CONVEX AND CONIC OPTIMIZATION
Homework 1

NILAVA METYA
nilava.metya@rutgers.edu
nm8188@princeton.edu

February 15, 2024

Problem 1

Let A be a real $m \times n$ matrix of rank r.

1. (a) Show that eigenvalues of $A^\top A$ are always nonnegative. (Hence singular values are well-defined as real, nonnegative scalars.)

   (b) Show that if A is symmetric then the singular values of A are the same as the absolute value of the eigenvalues of A.

2. Consider the optimization problem:
$$\min_{\substack{B \in \mathbb{R}^{m \times n} \\ \operatorname{rank} B \leq k}} \|A - B\|_2 .$$
Here $\|\cdot\|_2$ denotes the spectral norm of a matrix defined as $\|C\|_2 = \max_{\|x\|_2 = 1} \|Cx\|_2$. Show that the matrix $A_{(k)}$ is an optimal solution to the optimization problem above.

3. For $k = 40, 80, 120, 160$, use Matlab to compute $A_{(k)}$ as defined above. Report the value of $\left\|A - A_{(k)}\right\|_F$ in each case. (Include your code for this part and the next.)

4. Use the commands `subplot` and `imshow` (you need to import the matplotlib package in Python) to produce on the same figure the original image, as well as your compressed images $A_{(k)}$ for $k = 40, 80, 120, 160$. Label your subplots. In addition, produce two separate plots demonstrating (i) $\left\|A - A_{(k)}\right\|_F$ versus k, and (ii)"total savings" versus k. Total savings is to be interpreted as the answer to the question: How many fewer numbers do you need in order to store $A_{(k)}$ than you did to store A? Explain why this number is equal to $mn - (n + m + 1)k$. How much are you saving for $k = 160$?

5. Use the Matlab function `imwrite` or the Python command `image.save` to create two images from `imshow(A)` and `imshow(`$A_{(160)}$`)`. Can you tell them apart?

Solution

1. (a) Say $\lambda \in \mathbb{C}$ be an eigenvalue of $A^\top A$ with eigenvector $u \in \mathbb{C}^n$. Then $A^\top Au = \lambda u \implies \lambda \|u\|_2^2 = \overline{u}^\top A^\top Au = (\overline{Au})^\top Au = (Au)^\dagger Au = \|Au\|_2^2$. The non-negativity of $\|Au\|_2^2$ and $\|u\|_2^2$ proves the non-negativity of λ.
   Here, $\overline{u}$ is the coordinate-wise complex conjugate, and $(Au)^\dagger$ is the conjugate transpose. Note that we've used the fact that $\overline{A} = A$ because A has real entries.

   (b) Suppose A is an $n \times n$ real symmetric matrix. Then A can be diagonalized via an orthonormal basis, that is, there is a real matrix U such that $A = UDU^\top$, $U^\top U = UU^\top = \mathbf{1}_n$ and D is a diagonal matrix comprising of eigenvalues of A. One can rearrange the columns of U (the matrix of eigenvectors) such that their corresponding eigenvalues in D are in order of decreasing magnitude (i.e., decreasing after taking absolute value). So one can assume without loss of generality that D has entries $\lambda_1, \cdots, \lambda_n$ in this order where $|\lambda_1| \geq \cdots \geq |\lambda_n|$.
   So $A^\top A = UD^\top U^\top UDU^\top = UD^2U^\top$, whence the eigenvalues of $A^\top A$ are $\lambda_1^2 \geq \cdots \geq \lambda_n^2$. By what is mentioned in the assignment, σ_i is the square root of the i^{th} eigenvalue of $A^\top A$. In other words, $\sigma_i = \sqrt{\lambda_i^2} = |\lambda_i|$ in this case.

