Delft University of Technology
Faculty of Electrical Engineering, Mathematics, and Computer Science
Signal Processing Systems Section

EE 4530 APPLIED CONVEX OPTIMIZATION

30 January 2024, 13:30–16:30

Open book: copies of the book and the course slides allowed, as well as a one page cheat sheet. No other tools except a basic pocket calculator permitted.

Answer in English. Make clear in your answer how you reach the final result; the road to the answer is very important. Write your name and student number on each sheet and use one sheet per question.

Hint: Avoid losing too much time on detailed calculations, write down the general approach first.

Question 1 (13 points)

For each of the following sets or functions, explain if it is convex, concave, or neither convex nor concave. Prove it (by using the definition or some of the basic rules we have encountered in the course).

(a) The function $f(x) = \|x\|_p$ with $\text{dom} f \in \mathbb{R}^n$, for $p \geq 1$. Show it only through the epigraph formulation. **(2 points)**

(b) The function $f(X) = \text{tr}(XAX^\top B)$, with $X, A, B \in \mathbb{R}^{n\times n}$, and A, B positive semidefinite. **(2 points)**
Hint: a PSD matrix Z can be decomposed as $Z = Z^{1/2}Z^{1/2\top}$. Rewrite the trace as a norm.

(c) The binary cross entropy loss function:

$$f(w) = y\log(\sigma(w^\top x)) + (1-y)\log(1-\sigma(w^\top x))$$

with $w, x \in \mathbb{R}^n$, $y \in \mathbb{R}$. The function $\sigma(\cdot) : \mathbb{R} \to [0,1]$ is the sigmoid function, defined as $\sigma(z) = 1/(1+e^{-z})$. Use the second-order condition. **(3 points)**
Hint: $\sigma'(z) = \sigma(z)(1-\sigma(z))$

(d) The set $\mathcal{H} = \{H \in \mathbb{R}^{n\times n} | H = \sum_{k=0}^{K} h_k S^k, h_k \in \mathbb{R}\}$ where S is the delay matrix (the matrix containing 1s below the main diagonal and a 1 on the top-right corner). Do you know which set does $\mathcal{H}$ represent? **(1 point)**

(e) The set $\mathcal{S} = \{\theta | |e^{j\theta}| = 1\}$ for $\theta \in [0, 2\pi]$, where j is the imaginary unit. **(1 point)**

(f) The set of the gradients of a quadratic function, that is,:

$$\mathcal{G} = \{g \in \mathbb{R}^N | g = \nabla_x x^\top Ax \text{ for all } x \in \mathbb{R}^N\}$$

where $A \in \mathbb{R}^{N\times N}$ is negative semidefinite. **(2 points)**

(g) The set $\mathcal{S} = \{(x, f(x))\}$ for any $x \in \mathbb{R}$ and $f(x)$ a strictly convex function. **(2 points)**

Solution

(a) 2p It is convex. We can show this by showing that the epigraph of f is a convex set:

$$\mathrm{epi} f = \{(x,t) \in \mathbb{R}^{n+1} | \|x\|_p \leq t\}$$

Consider two distinct points $\bar{x}_1 = [x_1, t_1]^\top$ and $\bar{x}_2 = [x_2, t_2]^\top$, both belonging to epi$f$. We need to show that $\forall\ \theta \in [0,1]$ the point $\bar{x} := \theta \bar{x}_1 + (1-\theta)\bar{x}_2$ belongs to epi$f$. Expanding this expression we have:

$$\bar{x} = \theta \bar{x}_1 + (1-\theta)\bar{x}_2 = [\theta x_1 + (1-\theta)x_2, \theta t_1 + (1-\theta)t_2]$$

and:

$$\|\theta x_1 + (1-\theta)x_2\|_p \leq \theta \|x_1\|_p + (1-\theta)\|x_2\|_p \leq \theta t_1 + (1-\theta)t_2$$

from which we conclude that $\bar{x} \in \mathrm{epi} f$ and that f is convex.

(b) 2p It is convex. One way to show it is by rewriting the function $f(\cdot)$ as

$$f(X) = \mathrm{tr}(XA^{1/2}A^{1/2}X^\top B^{1/2}B^{1/2}) = \mathrm{tr}(B^{1/2}XA^{1/2}A^{1/2}X^\top B^{1/2}) = \|B^{1/2}XA^{1/2}\|_F^2$$

which is convex, since it is the composition of a convex function ($\|\cdot\|_F^2$) with an affine function of X.

