Let $p$ be a probability distribution on the interval $[0, 1]$. Let the $k$-th moment of $p$ be the expected value

$$\mu_k = \mathbb{E}[x^k] = \int_0^1 x^k p(x) dx$$

Prove that the $n \times n$ matrix $H$ with entry $H_{ij} = \mu_{i+j}$ is positive semidefinite.

**answer** Let $y \in \mathbb{R}^d$. We have

$$\langle y, Hy \rangle = \sum_{i=1}^d \sum_{j=1}^d y_i H_{i,j} y_j$$

$$= \sum_i \sum_j y_i \mathbb{E}[x^i] y_j$$

$$= \mathbb{E} \left[ \sum_i \sum_j y_i x^i y_j \right]$$

by linearity of the expectation

$$= \mathbb{E} \left[ \left( \sum_i y_i x^i \right)^2 \right] \geq 0$$
Consider the function
\[ f(x) = \frac{1}{2} \sum_{i=1}^{d} x_i^2 + \frac{1}{4d^2} \left( \sum_{i=1}^{d} x_i \right)^4 - \frac{1}{d} \left( \sum_{i=1}^{d} x_i \right)^2 \]

(a) Compute the gradient and Hessian of \( f \)

\[ \frac{\partial f}{\partial x_j} = x_j + \frac{1}{d^2} \left( \sum_{i=1}^{d} x_i \right)^3 - \frac{2}{d} \sum_{i=1}^{d} x_i \]

\[ \frac{\partial^2 f}{\partial x_j \partial x_k} = \delta_{j,k} + \frac{3}{d^2} \left( \sum_{i=1}^{d} x_i \right)^2 - \frac{2}{d} \]

(b) Compute the set of points where \( \nabla f = 0 \). For each point, determine if it is a local minimum, local maximum, or global minimum.

Suppose \( \nabla f(x) = 0 \). A necessary condition is that \( \sum_{j=1}^{d} \frac{\partial f}{\partial x_j} = 0 \), that is

\[ \sum_{j=1}^{d} x_j + \frac{1}{d} \left( \sum_{i=1}^{d} x_i \right)^3 - \frac{2}{d} \sum_{i=1}^{d} x_i = 0 \]

i.e.

\[ \frac{1}{d} \left( \sum_{i=1}^{d} x_i \right) \left( \left( \sum_{i=1}^{d} x_i \right)^2 - d \right) = 0 \]

thus either \( \sum_i x_i = 0 \) or \( \sum_i x_i = d^\frac{1}{2} \). When \( \sum_i x_i = 0 \), the condition \( \frac{\partial f}{\partial x_j} = 0 \) becomes \( x_j = 0 \). When \( \sum_i x_i = d^\frac{1}{2} \), the condition \( \frac{\partial f}{\partial x_j} = 0 \) becomes \( x_j + \frac{1}{d^2} d^2 - \frac{2}{d} d^\frac{1}{2} = 0 \), i.e. \( x_j = d^{-\frac{1}{2}} \). So we have the two following equilibria:

- \( x = 0 \). The Hessian is then
  \[ (\nabla^2 f(0))_{j,k} = \delta_{j,k} - \frac{2}{d} \]
  so
  \[ \nabla^2 f(0) = I - \frac{2}{d} \mathbf{1} \]
  where \( \mathbf{1} \) is the matrix whose entries are all ones. \( \nabla^2 f \) has eigenvalues +1, e.g. eigenvector \[ \begin{pmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \]
  and -1, e.g. eigenvector \[ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \]. Thus 0 is a saddle point (neither a local minimum nor a local maximum).

- \( x = d^{-\frac{1}{2}} \mathbf{1} \), where \( \mathbf{1} \) is the \( d \times 1 \) vector of ones. The Hessian is PSD since
  \[ (\nabla^2 f(d^{-\frac{1}{2}} \mathbf{1}))_{j,k} = \delta_{j,k} + \frac{3}{d^2} d - \frac{2}{d} = \delta_{j,k} + \frac{1}{d} \]
  thus
  \[ x^T \nabla^2 f(d^{-\frac{1}{2}} \mathbf{1}) x = \sum_i x_i^2 + \frac{1}{d} \left( \sum_i x_i \right)^2 \geq 0 \]
therefore $d^{-\frac{1}{4}} \mathbf{1}$ is a local minimum. Furthermore, it is a global minimum since $f(d^{-\frac{1}{4}} \mathbf{1}) = \frac{1}{2} + \frac{1}{4} - 1 = -\frac{1}{4}$. But for all $x$,

