Fall, 2024 CIS 515

Fundamentals of Linear Algebra and Optimization
Jean Gallier

Homework 6

December 2, 2024; Due December 11, 2024

Problem B1 (50 pts). Linear programming with box constraints is the following optimization problem:

$$\begin{aligned} &\text{minimize} && c^\top x \\ &\text{subject to} && Ax = b \\ &&& l \leq x \leq u, \end{aligned}$$

where A is an $m \times n$ matrix, $c, u, l, x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, with $l \leq u$ (which means that $l_i \leq u_i$, for $i = 1, \ldots, n$).

(1) (20 points) Prove that the dual of the above program is the following program:

$$\begin{aligned} &\text{maximize} && -\nu^\top b - \lambda_1^\top u + \lambda_2^\top l \\ &\text{subject to} && A^\top \nu + \lambda_1 - \lambda_2 + c = 0 \\ &&& \lambda_1 \geq 0, \quad \lambda_2 \geq 0. \end{aligned}$$

(2) (10 points) The primal problem in (1) can be reformulated by incorporating the constraints $l \leq x \leq u$ into the objective function by defining

$$f_0(x) = \begin{cases} c^\top x & \text{if } l \leq x \leq u \\ +\infty & \text{otherwise.} \end{cases}$$

The primal is reformulated as

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && Ax = b. \end{aligned}$$

Prove that the new dual function is given by

$$G(\nu) = \inf_{l \leq x \leq u} (c^\top x + \nu^\top (Ax - b)).$$

(3) (20 points) Given any real number $s \in \mathbb{R}$, let

$$s^+ = \max\{s, 0\}, \quad s^- = \max\{-s, 0\}.$$

Prove that for any fixed reals $s, \lambda, \mu \in \mathbb{R}$ with $\lambda \leq \mu$,

$$\inf_{\lambda \leq y \leq \mu} sy = \lambda s^+ - \mu s^-.$$

Hint. Consider the cases $s \geq 0$ and $s \leq 0$.

We extend the above operators to vectors $z \in \mathbb{R}^n$ componentwise by

$$z^+ = (z_1^+, \ldots . z_n^+), \quad z^- = (z_1^-, \ldots . z_n^-).$$

For any $w \in \mathbb{R}^n$, prove that

$$\inf_{l \leq x \leq u} x^\top w = l^\top w^+ - u^\top w^-.$$

Use the above to prove that

$$G(\nu) = -\nu^\top b + l^\top (A^\top \nu + c)^+ - u^\top (A^\top \nu + c)^-$$

and deduce that the dual program is the unconstrained problem

$$\text{maximize} \quad -\nu^\top b + l^\top (A^\top \nu + c)^+ - u^\top (A^\top \nu + c)^-$$

with respect to ν.

Problem B2 (120 pts). (1) Consider the determinant map, $f \colon \mathbf{M}_n(\mathbb{R}) \to \mathbb{R}$, given by

$$f(A) = \det(A), \quad A \in \mathbf{M}_n(\mathbb{R}).$$

For any matrix $B \in \mathrm{M}_n(\mathbb{R})$ (not necessarily invertible), let $\gamma \colon \mathbb{R} \to \mathbf{GL}(n, \mathbb{R})$ be the function given by

$$\gamma(t) = e^{tB}, \quad t \in \mathbb{R}.$$

Obviously, $\gamma(0) = I$. Geometrically, γ defines a curve in the group $\mathbf{GL}(n, \mathbb{R})$ passing through I at time $t = 0$. The function γ is differentiable, and by using the power series defining e^{tB} it is easily shown that

$$\gamma'(t) = Be^{tB},$$

so $\gamma'(0) = B$. In other words, the curve γ passes through I with velocity B. **You don't have to prove this fact** (Recall that when the domain space has dimension 1, we write $\gamma'(t) = d\gamma_1(t)$, the velocity vector at t.)

Let $g\colon \mathbb{R} \to \mathbb{R}$ be the function given by

$$g(t) = \det(\gamma(t)) = \det(e^{tB}), \quad t \in \mathbb{R}.$$

(1) Use the chain rule to prove that

$$d\det_I(B) = (\det \circ \gamma)'(0),$$

where $d\det_I$ is the derivative of the determinant function $\det\colon \mathrm{M}_n(\mathbb{R}) \to \mathbb{R}$ at I (the identity matrix).

(2) Prove that

$$d\det_I(B) = \mathrm{tr}(B),$$

the trace of B, for any matrix $B \in \mathrm{M}_n(\mathbb{R})$.

