \mathbf{X}\mathbf{s} = \mathbf{y} \end{array} \tag{3}$$

ii) Verify that the optimal solution is given by Equation 4.

$$\mathbf{X}^* = \mathbf{I} + \mathbf{y}\mathbf{y}^\top - \frac{1}{\mathbf{s}^\top\mathbf{s}}\mathbf{s}\mathbf{s}^\top \tag{4}$$

Solution:

i) Lagrangian:

$$L(\mathbf{X}, \boldsymbol{\nu}) = \mathbf{tr}(\mathbf{X}) - \log\det \mathbf{X} + \boldsymbol{\nu}^\top(\mathbf{X}\mathbf{s} - \mathbf{y})$$

KKT conditions:

1. $\mathbf{X} \succ 0, \quad \mathbf{X}\mathbf{s} = \mathbf{y}$ (primal constraints)
2. $\boldsymbol{\nu} \geq 0$ (dual constraints)
3. $\dfrac{\partial L(\mathbf{X}, \boldsymbol{\nu})}{\partial \mathbf{X}} = \mathbf{I} - (\mathbf{X}^{-1})^\top + \dfrac{\boldsymbol{\nu}\mathbf{s}^\top + \mathbf{s}\boldsymbol{\nu}^\top}{2} = 0$ (gradient should vanish)

Note that for the third term, i.e., $\frac{\partial}{\partial \mathbf{X}}\boldsymbol{\nu}^\top\mathbf{X}\mathbf{s}$, we use the fact that:

$$\mathbf{tr}(\mathbf{X}\boldsymbol{\nu}\mathbf{s}^\top) = \mathbf{tr}(\mathbf{s}^\top\mathbf{X}\boldsymbol{\nu}) = \mathbf{s}^\top\mathbf{X}\boldsymbol{\nu} = \mathbf{s}^\top\mathbf{X}^\top\boldsymbol{\nu} = \boldsymbol{\nu}^\top\mathbf{X}\mathbf{s} = \mathbf{tr}(\boldsymbol{\nu}^\top\mathbf{X}\mathbf{s}) = \mathbf{tr}(\mathbf{X}\mathbf{s}\boldsymbol{\nu}^\top)$$

Thus,

$$\begin{aligned} \frac{\partial}{\partial \mathbf{X}}\boldsymbol{\nu}^\top\mathbf{X}\mathbf{s} &= \frac{1}{2}\frac{\partial\left(\mathbf{tr}(\mathbf{X}\boldsymbol{\nu}\mathbf{s}^\top) + \mathbf{tr}(\mathbf{X}\mathbf{s}\boldsymbol{\nu}^\top)\right)}{\partial \mathbf{X}} \\ &= \frac{\boldsymbol{\nu}\mathbf{s}^\top + \mathbf{s}\boldsymbol{\nu}^\top}{2} \end{aligned}$$

ii) We know from the KKT condition 3. that

$$\mathbf{X}^{-1} = \mathbf{I} + \frac{\boldsymbol{\nu}\mathbf{s}^\top + \mathbf{s}\boldsymbol{\nu}^\top}{2} \tag{S3.1}$$

Multiplying both sides of equation S3.1 by $\mathbf{y}$, and use the primal condition $\mathbf{X}\mathbf{s} = \mathbf{y}$:

$$\mathbf{s} = \mathbf{X}^{-1}\mathbf{y} = \left(\mathbf{I} + \frac{\boldsymbol{\nu}\mathbf{s}^\top + \mathbf{s}\boldsymbol{\nu}^\top}{2}\right)\mathbf{y} \tag{S3.2}$$