2. Start with the SVD $A = U\Sigma V^\top$ where U, V are orthogonal $m\times, n \times n$ (respectively) matrices. So $U^\top, V^\top$ are also orthogonal. Note that for any $m \times n$ matrix C we have

$$\left\|U^\top C\right\| = \max_{\|x\|=1} \left\|U^\top Cx\right\| = \max_{\|x\|=1} \sqrt{x^\top C^\top UU^\top Cx} = \max_{\|x\|=1} \|Cx\| = \|C\|$$

   and

$$\|CV\| = \max_{\|x\|=1} \|CVx\| = \max_{\|Vx\|=1} \|CVx\| \overset{[y:=Vx]}{=} \max_{\|y\|=1} \|Cy\| = \|C\|$$

   where we used the facts that V is invertible and that $\|Vx\| = \sqrt{x^\top V^\top Vx} = \sqrt{x^\top x} = \|x\|$. This proves that

$$\min_{\substack{B \in \mathbb{R}^{m\times n} \\ \operatorname{rank} B \leq k}} \|A - B\| = \min_{\substack{B \in \mathbb{R}^{m\times n} \\ \operatorname{rank} B \leq k}} \left\|\Sigma - U^\top BV\right\| \overset{D:=U^\top BV}{=} \min_{\substack{D \in \mathbb{R}^{m\times n} \\ \operatorname{rank} D \leq k}} \|\Sigma - D\|$$

   because $U^\top, V$ are invertible and left or right multiplication (respectively) by them preserve rank. We are minimizing over the feasible set

$$\Omega_k \coloneqq \left\{X \in \mathbb{R}^{m\times n} \;\middle|\; \operatorname{rank} X \leq k\right\}.$$

   Let $\{e_i \in \mathbb{R}^n\}_{i=1}^n, \{f_i \in \mathbb{R}^m\}_{i=1}^m$ be the standard basis vectors. Then $\Sigma e_i = \sigma_i f_i$ (reminder: Σ is Sigma and not summation) for each $1 \leq i \leq s \coloneqq \min\{m, n\}$. Recall that we had ordered $\sigma_1 \geq \cdots \geq \sigma_s$ ($\sigma_i = 0$ for $i > r = \operatorname{rank} A$). Pick any $D \in \mathbb{R}^{m\times n}$ with at most rank k, whence $\ker D$ has dimension at least $n - k$. This means that $W \coloneqq \operatorname{span}\{e_1, \cdots, e_{k+1}\}$ nontrivially intersects $\ker D$ (otherwise their direct sum would have dimension $\geq n + 1$ which is more than the dimension of the ambient space $\mathbb{R}^n$). Let $v \in W \cap \ker D$

with $v \neq 0$. Write $v = \sum_{i=1}^{k+1} c_i e_i$. So

$$\begin{aligned}
\|(\Sigma - D)v\|^2 &= \|\Sigma v\|^2 \\
&= \left\| \sum_{i=1}^{k+1} c_i \sigma_i f_i \right\|^2 \\
&= \sum_{i=1}^{k+1} c_i^2 \sigma_i^2 && [\because \{f_i\} \text{ are orthonormal}] \\
&\geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} c_i^2 \\
&= \sigma_{k+1}^2 \|v\|^2 .
\end{aligned}$$

This suggests that $\|\Sigma - D\| \geq \sigma_{k+1}$ as D varies over Ω_k from definition of the $2-$norm. And we see that for $D^* := D_k^*$ being the $m \times n$ matrix with $D_{ij}^* = \begin{cases} \sigma_i & \text{if } 1 \leq i = j \leq k \\ 0 & \text{otherwise} \end{cases}$, $\Sigma - D_k^*$ is the matrix with diagonal entries σ_i at position (i, i) for $i \geq k+1$, and 0 everywhere, thus having norm σ_{k+1} (Problem 3(3) says that this norm is the maximum singular value, which is σ_{k+1} in this case). This proves that the abovementioned lower bound is achieved.