$$f(x) = \frac{1}{2} \left( \sum_{i=1}^{d} x_i^2 - \frac{1}{d} \left( \sum_{i} x_i \right)^2 \right) + \frac{1}{4d^2} \left( \sum_{i} x_i \right)^4 - \frac{1}{2d} \left( \sum_{i} x_i \right)^2$$

and the first term is $\geq 0$ by Cauchy-Schwarz. Thus

$$f(x) \geq \frac{1}{4d^2} \left( \sum_{i} x_i \right)^4 - \frac{1}{2d} \left( \sum_{i} x_i \right)^2$$

$$= \left( \frac{1}{2d} \left( \sum_{i} x_i \right)^2 - \frac{1}{2} \right)^2 - \frac{1}{4}$$

$$\geq -\frac{1}{4}$$

so $-\frac{1}{4}$ is in fact a global minimum.

(c) Consider the gradient method with exact line search starting at the point $x_0 = [1, -1, 0, \ldots, 0]^T$. Determine to which point in $\{ x : \nabla f(x) = 0 \}$ the algorithm converges. Explain your reasoning.

Starting at $x_0 = [1, -1, 0, \ldots, 0]^T$, the gradient is

$$\nabla f(x_0) = x_0$$

and line search minimizes $f$ on the line $x_0 + \alpha x_0$, that is, minimizes the function $h(\alpha) = f(x_0 + \alpha x_0)$. The derivative of $h$ at $\alpha$ is the directional derivative of $f$,

$$h'(\alpha) = \langle x_0, \nabla f(x_0 + \alpha x_0) \rangle = \langle x_0, x_0 + \alpha x_0 \rangle = 2(1 + \alpha)$$

therefore the minimum is attained at $\alpha = -1$, and $x_1 = x_0 - x_0 = 0$. The algorithm stops since the gradient is zero at zero.
3 Suppose that $f$ is quadratic and of the form $f(x) = \frac{1}{2}x^TQx - b^Tx$, where $Q$ is positive definite.

(a) Show that the Lipschitz condition $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ is satisfied with $L$ equal to the maximal eigenvalue of $Q$.

We have $\nabla f(x) = Qx - b$. Since $Q$ is diagonalizable, there exists an orthogonal matrix $M$ and a diagonal matrix $\Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{pmatrix}$ (with the $\lambda_i \geq 0$) such that $Q = M^T \Lambda M$. Then we have

$$\|\nabla f(x) - \nabla f(y)\|^2 = \|Q(x - y)\|^2 = (x - y)^T M^T \Lambda M (x - y) = u^T \Lambda u$$

where $u = M(x - y)$. The last equality follows from the fact that an orthogonal matrix preserves the euclidean norm (if $M$ is orthogonal, $\|Mx\|^2 = x^T M^T x = x^T x = \|x\|^2$) thus $\|u\| = \|y - x\|$.

(b) Consider the gradient method $x_{k+1} = x_k - sD\nabla f(x_k)$, where $D$ is positive definite. Show that the method converges to $x^* = Q^{-1}b$ for every starting point $x_0$ if and only if $s \in (0, \frac{2}{L})$ where $L$ is the maximum eigenvalue of $D^{\frac{1}{2}}QD^{\frac{1}{2}}$.

Using $\nabla f(x) = Qx - b$, we have

$$x_{k+1} = x_k - sD\nabla f(x_k) = x_k - sD(Qx_k - b) = (I - sDQ)x_k + sDb$$

we have

$x$ is a fixed point $\iff x = (I - sDQ)x + sDb$

$\iff DQx - Db = 0$

$\iff x = Q^{-1}b$ since $D$ and $Q$ are positive definite (thus invertible)