Now using $\mathbf{s}^\top \mathbf{y} = 1$ and equation S3.2, we obtain:

$$\begin{aligned}
1 = \mathbf{s}^\top \mathbf{y} &= \mathbf{y}^\top \left(\mathbf{I} + \frac{\boldsymbol{\nu}\mathbf{s}^\top + \mathbf{s}\boldsymbol{\nu}^\top}{2} \right)^\top \mathbf{y} \\
&= \mathbf{y}^\top \left(\mathbf{I} + \frac{\boldsymbol{\nu}\mathbf{s}^\top + \mathbf{s}\boldsymbol{\nu}^\top}{2} \right) \mathbf{y} \\
&= \mathbf{y}^\top \left(\mathbf{y} + \frac{\boldsymbol{\nu}\mathbf{s}^\top\mathbf{y} + \mathbf{s}\boldsymbol{\nu}^\top\mathbf{y}}{2} \right) \\
&= \mathbf{y}^\top\mathbf{y} + \frac{\mathbf{y}^\top\boldsymbol{\nu} + \mathbf{y}^\top\mathbf{s}\boldsymbol{\nu}^\top\mathbf{y}}{2} \\
&= \mathbf{y}^\top\mathbf{y} + \mathbf{y}^\top\boldsymbol{\nu}
\end{aligned}$$

$$\text{i.e.} \qquad \boldsymbol{\nu}^\top\mathbf{y} = \mathbf{y}^\top\boldsymbol{\nu} = 1 - \mathbf{y}^\top\mathbf{y} \tag{S3.3}$$

Plugging equation S3.3 into S3.2, we obtain the expression of $\boldsymbol{\nu}$:

$$\boldsymbol{\nu} = -2\mathbf{y} + (1 + \mathbf{y}^\top\mathbf{y})\mathbf{s} \tag{S3.4}$$

Substituting the expression for $\boldsymbol{\nu}$ (S3.4) into equation S3.1, we have:

$$\begin{aligned}
\mathbf{X}^{-1} &= \mathbf{I} + \frac{1}{2}(-2\mathbf{y}\mathbf{s}^\top - 2\mathbf{s}\mathbf{y}^\top + 2(1 + \mathbf{y}^\top\mathbf{y})\mathbf{s}\mathbf{s}^\top) \\
&= \mathbf{I} + (1 + \mathbf{y}^\top\mathbf{y})\mathbf{s}\mathbf{s}^\top - \mathbf{y}\mathbf{s}^\top - \mathbf{s}\mathbf{y}^\top
\end{aligned}$$

We verify that this is indeed the inverse of $\mathbf{X}^*$ given in equation 4:

$$\begin{aligned}
&\left(\mathbf{I} + (1 + \mathbf{y}^\top\mathbf{y})\mathbf{s}\mathbf{s}^\top - \mathbf{y}\mathbf{s}^\top - \mathbf{s}\mathbf{y}^\top\right) \mathbf{X}^* \\
&\quad = \left(\mathbf{I} + (1 + \mathbf{y}^\top\mathbf{y})\mathbf{s}\mathbf{s}^\top - \mathbf{y}\mathbf{s}^\top - \mathbf{s}\mathbf{y}^\top\right) \left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top - \frac{1}{\mathbf{s}^\top\mathbf{s}}\mathbf{s}\mathbf{s}^\top \right) \\
&\quad = \left(\mathbf{I} + \mathbf{y}\mathbf{y}^\top - \frac{1}{\mathbf{s}^\top\mathbf{s}}\mathbf{s}\mathbf{s}^\top \right) + (1 + \mathbf{y}^\top\mathbf{y})(\mathbf{s}\mathbf{s}^\top + \mathbf{s}\mathbf{y}^\top - \mathbf{s}\mathbf{s}^\top) \\
&\qquad - (\mathbf{y}\mathbf{s}^\top + \mathbf{y}\mathbf{y}^\top - \mathbf{y}\mathbf{s}^\top) - (\mathbf{s}\mathbf{y}^\top + (\mathbf{y}^\top\mathbf{y})\mathbf{s}\mathbf{y}^\top - \frac{1}{\mathbf{s}^\top\mathbf{s}}\mathbf{s}\mathbf{s}^\top) \\
&\quad = \mathbf{I}
\end{aligned}$$