Now, D^* looks like

$$D^* = \begin{bmatrix} \Sigma_{(k)} & 0_{k\times(n-k)} \\ 0_{(m-k)\times k} & 0_{(m-k)\times(n-k)} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_k \\ 0_{(m-k)\times k} \end{bmatrix} \begin{bmatrix} \Sigma_{(k)} & 0_{k\times(n-k)} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_k \\ 0_{(m-k)\times k} \end{bmatrix} \Sigma_{(k)} \begin{bmatrix} \mathbf{1}_k & 0_{k\times(n-k)} \end{bmatrix}$$

where $\mathbf{1}_k$ is the $k \times k$ identity matrix. But notice $U \begin{bmatrix} \mathbf{1}_k \\ 0_{(m-k)\times k} \end{bmatrix} = U_{(k)}$ and $V \begin{bmatrix} \mathbf{1}_k \\ 0_{(n-k)\times k} \end{bmatrix} = V_{(k)}$. Because we had transformed $D = U^\top B V$, the corresponding minimizer for the original problem (corresponding to D^*) is

$$B^* = U D^* V^\top = U_{(k)} \Sigma_{(k)} V_{(k)}^\top = A_{(k)}$$

which is what we wanted.

3. See code at end or the Jupyter notebook. The $A_{(k)}$ are computed in `[5]` of the code and stored in the `approx[]` array. Here're the Frobenius norms of the errors (which have computed in `[6]` of the code and stored in the `error[]` array):

$$\begin{aligned}
\left\|A - A_{(40)}\right\|_F &= 31.733288485394482 \\
\left\|A - A_{(80)}\right\|_F &= 19.016427727768185 \\
\left\|A - A_{(120)}\right\|_F &= 13.055854194249326 \\
\left\|A - A_{(160)}\right\|_F &= 9.511015173570117
\end{aligned}$$

4. See code at end or the Jupyter notebook. The corresponding part for comparing the compressed images in the code is `[7]`. `[8]` produces a plot for the Frobenius norms of the error vs k, and `[9]` produces a plot for total savings $(= mn - (m+n-1)k)$ vs k.

   We explain the total savings: Notice that fewer columns of U, Σ, V are used for constructing $A_{(k)}$. Namely, one only uses

   - m rows and k columns of U,

- n rows and k columns of V, and
- k rows and k columns of Σ, in which only the k diagonal entries are useful.

Thus $A_{(k)}$ is determined by the entries in $U_{(k)},$ (diagonal of) $\Sigma_{(k)},$ and $V_{(k)}$, and this number of necessary entries is $mk + k + nk = (m + n + 1)k$. Initially, A was described by mn reals. So total savings is the difference, namely $mn - (m + n + 1)k$. Total saving for $k = 160$ is $1075200 - (2081)160 = 742240$ which is approximately 69%.

5. A comparison has been shown in fig. 1. `[10]` of the code does this task. At first glance, one cannot compare at all. Only when one looks carefully, we can see the difference in sharpness of the two pictures the minor difference: towards the center of Conway's hair, a couple of hair strands can be seen in the original picture, which are 'dissolved' in $A_{(160)}$.

Problem 2

If "True," provide a proof. If "False," provide a counterexample and justify why your counterexample is valid.

1. A point $\overline{x} \in \mathbb{R}^n$ is a local minimum of a quadratic (i.e., degree-2) polynomial $p : \mathbb{R}^n \to \mathbb{R}$ if and only if there are no descent directions at $\overline{x}$.
2. A point $\overline{x} \in \mathbb{R}^n$ is a local minimum of a cubic (i.e., degree-3) polynomial $p : \mathbb{R}^n \to \mathbb{R}$ if and only if there are no descent directions at $\overline{x}$.
3. Suppose $\Omega \subseteq \mathbb{R}^n$ is a closed convex set and c is a vector in $\mathbb{R}^n$. Consider the problem of minimizing $c^\top x$ over Ω. If this problem has a finite optimal value, then it has an optimal solution.

Solution

1. **True**.
   We've already seen in class that if $\overline{x}$ is a local minima, then there is no descent direction of p at $\overline{x}$. I'll prove that the converse is true for quadratic functions (please note: **we'll write q for $\overline{x}$**).
   If $p : \mathbb{R}^n \to \mathbb{R}$ is a quadratic polynomial, there is a symmetric real matrix $A \in \mathbb{R}^{n\times n}$, a real vector $b \in \mathbb{R}^n$ and a scalar $c \in \mathbb{R}$ such that $p(x) = x^\top Ax - 2b^\top x + c$.