writing $x^* = Q^{-1}b$

$$x_{k+1} - x^* = (I - sDQ)x_k + sDb - Q^{-1}b$$

$$= (I - sDQ)(x_k - Q^{-1}b)$$

$$= (I - sDQ)(x_k - x^*)$$

and with the change of variable $u_k = Q^{-\frac{1}{2}}x_k$, we have

$$u_{k+1} - u^* = D^{-\frac{1}{2}}(I - sDQ)(x_k - Q^{-1}b)$$

$$= D^{-\frac{1}{2}}(I - sDQ)(D^{\frac{1}{2}}u_k - D^{\frac{1}{2}}u^*)$$

$$= (I - sD^{\frac{1}{2}}QD^{\frac{1}{2}})(u_k - u^*)$$
and by induction
\[ u_k - Q^{-1}b = (I - sD^{\frac{1}{2}}QD^{\frac{1}{2}})^k(u_0 - D^{\frac{1}{2}}Q^{-1}b) \]
where \( I - sD^{\frac{1}{2}}QD^{\frac{1}{2}} \) is symmetric. Therefore the algorithm converges for any initial condition if and only if all eigenvalues of \( I - sD^{\frac{1}{2}}QD^{\frac{1}{2}} \) are less than one in absolute value, that is
\[ -I < I - sD^{-\frac{1}{2}}QD^{\frac{1}{2}} < I \]
i.e. \( 0 < sD^{\frac{1}{2}}QD^{\frac{1}{2}} < 2I \), i.e. \( s > 0 \) and \( s\lambda_{\text{max}}(D^{\frac{1}{2}}QD^{\frac{1}{2}}) < 2 \), i.e. \( s \in (0, \frac{\lambda_{\text{max}}(D^{\frac{1}{2}}QD^{\frac{1}{2}})}{2}) \).
A matrix $A$ is completely positive if there exists a matrix $X$ with only nonnegative entries satisfying $A = XX^T$.

(a) Show that the set of completely positive matrices forms a proper convex cone. Let $C$ be the set of such matrices. Let $A, B \in C$. Then there exist $X_A, X_B$ with non-negative entries such that $A = X_A X_A^T$ and $B = X_B X_B^T$. Then for all $\alpha, \beta > 0$,

$$(\alpha A | \beta B)(\alpha A | \beta B)^T = \alpha AA^T + \beta BB^T = \alpha A + \beta B$$

where $(\alpha A | \beta B)$ is a matrix with positive entries, therefore $\alpha A + \beta B \in C$, and $C$ is a cone. $C$ is proper since it is

- pointed: $A \in C \cap -C$ implies $A$ is both positive semidefinite and negative semidefinite, so $A = 0$, therefore $C \cap -C = \{0\}$

(b) Compute the dual cone.

The dual cone $C^*$ is defined as

$$(M, A) = \text{trace}(M A) = \sum_{i,j} M_{ij} A_{ij}.$$ Let $M = \{M : \forall x \in (\mathbb{R}_+)^d, x^T M x \geq 0\}$ (in particular $N$ is not assumed to be symmetric). The claim is that $C = M$. First, we observe that if $A = XX^T$ for some $X$ with nonnegative entries, then writing $X = [x_1 | \ldots | x_d]$ where each $x_i$ is a vector in the non-negative orthant,

$$\langle M, A \rangle = \text{trace}(M XX^T) = \text{trace}(X^T M X) = \text{trace} \left( \begin{pmatrix} x_1^T \\ \vdots \\ x_d^T \end{pmatrix} (M x_1 | \ldots | M x_d) \right) = \sum_{i=1}^d x_i^T M x_i$$

therefore

$$M \in C^* \iff \langle M, A \rangle \geq 0 \forall A \in C \iff \sum_i x_i^T M x_i \geq 0 \forall x_1, \ldots, x_d \in \mathbb{R}_+^d \iff x^T M x \geq 0 \forall x \in \mathbb{R}_+^d \iff M \in M$$

which proves the claim.
Let the support function of a set \( C \) be defined as
\[
S_C(x) = \sup_{y \in C} x^T y
\]

(a) Show that \( S_C \) is convex.