Finally, we verify that $\mathbf{X}^* \succ 0$:

$$\mathbf{X}^* = \mathbf{I} + \mathbf{y}\mathbf{y}^\top - \frac{1}{\mathbf{s}^\top\mathbf{s}}\mathbf{s}\mathbf{s}^\top = \left(\mathbf{I} + \frac{\mathbf{y}\mathbf{s}^\top}{\|\mathbf{s}\|_2} - \frac{\mathbf{s}\mathbf{s}^\top}{\mathbf{s}^\top\mathbf{s}} \right) \left(\mathbf{I} + \frac{\mathbf{y}\mathbf{s}^\top}{\|\mathbf{s}\|_2} - \frac{\mathbf{s}\mathbf{s}^\top}{\mathbf{s}^\top\mathbf{s}} \right)^\top$$

Exercise 4: Estimating covariance and mean

(Recall Exercise 4 in Sheet 3) We consider the problem of estimating the covariance matrix $\boldsymbol{\Sigma}$ and the mean $\boldsymbol{\mu}$ of a Gaussian probability density function as given in Equation 5 based on N independent samples $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N \in \mathbb{R}^n$.

i) We first consider the estimation problem when there are no additional constrains on $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ be the sample mean and covariance as defined in Equation 6. Show that the log-likelihood function given in Equation 7 can be expressed as in Equation 8 and use this expression to show that if $\hat{\boldsymbol{\Sigma}} \succ 0$, then the ML estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ are unique and given by the sample covariance and sample mean.

ii) The log-likelihood function includes a convex term ($-\log\det \boldsymbol{\Sigma}$) so it is not obviously concave. Show that $\mathcal{L}$ is concave, jointly in $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ in the region defined by $\boldsymbol{\Sigma} \preceq 2\hat{\boldsymbol{\Sigma}}$. This means we can use convex optimization to compute simultaneous ML estimates of $\boldsymbol{\Sigma}$ and μ, subject to convex constraints, as long as the constraints include $\boldsymbol{\Sigma} \preceq 2\hat{\boldsymbol{\Sigma}}$, i.e. the estimate $\boldsymbol{\Sigma}$ must not exceed twice the unconstrained ML estimate.

$$p(\mathbf{x} \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}) = (2\pi)^{-\frac{n}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{5}$$

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^\top\end{aligned} \tag{6}$$

$$\mathcal{L}(\boldsymbol{\Sigma}, \boldsymbol{\mu}) = -\frac{Nn}{2}\log(2\pi) - \frac{N}{2}\log\det\boldsymbol{\Sigma} - \frac{1}{2}\sum_{k=1}^{N}(\mathbf{x}_k-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) \tag{7}$$

$$\mathcal{L}(\boldsymbol{\Sigma}, \boldsymbol{\mu}) = \frac{N}{2}\left(-n\log(2\pi) - \log\det\boldsymbol{\Sigma} - \mathbf{tr}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) - (\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right) \tag{8}$$

Solution:

i) Comparing Equation 7 and 8, we need to show that:

$$\begin{aligned}
\sum_{k=1}^{N}(\mathbf{x}_k-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}) &= N\mathbf{tr}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) + N(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}) \\
\iff \mathbf{tr}\left(\sum_{k=1}^{N}(\mathbf{x}_k-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})\right) &= N\mathbf{tr}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) + N\mathbf{tr}\left((\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})\right) \\
\iff \mathbf{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{N}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^\top\right) &= N\mathbf{tr}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) + N\mathbf{tr}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top\right) \\
&= N\mathbf{tr}\left(\boldsymbol{\Sigma}^{-1}\left(\hat{\boldsymbol{\Sigma}} + (\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top\right)\right)
\end{aligned}$$

It is equivalent to prove:

$$\sum_{k=1}^{N}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^\top = N\left(\hat{\boldsymbol{\Sigma}} + (\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top\right)$$