(c) 3p It is concave. We can prove this by showing that the Hessian $\nabla^2 f(w) \preceq 0$, i.e., it is negative semidefinite. First, we compute the gradient $\nabla f(w)$. Denote with $z \in \mathbb{R}$ the value $z = w^\top x$; then we have by the chain rule that:

$$\frac{df}{dw_j} = \frac{df}{dz}\frac{dz}{dw_j}.$$

The factor $\frac{dz}{dw_j}$ is simply $\frac{dz}{dw_j} = x_j$. The factor $\frac{df}{dz}$ is:

$$\begin{aligned}
\frac{df}{dz} &= y\frac{\sigma'(z)}{\sigma(z)} + (1-y)\frac{-\sigma'(z)}{1-\sigma(z)} \\
&= y\frac{\sigma(z)(1-\sigma(z))}{\sigma(z)} + (1-y)\frac{-(\sigma(z)(1-\sigma(z))}{1-\sigma(z)} \\
&= y(1-\sigma(z)) + (1-y)(-\sigma(z) \\
&= y - \sigma(z)
\end{aligned}$$

So we have

$$\frac{df}{dw_j} = (y - \sigma(w^\top x))x_j$$

and the gradient is $\nabla f(w) = (y - \sigma(w^\top x))x$.

We compute the (jk)th entry of the Hessian through the second-order partial derivative:

$$\begin{aligned}
\frac{df}{dw_j w_k} &= -\frac{d\sigma(w^\top x))x_j}{dw_k} && (1) \\
&= -\frac{d\sigma(z))}{dz}\frac{dz}{dw_k}x_j && (2) \\
&= -\sigma(w^\top x)(1-\sigma(w^\top x))x_j x_k && (3)
\end{aligned}$$

It is then easy to see that:

$$\nabla^2 f(w) = -\sigma(w^\top x)(1-\sigma(w^\top x))xx^\top \tag{4}$$

which is negative semidefinite, since for any vector $y \in \mathbb{R}^n$, we have:

$$\begin{aligned} y^\top \nabla^2 f(w) y &= -\sigma(w^\top x)(1-\sigma(w^\top x))y^\top x x^\top y &(5)\\ &= -\sigma(w^\top x)(1-\sigma(w^\top x))(y^\top x)^2 &(6)\end{aligned}$$

which is always non-positive since both the terms $\sigma(w^\top x)$ and $(1-\sigma(w^\top x))$ are non-negative, and so is their product.

(d) 1p It is a convex set, which can be easily shown by using the definition of convexity of a set. The set $\mathcal{H}$ is the set of circulant matrices.

(e) 1p It is obviously convex, since for any θ it holds $|e^{j\theta}| = 1$.

(f) 2p It is convex. We show this in two ways.

Affine function. Consider the set $\mathcal{S} = \mathbb{R}^N$, i.e., the set of all Euclidean vectors of dimension N. Clearly $\mathcal{S}$ is convex. We know that the image of a convex set under an affine function is convex. Since the gradient of a quadratic function is linear (hence affine) in x, we have that the set of gradients is also a convex set. Formally:

$$\mathcal{S} = \mathbb{R}^N \text{ convex } \implies f(\mathcal{S}) = \{f(x) \mid x \in \mathcal{S}\} \text{ convex}$$

where $f(\mathcal{S}) = \{Ax \mid x \in \mathcal{S}\}$

By definition. The set $\mathcal{G}$ is the set of vectors

$$\mathcal{G} = \{Ax \mid x \in \mathbb{R}^N\}$$

Taking any two gradients $g_1 = Ax_1$ and $g_2 = Ax_2$, it holds:

$$\theta g_1 + (1-\theta)g_2 = A\theta x_1 + A(1-\theta)x_2 = A(\theta x_1 + (1-\theta)x_2) \in \mathcal{G}.$$

Since the vector $\theta x_1 + (1-\theta)x_2 \in \mathbb{R}^N$, its gradient belongs to $\mathcal{G}$.

(g) 2p It is not convex. It is the set of points constituting the graph of a function. The only possible way for the graph of a function to be convex is when the graph is an affine set (a straight line). But since the function f is strictly convex, it cannot be affine in any part of its domain.