   Consider the following statements for each point $x \in \mathbb{R}^n$ and each function $f : \mathbb{R}^n \to \mathbb{R}$:

   $X(f, x)$: $x \in \mathbb{R}^n$ is not a local minima of $f : \mathbb{R}^n \to \mathbb{R}$.

   $Y(f, x)$: There is no descent direction of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$.

   Lemma 1
   If $\frac{1}{2}\nabla^2 p(x) = A$ is diagonal (where p is the above quadratic polynomial), say this is diag$(\lambda_1, \cdots, \lambda_n)$, then there is no $q \in \mathbb{R}^n$ such that both $X(p, q), Y(p, q)$ are true.

   Proof. For the sake of contradiction, assume $\exists q \in \mathbb{R}^n$ is such that both $X(p, q), Y(p, q)$ are true. $Y(p, q) \implies \nabla p(q) = 0 \implies Aq = b \implies \lambda_i q_i = b_i \forall i$.

   Case 1*: Some eigenvalue of A is negative.* WLOG, say $\lambda_1 < 0$. Then we claim that e_1 is a descent direction. Indeed, if $\alpha \in (0, 1)$ then

   $$\begin{aligned} p(q + \alpha e_1) - p(q) &= \alpha^2 e_1^\top Ae_1 + 2\alpha q^\top Ae_1 - 2\alpha b^\top e_1 \\ &= \alpha^2\lambda_1 + 2\alpha(\lambda_1 q_1 - b_1) \\ &= \alpha^2\lambda_1 < 0. \end{aligned}$$

   [So $\overline{\alpha} = 1$ here]. This proves $Y(p, q)$ false.

   Case 2*: All eigenvalues of A are non-negative.* Then $A \succeq 0$. For any $y \in \mathbb{R}^n$

   $$\begin{aligned} p(q + y) - p(q) &= y^\top Ay + 2q^\top Ay - 2b^\top y \\ &= y^\top Ay + 2y^\top Aq - 2b^\top y \\ &= y^\top Ay + 2y^\top b - 2b^\top y \\ &= y^\top Ay \geq 0 \end{aligned}$$

   where the last inequality is true because $A \succeq 0$. Taking $y = a - q$ gives $p(a) \leq (q)$, and this inequality holds for all $a \in \mathbb{R}^n$ (because it holds for every $y \in \mathbb{R}^n$). Therefore, q is a local minima of p, proving $X(p, q)$ false.

   These two cases are exhaustive and prove that $X(p, q), Y(p, q)$ cannot be simultaneously true. ■

Lemma 2
There is no $q \in \mathbb{R}^n$ such that both $X(p,q), Y(p,q)$ are true, where p is the aforementioned quadratic polynomial.

Proof. Since A, is real symmetric, there is an orthogonal matrix U (so $U^\top U = UU^\top = \mathrm{Id}_n$) such that $A = UDU^\top$ where $D = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ is the diagonal comprising of eigenvalues of A. Define $g(x) = p(Ux)$. Then

$$\begin{aligned} g(x) &= x^\top U^\top (UDU^\top) Ux - 2b^\top Ux + c \\ &= x^\top Dx - 2\tilde{b}^\top x + c \end{aligned}$$

where $\tilde{b} = U^\top b$. it satisfies that $\nabla^2 g(x)$ is diagonal for every $x \in \mathbb{R}^n$, so by Lemma 1 there is no point x such that $X(g,x), Y(g,x)$ are both true.

Fix a $q \in \mathbb{R}^n$. Then at least one of $X(g, U^\top q)$ or $Y(g, U^\top q)$ is false. Note $g(U^\top q) = p(q)$.
If $X(g, U^\top q)$ is false, then $U^\top q$ is a local minima of g. So $\exists \delta > 0$ such that if $\left\|y - U^\top q\right\| < \delta$ then $p(Uy) = g(y) \geq g(U^\top q) = p(q)$. So if $y \in \mathbb{R}^n$ satisfies $\|y - q\| < \delta$ then using the fact that $\left\|U^\top y - U^\top q\right\| = \|y - q\| < \delta$ (orthogonal matrices preserve the $2-$norm) we get $p(y) = g(U^\top y) \geq g(U^\top q) = p(q)$ whence q is a local minima for p. So $X(p,q)$ is false.
If $Y(g, U^\top q)$ is false, then there is a (descent) direction d and $\overline{\alpha} > 0$ such that $g(U^\top q + \alpha d) < g(U^\top q) = p(q)$. Noting that $g(U^\top q + \alpha d) = p(q + \alpha(Ud))$ and setting $\tilde{d} \coloneqq Ud$ gives $p(q + \alpha\tilde{d}) < p(q) \forall \alpha \in (0, \alpha)$, whence p has a descent direction $\tilde{d}$ at q. This means $Y(p,q)$ is false.