Simply multiplying out brackets:

$$\begin{aligned}
\sum_{k=1}^{N}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^\top &= \sum_{k=1}^{N}\mathbf{x}_k\mathbf{x}_k^\top - N\boldsymbol{\mu}\hat{\boldsymbol{\mu}}^\top - N\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^\top + N\boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= \sum_{k=1}^{N}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^\top + N\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top - N\boldsymbol{\mu}\hat{\boldsymbol{\mu}}^\top - N\hat{\boldsymbol{\mu}}\boldsymbol{\mu}^\top + N\boldsymbol{\mu}\boldsymbol{\mu}^\top \\
&= N\hat{\boldsymbol{\Sigma}} + N(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top
\end{aligned}$$

Thus, we verify the equivalence of Equation 7 and 8.

Now, let's maximize $\mathcal{L}(\boldsymbol{\Sigma}, \boldsymbol{\mu})$ in Equation 8. While it is not a concave function of $\boldsymbol{\Sigma}$ in general, we know that the gradient should vanish at the global optimizer (but not conversely).

Setting the gradient w.r.t. $\boldsymbol{\Sigma}^{-1}$ $\left(\frac{\partial \mathcal{L}(\boldsymbol{\Sigma},\boldsymbol{\mu})}{\partial \boldsymbol{\Sigma}^{-1}} = 0 \implies \frac{\partial \mathcal{L}(\boldsymbol{\Sigma},\boldsymbol{\mu})}{\partial \boldsymbol{\Sigma}} = 0\right)$ and $\boldsymbol{\mu}$ to be 0, we obtain:

$$\frac{\partial \mathcal{L}(\boldsymbol{\Sigma},\boldsymbol{\mu})}{\partial \boldsymbol{\Sigma}^{-1}} = \boldsymbol{\Sigma}^\top - \left(\hat{\boldsymbol{\Sigma}} + (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top\right)^\top = 0$$
$$\frac{\partial \mathcal{L}(\boldsymbol{\Sigma},\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}) = 0$$

which has unique solution ($\hat{\boldsymbol{\Sigma}} \succ 0$ guarantees that $\boldsymbol{\Sigma}$ is nonsingular):

$$\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}} + (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top$$
$$\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$$

ii) We show that the function

$$f(\boldsymbol{\Sigma}) = -\log\det \boldsymbol{\Sigma} - \mathbf{tr}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}})$$

is concave in $\boldsymbol{\Sigma}$ for $0 \prec \boldsymbol{\Sigma} \prec 2\hat{\boldsymbol{\Sigma}}$. This will establish concavity of the log-likelihood function because the remaining term of $\mathcal{L}$ is concave in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

The gradient and Hessian of f are given by:

$$\nabla f(\boldsymbol{\Sigma}) = -\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$$
$$\nabla^2 f(\boldsymbol{\Sigma})[V] = \boldsymbol{\Sigma}^{-1}V\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}V\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}V\boldsymbol{\Sigma}^{-1}$$

where by $\nabla^2 f(\boldsymbol{\Sigma})[V]$ we mean:

$$\nabla^2 f(\boldsymbol{\Sigma})[V] = \left.\frac{d}{dt}\nabla f(\boldsymbol{\Sigma} + tV)\right|_{t=0}$$

We show that

$$\mathbf{tr}(V\nabla^2 f(\Sigma)[V]) = \left.\frac{d^2}{dt^2} f(\boldsymbol{\Sigma} + tV)\right|_{t=0} \leq 0$$

for all V. We have

$$\begin{aligned}\mathbf{tr}(V\nabla^2 f(\boldsymbol{\Sigma})[V]) &= \mathbf{tr}(V\boldsymbol{\Sigma}^{-1}V\boldsymbol{\Sigma}^{-1}) - 2\mathbf{tr}(V\boldsymbol{\Sigma}^{-1}V\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) \\ &= \mathbf{tr}\left((\boldsymbol{\Sigma}^{-1/2}V\boldsymbol{\Sigma}^{-1/2})^2(\mathbf{I} - 2\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2})\right) \\ &\leq 0\end{aligned}$$

for all V if

$$2\boldsymbol{\Sigma}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1/2} \succeq \mathbf{I}$$

i.e., $\boldsymbol{\Sigma} \preceq 2\hat{\boldsymbol{\Sigma}}$.