Question 2 (9 points)

Assume we are given a dataset of N points $\mathcal{D} = \{x_1, \dots, x_N\}$, where $x_i \in \mathbb{R}^n$. We would like to reduce the dimensionality of our dataset $\mathcal{D}$ (which is now $N \times n$). To achieve this goal, we want to project our points into a lower dimensional subspace of dimension $d \ll n$, obtaining a projected dataset $\mathcal{D}_p = \{z_1, \dots, z_N\}$, where $z_i \in \mathbb{R}^d$. For now, we can assume that $d = 1$, so that $\mathcal{D}_p$ is of dimension $N \times 1$ (every point x_i is now described by the scalar z_i).

The problem statement now is the following: Find the $n \times 1$ projection vector u such that the variance of the points $z_i = u^T x_i$ which yield $\mathcal{D}_p$ is maximized. This way, the diversity of the points in $\mathcal{D}$ is retained as much as possible in $\mathcal{D}_p$. Note that there is a scaling ambiguity related to u; as a result we will assume that u has unit energy.

(a) Formulate the problem statement as an optimization problem and draw a visual sketch of the problem considering $n = 2$. **(2 points)**

(b) Rewrite the problem in matrix-vector form as a function of Σ_x, with $\Sigma_x = 1/N \sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})^T$ the (empirical) covariance matrix of the data in $\mathcal{D}$, with $\overline{x}$ the mean value $\overline{x} = 1/N \sum_{i=1}^{N} x_i$. Is the problem convex? Why? **(2 points)**

(c) Given that problem (b) results in a quadratic objective with a unitary ℓ_2-norm constraint, find the solution of the problem with the KKT conditions. **(3 points)**

(d) Derive also a projected gradient algorithm to reach this solution. **(2 points)**

Solution

(a) The projection of a point $x_i \in \mathbb{R}^n$ into the vector $u \in \mathbb{R}^n$ to obtain the point $z_i \in \mathbb{R}$ can be computed as $x_i^\top u$, if we consider u to be an orthonormal vector, i.e., $\|u\| = 1$. Denote with $\overline{x}$ the mean value $\overline{x} = 1/N \sum_{i=1}^{N} x_i$. It follows that the mean value $\overline{z}$ is

$$\overline{z} = \frac{1}{N} \sum_{i=1}^{N} z_i = \frac{1}{N} u^\top \sum_{i=1}^{N} x_i = u^\top \overline{x}. \tag{7}$$

The variance of the points $\{z_1, \dots, z_N\}$ is $1/N \sum_{i=1}^{N} (z_i - \overline{z})^2$. Since it depends on u, we have that the function we want to maximize is $f(u) = 1/N \sum_{i=1}^{N} (u^\top x_i - u^\top \overline{x})^2$. Thus the optimization problem we want to solve is:

$$\begin{aligned} u := \underset{u}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \left(u^\top (x_i - \overline{x}) \right)^2 \\ \text{s.t. } \|u\|_2 = 1 \end{aligned} \tag{8}$$

(b) We can rewrite $f(u)$ as:

$$f(u) = \frac{1}{N} \sum_{i=1}^{N} u^\top (x_i - \overline{x})(x_i - \overline{x})^\top u = u^\top \Sigma_x u, \tag{9}$$

where $\Sigma_x = 1/N \sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^\top$. Problem (8) is then:

$$\begin{aligned} u := \underset{u}{\operatorname{argmax}} \quad & u^\top \Sigma_x u \\ \text{s.t.} \quad & \|u\|_2 = 1 \end{aligned} \tag{10}$$

which is not convex due to the constraint set being not convex.

(c) Despite the problem being not convex, we know how to find a solution, since it is an eigenvalue problem. Moreover, in this case, it coincides with the principal component analysis (PCA) technique by considering only the first principal component. The Lagrangian is:

$$L(u, \lambda) = u^\top \Sigma_x u - \lambda(\|u\|_2^2 - 1) \tag{11}$$

where λ is the Lagrange multiplier associated to the equality constraint. The KKT conditions are:

- (primal feasibility) $\|u\|_2^2 = 1$
- (stationarity) $\nabla_u L(u, \lambda) = 0 \Rightarrow 2\Sigma_x u - 2\lambda u = 0 \Rightarrow \Sigma_x u = \lambda u$

which shows u needs to be an eigenvector of Σ_x (and λ the associated eigenvalue). Since we want to maximize $f(u)$, the vector u maximizing $f(u)$ is the eigenvector associated to the maximum eigenvalue of Σ_x.