Since at least one of $X(g, U^\top q), Y(g, U^\top q)$ is false, we conclude that at least one of $X(p,q), Y(p,q)$ is false. ■

2. **False**.
It is true (as seen in class) that if p has a local minima at $\overline{x}$, then there is no descent direction of p at $\overline{x}$. We show a counterexample for the converse. Consider the function $p : \mathbb{R}^2 \to \mathbb{R}$ given by $p(x_1, x_2) = x_2(x_2 - x_1^2)$. We prove that p has no descent direction at $\overline{x} = (0,0)$, yet isn't a local minima. *A wider class of examples can be found from Theorem 3 in the Appendix of this problem.*

Let's see why it's *not a local minima*. Indeed $p(\overline{x}) = 0$ and the sequence of points $\left\{q_k = \left(\frac{1}{k}, \frac{1}{2k^2}\right)\right\}_{k \in \mathbb{N}}$ converge to $\overline{x} = (0,0)$ (because their norm $\frac{1}{k} \cdot \sqrt{1 + \frac{1}{2k^2}}$ converges to 0 as $k \to \infty$) and yet $p(q_k) = \frac{-1}{4k^4} < 0 \forall k \in \mathbb{N}$, proving that there is no ball around $\overline{x}$ where all points have functional value (under p) at least $p(\overline{x}) = 0$.

Now we show that there is *no descent direction of p at $\overline{x}$*. Suppose (a,b) is a descent direction. Define $x_\alpha \coloneqq (\alpha a, \alpha b)$ for $\alpha \in \mathbb{R}$. Then there is some $\overline{\alpha} > 0$ such that $\alpha \in (0, \overline{\alpha}) \implies p(x_\alpha) = \alpha b(\alpha b - \alpha^2 a^2) < 0$. Because $\alpha \neq 0$, this is equivalent to saying $b(b - \alpha a^2) < 0$. Note that $b \neq 0$ because the inequality is strict. If $b < 0$ then $b - \alpha a^2 < -\alpha a^2 \leq 0$ (by adding $-\alpha a^2$ on both sides of the inequality and noting that $\alpha, a^2 \geq 0$) implying that $p(x_\alpha) > 0$. If $b > 0$, then taking $\alpha \in \left(0, \frac{1}{2}\min\left\{\frac{b}{a^2}, \overline{\alpha}\right\}\right)$ we get $b - \alpha a^2 > b - \frac{b}{a^2} \cdot a^2 = 0$ whence $p(x_\alpha) > 0$. This contradicts the fact that (a,b) is a descent direction.

3. **False**.
Consider the set $\Omega = \left\{(x_1, x_2) \in \mathbb{R}^2 \;\middle|\; x_1 \geq 0, x_1 x_2 \geq 1\right\}$ with $c = (0,1)$.
Ω is closed because Ω is the intersection of inverse image of $[1, \infty)$ under the function $(x_1, x_2) \mapsto x_1 x_2$ and the inverse image of $[0, \infty)$ under the function $(x_1, x_2) \to x_1$, and both of these are continuous functions (we've used the fact that inverse of a closed set, under a continuous function, is a closed set; and that the intersection of two closed sets is closed).
This is also convex, which is seen as follows. Let $(a,b), (u,v) \in \Omega, \lambda \in [0,1]$. Note that $a, b, u, v \geq 0$. Then taking $\mu = 1 - \lambda \in [0,1]$ gives $\lambda a + \mu u \geq 0$ because $\lambda, \mu, a, u \geq 0$, and $(\lambda a + \mu u)(\lambda b + \mu v) =$