Exercise 5: Estimating mean and variance

Consider a random variable $x \in \mathbb{R}$ with density p, which is normalized, i.e. has zero mean and unit variance. Consider a random variable $y = \frac{x+b}{a}$ obtained by an affine transformation of x, where $a > 0$. The random variable y has mean $\frac{b}{a}$ and variance $\frac{1}{a^2}$. As a and b vary over the non-negative real numbers $\mathbb{R}_+$ and the real numbers $\mathbb{R}$, respectively, we generate a family of densities obtained from p by scaling and shifting, uniquely parametrized by mean and variance.

i) Show that if p is log-concave, then finding the ML estimate of a and b, given samples $y_1, ..., y_n$ of y is a convex problem.

ii) As an example, work out an analytical solution for the ML estimates of a and b, assuming p is a normalized Laplacian density $p(x) = \exp(-2|x|)$.

Solution:

i) Applying the rule of density transformation, the density of y is given by:

$$p_y(u) = ap(au - b)$$

The log-likelihood function is:

$$\log p_y(u) = \log a + \log p(au - b)$$

If p is log-concave, then the log-likelihood function is a concave function of a (it's sum of concave functions of a) and b.

Given n samples $y_1, ..., y_n$, the log-likelihood is:

$$\sum_{i=1}^{n} \log p_y(y_i) = n \log a + \sum_{i=1}^{n} \log p(ay_i - b)$$

ML estimation= maximizing a concave (log-likelihood) function over a convex set $(\mathbb{R}+, \mathbb{R})$, and is thus a convex problem.

ii) For the Laplace distribution, the log-likelihood is:

$$\sum_{i=1}^{n} \log p_y(y_i) = n \log a - 2 \sum_{i=1}^{n} |ay_i - b|$$

The ML estimates solve:

$$\text{maximize} \qquad n \log a - 2 \sum_{i=1}^{n} |ay_i - b| \tag{S5.1}$$

Let $c = b/a$, (S5.1) transforms to:

$$\text{maximize} \qquad n \log a - 2a \sum_{i=1}^{n} |y_i - c| \tag{S5.2}$$

Solving for a and c, we obtain that:

Maximize S5.2 w.r.t. c : $\qquad c = \text{median of } y$

Plugging in the solution of c into S5.2 and maximize w.r.t. a : $\qquad a = \dfrac{n}{2 \sum_{i=1}^{n} |y_i - c|}$

Exercise 6: Robust linear classification

Consider the robust linear classification problem given in problem 9 where we seek an affine function $f(x) = \mathbf{w}^\top \mathbf{x} - b$ that separates the two sets of points $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and $\{\mathbf{y}_1, ..., \mathbf{y}_M\}$. This means that $\mathbf{w}^\top \mathbf{x}_i - b > 0$ for $i = 1, ..., N$ and $\mathbf{w}^\top \mathbf{y}_j - b < 0$ for $j = 1, ..., M$.

$$
\begin{aligned}
&\text{maximize } t \\
&\text{subject to } \mathbf{w}^\top \mathbf{x}_i - b \geq t, \;\; i = 1, ..., N \\
&\qquad\qquad\;\; \mathbf{w}^\top \mathbf{y}_i - b \leq -t, \;\; i = 1, ..., M \\
&\qquad\qquad\;\; ||\mathbf{w}||_2 \leq 1
\end{aligned} \tag{9}
$$

i) Show that the optimal value t^* is positive if and only if the two sets of points can be linearly separated. When the two sets of points can be linearly separated, show that the inequality $||\mathbf{w}||_2 \leq 1$ is tight, i.e., we have $||\mathbf{w}^*||_2 = 1$ for the optimal $\mathbf{w}^*$.