Hint. Use the fact that $\det(e^M) = e^{\mathrm{tr}(M)}$ for any matrix $M \in \mathrm{M}_n(\mathbb{R})$.

(3) Prove that

$$d\det_A(B) = \det(A)\mathrm{tr}(A^{-1}B),$$

for any $A \in \mathbf{GL}(n, \mathbb{R})$ and any matrix $B \in \mathrm{M}_n(\mathbb{R})$.

Hint. Find a curve $\gamma\colon \mathbb{R} \to \mathbf{GL}(n, \mathbb{R})$ such that $\gamma(0) = A$ and $\gamma'(0) = B$ and use the chain rule.

(4) Proposition 3.5 (Vol II) shows that for any continuous bilinear map $f\colon E_1 \times E_2 \to F$, for every $(a, b) \in E_1 \times E_2$, the derivative $\mathrm{D}f_{(a,b)}$ exists and is given by

$$\mathrm{D}f_{(a,b)}(u, v) = f(u, b) + f(a, v),$$

for all $(u, v) \in E_1 \times E_2$.

It can be shown (and you need not prove it, unless you decide to solve the extra credit problem) that for any continuous multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, for any $(a_1, \ldots, a_n) \in E_1 \times \cdots \times E_n$, the derivative $\mathrm{D}f_{(a_1,\ldots,a_n)}$ exists and is given by

$$\begin{aligned} \mathrm{D}f_{(a_1,\ldots,a_n)}(u_1, \ldots, u_n) &= f(u_1, a_2, a_3, \ldots, a_n) + f(a_1, u_2, a_3, \ldots, a_n) + \cdots \\ &\quad + f(a_1, a_2, a_3, \ldots, a_{n-1}, u_n) \\ &= \sum_{k=1}^{n} f(a_1, \ldots, a_{k-1}, u_k, a_{k+1}, \ldots, a_n), \end{aligned}$$

for all $(u_1, \ldots, u_n) \in E_1 \times \cdots \times E_n$.

By definition, for every $a = (a_1, \ldots, a_n) \in E_1 \times \cdots \times E_n$, the map $\mathrm{D}f_a$ is a continuous linear map from $E_1 \times \cdots \times E_n$ to F, namely, $\mathrm{D}f_a \in \mathcal{L}(E_1 \times \cdots \times E_n, F)$. The map $\mathrm{D}f\colon E_1 \times \cdots \times E_n \to \mathcal{L}(E_1 \times \cdots \times E_n, F)$ given by $a \mapsto \mathrm{D}f_a$ is linear and continuous for $n = 2$, but it is not linear for $n \geq 3$. It is also not multilinear for $n \geq 2$, but it can still be shown that it is continuous (you need not prove it, unless you decide to solve the extra credit problem).

Using the above facts, prove (quickly, this is easy) that for *any* matrix $A \in \mathrm{M}_n(\mathbb{R})$ and any matrix $B \in \mathrm{M}_n(\mathbb{R})$, the derivative $d\det_A$ exists and is given by

$$\begin{aligned} d\det_A(B) &= \det(B^1, A^2, A^3, \dots, A^n) + \det(A^1, B^2, A^3, \dots, A^n) + \cdots \\ &\quad + \det(A^1, A^2, A^3, \dots, A^{n-1}, B^n) \\ &= \sum_{k=1}^{n} \det(A^1, \dots, A^{k-1}, B^k, A^{k+1}, \dots, A^n), \end{aligned}$$

where $A^1, \dots, A^n$ are the columns of A and $B^1, \dots, B^n$ are the columns of B. Furthermore, the map $d\det\colon \mathrm{M}_n(\mathbb{R}) \to \mathcal{L}(\mathrm{M}_n(\mathbb{R}), \mathbb{R})$ given by $A \mapsto d\det_A$ is continuous.

Therefore, $d\det_A$ exists even if A is not invertible, but we would like to find a more "friendly" and more explicit expression for it. There such an explicit formula involving the adjugate matrix $\widetilde{A}$ of A from Section 6.4, Definition 6.9.

(5) (**Extra Credit 40 pts**) Prove that for any continuous multilinear map $f\colon E_1 \times \cdots \times E_n \to F$, for any $a = (a_1, \dots, a_n) \in E_1 \times \cdots \times E_n$, the derivative $\mathrm{D}f_{(a_1,\dots,a_n)}$ exists and is given by

$$\begin{aligned} \mathrm{D}f_{(a_1,\dots,a_n)}(u_1, \dots, u_n) &= f(u_1, a_2, a_3, \dots, a_n) + f(a_1, u_2, a_3, \dots, a_n) + \cdots \\ &\quad + f(a_1, a_2, a_3, \dots, a_{n-1}, u_n) \\ &= \sum_{k=1}^{n} f(a_1, \dots, a_{k-1}, u_k, a_{k+1}, \dots, a_n), \end{aligned}$$

for all $u = (u_1, \dots, u_n) \in E_1 \times \cdots \times E_n$.