(d) We can applying some projected gradient (ascent) steps. By considering an initial starting point $u^{(0)}$, the algorithm would have the following iterates for $k \geq 1$:

$$w^{(k)} = u^{(k-1)} + 2\alpha \Sigma_x u^{(k-1)} \tag{12}$$

$$u^{(k)} = \text{Proj}_{\|\cdot\|_2 = 1}(w^{(k)}) = \frac{w^{(k)}}{\|w^{(k)}\|_2} \tag{13}$$

Question 3 (8 points)

Assume we have a composition function $f(x) = g(h(x))$, where $h : \mathbb{R}^n \to \mathbb{R}$ is convex, and $g : \mathbb{R} \to \mathbb{R}$ is increasing and convex. Let us also assume that all these functions are twice differentiable.

(a) Is $f(x)$ convex? Why, why not? **(1 point)**

(b) Let p_f* and x_f* be the optimal value and optimal point for the problem $\min_x f(x)$. Similarly, let p_h* and x_h* be the optimal value and optimal point for the problem $\min_x h(x)$. What is the relationship between x_h* and x_f*? Is p_f* less than, greater than, or equal to p_h*? **(1 point)**

(c) Write down the gradient descent step for $\min_x f(x)$ and $\min_x h(x)$. Prove that with exact line search the iterates are the same. **(2 points)**

(d) Repeat part (c) for the Newton step. **(3 points)**

Hint: use the matrix inversion lemma:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \text{ , where}$$
$$A \in \mathbb{R}^{n\times n},\ C \in \mathbb{R}^{k\times k},\ U \in \mathbb{R}^{n\times k},\ V \in \mathbb{R}^{k\times n}, \text{ and in our case } k = 1$$

(e) Are the iterates the same if using backtracking line search? Explain. **(1 point)**

Solution

(a) 1p Convex: it is a composition of a non-decreasing convex and a convex function.

(b) 1p The optimal points are the same: due to the fact that g is increasing, the composition function cannot take a smaller value at any other point than at the optimal (minimum) point of inner function. The optimal values of course can be different, but without knowing g we cannot tell which is larger.

(c) 2p The gradient step for $h(x)$ is $-\nabla h(x)$. The gradient step for $f(x)$ is $-g'(h(x))\nabla h(x)$, where $g'(h(x))$, i.e. the first derivative is a positive scalar. Therefore, the gradient step moves in the same direction in both cases. The exact line search will find the same optimal point along this search direction, therefore, the iterates are equal.

(d) 3p The Newton step for $h(x)$ is

$$-\nabla^2 h(x)^{-1}\nabla h(x). \tag{14}$$

.

The Hessian for f is:

$$g''(h(x))\nabla h(x)\nabla h(x)^T + g'(h(x))\nabla^2 h(x) \tag{15}$$

Therefore, the Newton step for f is (up to a scaling):

$$-(a\nabla h(x)\nabla h(x)^T + \nabla^2 h(x))^{-1}\nabla h(x), \tag{16}$$

where $a = g''(h(x))/g'(h(x))$. Using the matrix inversion lemma, this can be written as

$$-\nabla^{-2}h(x)\nabla h(x)+\nabla^{-2}h(x)\nabla h(x)(a^{-1}+\nabla h(x)^T\nabla^{-2}h(x)\nabla h(x))^{-1}\nabla h(x)^T\nabla^{-2}h(x)\nabla h(x), \tag{17}$$

which can be simplified as

$$(-1+b)\nabla^{-2}h(x)\nabla h(x), \tag{18}$$

where $b = (a^{-1} + \nabla h(x)^T\nabla^{-2}h(x)\nabla h(x))^{-1}\nabla h(x)^T\nabla^{-2}h(x)\nabla h(x)$. Again, this is a positive multiple of the Newton step for h, therefore, the iterates are the same.

(e) 1p Backtracking line search starts with a given step size which is iteratively reduced until a stopping condition is satisfied. As such, it depends on the norm (and not only the direction) of the search direction. Therefore, the iterates will be different.