ii) Using the change of variables $\tilde{\mathbf{w}} = \frac{\mathbf{w}}{t}, \tilde{b} = \frac{b}{t}$, prove that problem 9 is equivalent to the quadratic program given in 10.

$$
\begin{aligned}
&\text{minimize } ||\tilde{\mathbf{w}}||_2 \\
&\text{subject to } \tilde{\mathbf{w}}^\top \mathbf{x}_i - \tilde{b} \geq 1, \;\; i = 1, ..., N \\
&\qquad\qquad\;\; \tilde{\mathbf{w}}^\top \mathbf{y}_i - \tilde{b} \leq -1, \;\; i = 1, ..., M
\end{aligned} \tag{10}
$$

Solution:

i) "$\Longrightarrow$":

As $t^* > 0$, we have, for all $\mathbf{x}_i$, $\mathbf{y}_i$:

$$\mathbf{w}^{*\top} \mathbf{x}_i \geq t^* + b^* > b^* > b^* - t^* \geq \mathbf{w}^{*\top} \mathbf{y}_i$$

Hence, $\mathbf{w}^*$ and b^* define a separating hyperplane.

"$\Longleftarrow$": If $\mathbf{w}$ and b define a separating hyperplane, then there is a positive t satisfying the constraints, and thus the optimal value of $t^* \geq t$ is positive.

Next, we prove by contraposition that $||\mathbf{w}^*||_2 = 1$ for the optimal $\mathbf{w}^*$.

Let ($\mathbf{w}_1$ (with $\|\mathbf{w}_1\|_2 < 1$), b_1, t_1) be a feasible solution of the problem, then we have:

$$
\begin{aligned}
\mathbf{w}_1^\top \mathbf{x}_i - b_1 \geq t_1 \quad &\Longleftrightarrow \quad \frac{\mathbf{w}_1^\top}{\|\mathbf{w}_1\|_2} \mathbf{x}_i - \frac{b_1}{\|\mathbf{w}_1\|_2} \geq \frac{t_1}{\|\mathbf{w}_1\|_2} \qquad \forall i = 1, ..., N \\
\mathbf{w}_1^\top \mathbf{y}_i - b_1 \leq -t_1 \quad &\Longleftrightarrow \quad \frac{\mathbf{w}_1^\top}{\|\mathbf{w}_1\|_2} \mathbf{y}_i - \frac{b_1}{\|\mathbf{w}_1\|_2} \leq -\frac{t_1}{\|\mathbf{w}_1\|_2} \qquad \forall i = 1, ..., M
\end{aligned}
$$

i.e., ($\frac{\mathbf{w}_1}{\|\mathbf{w}_1\|_2}$, $\frac{b_1}{\|\mathbf{w}_1\|_2}$, $\frac{t_1}{\|\mathbf{w}_1\|_2}$) is a feasible solution of the problem, and in particular, $\frac{t_1}{\|\mathbf{w}_1\|_2} > t_1$.

Hence, if $\|\mathbf{w}\|_2 < 1$ then it could not be the optimal $\mathbf{w}$.
$\Longrightarrow \quad ||\mathbf{w}^*||_2 = 1$ for the optimal $\mathbf{w}^*$.

ii) Suppose $\mathbf{w}, b, t$ are feasible in problem 9, with t > 0. Then $\tilde{\mathbf{w}}, \tilde{b}$ are feasible in the QP (10), with objective value $\|\tilde{\mathbf{w}}\|_2 = \|\mathbf{w}\|_2/t$.

Conversely, if $\tilde{\mathbf{w}}, \tilde{b}$ are feasible in the QP (10), then $t = 1/\|\tilde{\mathbf{w}}\|_2, \mathbf{w} = \tilde{\mathbf{w}}/\|\tilde{\mathbf{w}}\|_2, b = \tilde{b}/\|\tilde{\mathbf{w}}\|_2$, are feasible in problem 9, with objective value $t = 1/\|\tilde{\mathbf{w}}\|_2$.

References

[1] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.