Fix \( x, x' \in \mathbb{R}^d \), and \( t \in [0, 1] \). We have
\[
S_C(tx + (1-t)x') = \sup_{y \in C} (tx + (1-t)x')^T y = \sup_{y \in C} tx^T y + \sup_{y' \in C} ty'^T y' = tS_C(x) + (1-t)S_C(x')
\]

(b) Show that \( S_{A+B} = S_A + S_B \)

Let \( A + B = \{x + y, x \in A, y \in B\} \). Let \( x \in \mathbb{R}^d \).
\[
S_{A+B}(x) = \sup_{y \in A+B} x^T y = \sup_{y_A \in A, y_B \in B} x^T (y_A + y_B) = \sup_{y_A \in A} x^T y_A + \sup_{y_B \in B} x^T y_B
\]

(c) Show that \( S_{A \cup B} = \max\{S_A, S_B\} \) since \( A \subseteq A \cup B \), we have
\[
\sup_{y \in A \cup B} x^T y \geq \sup_{y \in A} x^T y
\]
i.e. \( S_{A \cup B}(x) \geq S_A(x) \). Similarly, \( S_{A \cup B}(x) \geq S_B(x) \), thus \( S_{A \cup B} \geq \max(S_A, S_B) \). For the reverse inequality, by definition of the sup, there exists a sequence \((y_n)\) in \( A \cup B \) such that \( x^T y_n \to S_{A \cup B}(x) \). But for all \( n \), if \( y_n \in A \) then \( x^T y_n \leq S_A(x) \), and if \( y_n \in B \) then \( x^T y_n \leq S_B(x) \), so in both cases \( x^T y_n \leq \max(S_A(x), S_B(x)) \). Taking the limit, we have
\[
S_{A \cup B}(x) = \lim_{n} x^T y_n \leq \max(S_A(x), S_B(x))
\]

(d) Let \( B \) be closed and convex. Show that \( A \subseteq B \) if and only if \( S_A(x) \leq S_B(x) \) for all \( x \).

Suppose \( A \subseteq B \). Then for all \( x \),
\[
S_A(x) = \sup_{y \in A} x^T y \leq \sup_{y \in B} x^T y \quad \text{since the supremum is smaller on a smaller set}
\]
\[
= S_B(x)
\]

Conversely, if \( A \not\subseteq B \), then there exists \( x \in A \) with \( x \not\in B \). Since \( B \) is closed and convex, \( x \) and \( B \) can be separated (strictly) by a hyperplane, that is there exists a hyperplane \( H = \{u : u^T u = \alpha\} \) such that \( x^T u > \alpha \) and \( y^T u < \alpha \) for all \( y \in B \). In particular, \( S_B(u) = \sup_{y \in B} u^T y \leq \alpha \), and \( S_A(u) = \sup_{y \in A} u^T y \geq u^T x > \alpha \). Therefore \( S_A(u) > S_B(u) \).
(a) Suppose that \( f : \mathbb{R}^d \to \mathbb{R} \) is convex and concave. Show that \( f \) must be an affine function.

Since \( f \) is convex, for all \( x, y \in \mathbb{R}^d \) and \( t \in [0, 1] \)

\[
f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)
\]

the reverse inequality is also true since \( f \) is concave. Thus we have equality, and

\[
f(tx + (1-t)y) = tf(x) + (1-t)f(y) \quad \text{(1)}
\]

Let \( g(x) = f(x) - f(0) \). To show that \( f \) is affine, it suffice to show that \( g \) is linear, that is satisfies scalar multiplication and vector addition properties.

- Scalar multiplication: we want to show

\[
\forall x \in \mathbb{R}^d, \forall \alpha \in \mathbb{R}, \quad g(\alpha x) = \alpha g(x) \quad \text{(2)}
\]

  - Suppose \( \alpha \in [0, 1] \). Then applying (1) to \( x \) and \( 0 \) with \( t = \alpha \), we have

    \[
f(\alpha x) = \alpha f(x) + (1-\alpha)f(0)
    \]

    thus \( f(\alpha x) - f(0) = \alpha(f(x) - f(0)) \), i.e. \( g(\alpha x) = \alpha g(x) \).

  - Suppose \( \alpha > 1 \). Then applying (1) to \( \alpha x \) and \( 0 \) with \( t = \frac{1}{\alpha} \), we have

    \[
f\left(\frac{1}{\alpha} \alpha x\right) = \frac{1}{\alpha} f(\alpha x) + (1 - \frac{1}{\alpha})f(0)
    \]

    thus \( \alpha(f(x) - f(0)) = f(\alpha x) - f(0) \), i.e. \( g(\alpha x) = \alpha g(x) \). This proves (2) for all \( \alpha \geq 0 \).