Hint. Generalize the proof of Proposition 3.5 (Vol II).

Prove that $\mathrm{D}f$ (a map from $E_1 \times \cdots \times E_n$ to $\mathcal{L}(E_1 \times \cdots \times E_n, F)$) is continuous.

Hint. To prove that $\mathrm{D}f$ is continuous, first observe that $\mathrm{D}f$ is the sum of the n functions $(\mathrm{D}f)^1, \dots, (\mathrm{D}f)^n$, with $(\mathrm{D}f)^k$ from $E_1 \times \cdots \times E_n$ to $\mathcal{L}(E_1 \times \cdots \times E_n, F)$ given by

$$(\mathrm{D}f)^k_{(a_1,\dots,a_n)}(u_1, \dots, u_n) = f(a_1, \dots, a_{k-1}, u_k, a_{k+1}, \dots, a_n).$$

The function $(\mathrm{D}f)^k$ is independent of the variable a_k, so it is *not* multilinear, but its restriction to $E_1 \times \cdots \times E_{k-1} \times E_{k+1} \times \cdots \times E_n$ is $(n-1)$-multilinear, so if we can show that this restriction is continuous, then $(\mathrm{D}f)^k$ itself will be continuous. To simplify notation, write $\mathcal{E}_k = E_1 \times \cdots \times E_{k-1} \times E_{k+1} \times \cdots \times E_n$. We also use the notation $(\mathrm{D}f)^k$ to denote the restriction of $(\mathrm{D}f)^k$ to $\mathcal{E}_k$.

Show that the operator norm $\left\|(\mathrm{D}f)^k\right\|$ of the restriction of $(\mathrm{D}f)^k$ to $\mathcal{E}_k$ satisfies the inequality

$$\left\|(\mathrm{D}f)^k\right\| \leq \|f\|,$$

where $\|f\|$ is the norm of the multilinear map f (for norms of linear and multilinear maps, see Section 2.6, Vol. II).

(6) Prove that for *any* matrix $A \in \mathrm{M}_n(\mathbb{R})$, not necessarily invertible, there is a convergent sequence $(A_k)_{k\geq 1}$ of *invertible* matrices $A_k \in \mathrm{GL}(n, \mathbb{R})$ whose limit is A. To prove this, it is convenient to use the Frobenius norm or the operator 2-norm (the spectral norm). You need to construct a sequence of invertible matrices A_k such that

$$\lim_{k\mapsto\infty} \|A - A_k\| = 0.$$

Hint. Use a convenient factorization of A.

(7) Recall the definition of the *adjugate matrix* $\widetilde{A}$ of an $n \times n$ matrix A and the fact that if A is invertible, then by Proposition 6.7 (see Vol I),

$$A^{-1} = (\det(A))^{-1}\widetilde{A}.$$

Using the above, (3) is rewritten as

$$d\det_A(B) = \mathrm{tr}(\widetilde{A}B),$$

for any $A \in \mathbf{GL}(n, \mathbb{R})$ and any matrix $B \in \mathrm{M}_n(\mathbb{R})$. Use (6) to prove that

$$d\det_A(B) = \mathrm{tr}(\widetilde{A}B),$$

for *any* matrix $A \in \mathrm{M}_n(\mathbb{R})$ (not necessarily invertible) and any matrix $B \in \mathrm{M}_n(\mathbb{R})$.

(8) Let $\mathrm{GL}^+(n, \mathbb{R})$ be the subgroup of $\mathrm{GL}(n, \mathbb{R})$ consisting of all matrices A such that $\det(A) > 0$. It can be shown that this subgroup is open in $\mathrm{M}_n(\mathbb{R})$. Consider the function $\ell\colon \mathrm{GL}^+(n, \mathbb{R}) \to \mathbb{R}$ given by

$$\ell(A) = \log\det(A).$$

Prove that

$$d\ell_A(B) = \mathrm{tr}(A^{-1}B)$$

for all $A \in \mathrm{GL}^+(n, \mathbb{R})$ and all $B \in \mathrm{M}_n(\mathbb{R})$.

Remark: The function $\log\det$ is a *barrier* function used in convex optimization.