Question 4 (10 points)

Given $x_j \in \mathbb{R}^n$ and $y_j \in \{-1, 1\},\ j = 1, \ldots, m$, find the hyperplane $a^T x + b = 0$, with $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, that separates the data points x_j with $y_j = -1$ from the data points x_j with $y_j = 1$. To find such a separating hyperplane, the following so-called support vector machine (SVM) problem can be solved:

$$(a^\star, b^\star, \epsilon^\star) = \arg\min_{a,b,\epsilon} \frac{1}{2}\|a\|^2 + c\sum_{j=1}^{m} \epsilon_j \tag{19}$$

$$\text{subject to } y_j(a^T x_j + b) \geq 1 - \epsilon_j,\ j = 1, \ldots, m \tag{20}$$

$$\epsilon_j \geq 0,\ j = 1, \ldots, m \tag{21}$$

where the variables ϵ_j are included to accommodate data points x_j that are not linearly separable.

(a) Give the KKT conditions for this problem. **(3 points)**

(b) Prove that the optimal direction of the hyperplane is given by $a^* = \sum_{j=1}^{m} \mu_j^* y_j x_j$ where $\mu^* = [\mu_1^*, \ldots, \mu_m^*]^T$ is the Lagrange multiplier related to (20). **(2 points)**

(c) Show that the data points x_j that are either within the slab or poorly classified correspond to $\mu_j = c$. **(2 points)**

(d) Show that μ^* is obtained by solving

$$\text{minimize } \mu^T Q \mu - \mathbf{1}^T \mu \tag{22}$$

$$\text{subject to } y^T \mu = 0 \tag{23}$$

$$\mu_j \in [0, c], j = 1, \ldots, m \tag{24}$$

with $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^m$, $y = (y_1, \ldots, y_m)^T \in \mathbb{R}^m$, and $Q \in \mathbb{R}^{m \times m}$. Give an expression of the matrix Q. **(3 points)**

Solution

(a) The Lagrangian $L : \mathbf{R}^n \times \mathbf{R} \times \mathbf{R}^m \times \mathbf{R}^m$ associated with the problem is

$$L(a, b, \epsilon, \mu, \nu) = \frac{1}{2}\|a\|_2^2 + c\sum_{j=1}^{m} \epsilon_j + \sum_{j=1}^{m} \mu_j(1 - \epsilon_j - y_j(a^T x_j + b)) - \sum_{j=1}^{m} \nu_j \epsilon_j.$$

The KKT conditions are

$$\begin{aligned}
&1 - \epsilon_j^\star - y_j(x_j^T a^\star + b^\star) \leq 0 \quad \text{and} \quad \epsilon_j^\star \geq 0, j = 1, \ldots, m \\
&\mu_j^\star \geq 0 \quad \text{and} \quad \nu_j^\star \geq 0,\ j = 1, \ldots, m \\
&\mu_j^\star(1 - \epsilon_j^\star - y_j(x_j^T a^\star + b)) = 0 \text{ and } \nu_j^\star \epsilon_j^\star = 0,\ j = 1, \ldots, m \\
&\begin{bmatrix} a \\ 0 \\ c\mathbf{1} \end{bmatrix} + \sum_{j=1}^{m} \mu_j^\star \begin{bmatrix} -y_j x_j \\ -y_j \\ -\mathbf{e}_j \end{bmatrix} + \sum_{j=1}^{m} \nu_j^\star \begin{bmatrix} 0 \\ 0 \\ -\mathbf{e}_j \end{bmatrix} = 0,
\end{aligned}$$

where $\mathbf{e}_j \in \mathbf{R}^\mathbf{m}$ is the jth column of the identity matrix of size $m \times m$.

(b) We have

$$\frac{\partial L}{\partial a} = 0 \Rightarrow a^\star = \sum_{j=1}^{m} \mu_j^\star y_j x_j;$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow -\sum_{j=1}^{m} \mu_j y_j = 0$$

$$\frac{\partial L}{\partial \epsilon_j} = 0 \Rightarrow \mu_j = c - \nu_j;$$

Since $\nu_j \geq 0$ and $\mu_j \geq 0$, we have $\mu_j \in [0, c]$.

(c) For misclassified data points $\epsilon_j > 0$, then according to the KKT conditions, the corresponding $\nu_j = 0$. This means $\mu_j = c - \nu_j = c$.

(d) Using the above expressions in L, and writing the dual problem we obtain $\mu^\star$ by solving

$$\begin{aligned} \text{minimize } & \mu^T Q \mu - \mathbf{1}^T \mu && (25) \\ \text{subject to } & y^T \mu = 0 \\ & \mu_j \in [0, c], j = 1, \ldots, m && (26) \end{aligned}$$

where $Q = \mu^T \mathbf{diag}(y) X^T X \mathbf{diag}(y) \mu$ with $X = [x_1, \ldots, x_m]$.