  - Now suppose \( \alpha < 0 \). We first show \( g(-x) = -g(x) \). Apply (1) to \( -x \) and \( x \) with \( t = \frac{1}{2} \). Then

    \[
f(0) = \frac{1}{2} f(x) + \frac{1}{2} f(-x)
    \]

    thus \( f(x) - f(0) = -(f(x) - f(0)) \), i.e. \( g(-x) = -g(x) \). Now for any \( \alpha < 0 \)

    \[
g(\alpha x) = g(-\alpha(-x)) = -\alpha g(-x) = -\alpha(-g(x))
    \]

which proves the scalar multiplication property.

- Vector addition: let \( x, y \in \mathbb{R}^d \). We have

\[
g(x + y) = f(x + y) - f(0)
\]

\[
= f\left(\frac{1}{2} 2x + \frac{1}{2} 2y\right) - f(0)
\]

\[
= \frac{1}{2} f(2x) + \frac{1}{2} f(2y) - f(0) \quad \text{by (1)}
\]

\[
= \frac{1}{2}(g(2x) + g(2y))
\]

\[
= \frac{1}{2}(2g(x) + 2g(y)) \quad \text{by the scalar multiplication property}
\]

\[
= g(x) + g(y)
\]
(b) Suppose that \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is convex and bounded above. Show that \( f \) must be a constant function.

We prove the contrapositive. Suppose \( f \) is not constant, then there exist \( x \) and \( y \), distinct, with \( f(y) > f(x) \). Let \( \lambda > 1 \), and apply the convex inequality at \( x + \lambda(y - x) \) and \( x \) with \( t = \frac{1}{\lambda} \). Then

\[
f(\frac{1}{\lambda}(x + \lambda(y - x)) + (1 - \frac{1}{\lambda})x) \leq \frac{1}{\lambda} f(x + \lambda(y - x)) + (1 - \frac{1}{\lambda})f(x)
\]

i.e.

\[
f(y) \leq \frac{1}{\lambda} f(x + \lambda(y - x)) + (1 - \frac{1}{\lambda})f(x)
\]

rearranging,

\[
f(x + \lambda(y - x)) \geq \lambda(f(y) - f(x)) + f(x)
\]

letting \( \lambda \rightarrow \infty \), the RHS grows to infinity, thus \( f \) is unbounded.

(c) Suppose \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) is strongly convex and Lipschitz. Show no such \( f \) exists.

Suppose \( f \) is strongly convex. If \( f \) is differentiable at some \( x \), then there exists \( m > 0 \) such that for all \( y \)

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \| y - x \|^2
\]

but since \( f \) is Lipschitz, there exists \( L > 0 \) such that for all \( y \)

\[
|f(y) - f(x)| \leq L \| y - x \|
\]

in particular, we must have

\[
\langle \nabla f(x), y - x \rangle + \frac{m}{2} \| y - x \|^2 \leq f(y) - f(x) \leq L \| y - x \|
\]

for all \( y \), which is impossible: for example, take \( u \perp \nabla f(x) \) with \( \| u \| = 1 \), and \( y = x + \alpha u \). Then we need to have for all \( \alpha \)

\[
\langle \nabla f(x), tu \rangle + \frac{m}{2} \alpha^2 \| u \|^2 \leq L \alpha \| u \|
\]

i.e.

\[
\frac{m}{2} \alpha^2 \leq L \alpha
\]

for all \( \alpha \), which is impossible. Therefore we have a contradiction and no such function exists.

---

1 If \( f \) is nowhere differentiable (I don’t know whether a nowhere differentiable convex function exists...), then the strong convexity condition becomes: for all \( x, y \) and for all \( t \in (0, 1) \)

\[
f(y) \geq f(x) + \frac{f(x + t(y - x)) - f(x)}{t} + \frac{m}{2} (1 - t) \| y - x \|^2
\]

and by the Lipschitz condition,

\[
f(x + t(y - x)) - f(x) \geq -Lt \| y - x \|
\]

thus

\[
f(y) \geq f(x) - L \| y - x \| + \frac{m}{2} (1 - t) \| y - x \|^2
\]

for all \( t \). Taking \( t \to 0 \), we obtain \( f(y) \geq f(x) - L \| y - x \| + \frac{m}{2} \| y - y \|^2 \), so

\[
-L \| y - x \| + \frac{m}{2} \| y - y \|^2 \leq f(y) - f(x) \leq L \| y - x \|
\]

so \( \frac{m}{2} \| y - x \|^2 \leq 2L \| y - x \| \) for all \( y \), which is impossible.