Problem B3 (20). Let A be an $n \times n$ real symmetric matrix, B an $n \times n$ symmetric positive definite matrix, and let $b \in \mathbb{R}^n$.

Prove that a necessary condition for the function J given by

$$J(v) = \frac{1}{2}v^\top Av - b^\top v$$

to have an extremum in $u \in U$, with U defined by

$$U = \{v \in \mathbb{R}^n \mid v^\top Bv = 1\},$$

is that there is some $\lambda \in \mathbb{R}$ such that

$$Au - b = \lambda Bu.$$

Hint. Express the definition of U as

$$U = \{v \in \mathbb{R}^n \mid \varphi(v) = 0\},$$

with

$$\varphi(v) = \frac{1}{2} - \frac{1}{2} v^\top B v.$$

Extra credit (20 points). Prove that there is a symmetric positive definite matrix S such that $B = S^2$. Prove that if $b = 0$, then λ is an eigenvalue of the symmetric matrix $S^{-1}AS^{-1}$.

Remark: If $b \neq 0$, solving for λ is a lot harder.

Problem B4 (10 pts). Verify the formula

$$(X^\top X + KI_n)^{-1} X^\top = X^\top (XX^\top + KI_m)^{-1},$$

where X is a real $m \times n$ matrix and $K > 0$. You may assume without proof that both $X^\top X + KI_n$ and $XX^\top + KI_m$ are invertible (because they are symmetric positive definite).

Problem B5 (40 pts). Consider the method of ridge regression to learn an affine function $f(x) = x^\top w + b$ instead of a linear function $f(x) = x^\top w$, where $b \in \mathbb{R}$. We have the following optimization program

Program (RR3):

$$\begin{aligned}
&\text{minimize} \quad \xi^\top \xi + K w^\top w \\
&\text{subject to} \\
&\qquad\qquad y - Xw - b\mathbf{1} = \xi,
\end{aligned}$$

with $y, \xi, \mathbf{1} \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$. Note that in Program (**RR3**) minimization is performed over ξ, w and b, but b is not penalized in the objective function.

(1) This problem can be solved directly by computing the quadratic functional $J(w, b) = \xi^\top \xi + K w^\top w$ in terms of w and b. Prove that

$$J(w, b) = \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} - 2 \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix} + y^\top y.$$

(2) Prove that the matrix

$$A = \begin{pmatrix} X^\top X + KI_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix}$$

is symmetric positive definite. You can either use an argument involving a Schur complement (Chapter 7 of Vol II, linalg-II), or proceed as follows.

Let g be the function given by

$$\begin{aligned} g(w, b) &= \begin{pmatrix} w^\top & b \end{pmatrix} \begin{pmatrix} X^\top X + K I_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} \\ &= w^\top (X^\top X + K I_n) w + 2 w^\top X^\top \mathbf{1}_m b + \mathbf{1}_m^\top \mathbf{1}_m b^2. \end{aligned}$$

Then A is symmetric positive definite iff $(w, b) \neq 0$ implies that $g(w, b) > 0$. Prove that if $w \neq 0$ and $b = 0$, then $g(w, 0) > 0$. If $b \neq 0$, for b fixed the function $w \mapsto g(w, b)$ is strictly convex because $X^\top X + K I_n$ is SPD, so it has a unique minimum obtained by setting the gradient $\nabla_w g$ to 0. Find the value w^* for which $\nabla_w g = 0$, and compute the corresponding minimum value $g(w^*, b)$. Prove that $g(w^*, b)$ is of the form $T b^2$ and compute T (T happens to be the Schur complement of $X^\top X + K I_n$ in A). Prove that $T > 0$, so that if $b \neq 0$, then $g(w^*, b) > 0$. Deduce that A is symmetric positive definite.

(3) Prove that the function $J(w, b)$ has a unique minimum obtained by setting its gradient to zero, which yields the system

$$\begin{pmatrix} X^\top X + K I_n & X^\top \mathbf{1}_m \\ \mathbf{1}_m^\top X & m \end{pmatrix} \begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} X^\top y \\ \mathbf{1}_m^\top y \end{pmatrix}. \tag{$*_1$}$$

Prove that the solution (w, b) of the above system agrees with the solutions given by the system associated with the dual, namely

$$\begin{pmatrix} X X^\top + K I_m & \mathbf{1}_m \\ \mathbf{1}_m^\top & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ b \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \tag{$*_2$}$$

with

$$w = X^\top \alpha.$$

TOTAL: 240 points + 60 extra credit.