$\lambda^2 a + \lambda\mu(bu + av) + \mu^2 v \geq \lambda^2 + \lambda\mu\left(\frac{u}{a} + \frac{a}{u}\right) + \mu^2 \geq \lambda^2 + 2\lambda\mu + \mu^2 = 1$. So $\lambda(a, b) + (1 - \lambda)(u, v) \in \Omega$. Then the objective function is just $f(x_1, x_2) = (0, 1)^\top (x_1, x_2) = x_2$. This gives $\inf_{(x_1,x_2)\in\Omega} x_2 = 0$ (each x_2 is non-negative and, for example, consider the sequence of points $(n, 1/n)$) but there is no (x_1, x_2) such that $x_2 = 0$ because of the given conditions (otherwise $1 \leq x_1 \cdot 0 = 0$). So optimal value exists, but optimal solution doesn't.

Appendix

Theorem 3
Let $f, g : \mathbb{R} \to \mathbb{R}$ be $\mathscr{C}^1$ functions such that such that

- $\exists \varepsilon > 0$ such that $(f - g)^{-1}(0) \cap (0, \varepsilon) = \varnothing$,
- $f(0) = g(0) = 0$,
- $f'(0) = g'(0)$ (call it d), and
- $\exists \delta > 0$ such that d does not lie strictly between $\frac{f(t)}{t}$ and $\frac{g(t)}{t}$ for every $t \in (0, \delta)$,

then $h : \mathbb{R}^2 \to \mathbb{R}$ given by $h(x, y) = (y - f(x))(y - g(x))$ has no descent direction at $(0, 0)$ and $(0, 0)$ is not a local minima of h.

Proof. Suppose $(a, b) \neq (0, 0)$ is a descent direction at $(0, 0)$, for the sake of contradiction. So there is some $\overline{\alpha}$ such that $h(\alpha a, \alpha b) < h(0, 0) \forall \alpha \in (0, \overline{\alpha})$. We can assume, WLOG, that $0 < a^2 + b^2 < \varepsilon$ (for example, by replacing (a, b) with $(\lambda a, \lambda b)$ such that $\lambda < \frac{\sqrt{\varepsilon}}{\overline{\alpha}\sqrt{a^2+b^2}}$). One must have $f > g$ or $g > f$ on $(0, \varepsilon)$ because otherwise, by continuity, there would be a non-zero point in $(0, \varepsilon) \subset B^o(0, \varepsilon)$ which is a zero of the function $f - g$. WLOG, assume $f > g$ on $(0, \varepsilon)$. This implies $0 > h(a, b) = (b - f(a))(b - g(a))$, whence $g(a) < b < f(a)$. In fact, it is true that $0 > h(\alpha a, \alpha b) = (\alpha b - f(\alpha a))(\alpha b - g(\alpha a)) \forall \alpha \in (0, \overline{\alpha})$ by definition. From this it follows that $a \neq 0$ (because if $a = 0$ then $h(0, \alpha b) = (\alpha b)^2 > 0$). We thus have $g(\alpha a) < \alpha b < f(\alpha a)$ which gives that $b = da$ by using the sandwich theorem. Then note that if $\alpha = \frac{1}{2}\min\left\{\frac{\delta}{a}, \overline{\alpha}\right\}$ we have $h(\alpha a, \alpha b) = (\alpha b - f(\alpha a))(\alpha b - g(\alpha a)) = (\alpha a d - f(\alpha a))(\alpha a d - g(\alpha a)) = t^2\left(d - \frac{f(t)}{t}\right)\left(d - \frac{g(t)}{t}\right)$ where $t = \alpha a < \delta$ so this quantity is ≥ 0 by the assumption in the last bullet point, however this contradicts the fact that $h(\alpha a, \alpha b) < 0$.

Now we show that $(0, 0)$ not a local minima. Since $f > g$ on $(0, \varepsilon)$, one can choose points $x_n = \frac{1}{n}$ and $y_n \in (g(x_n), f(x_n))$ and observe that $\lim y_n = 0$ by continuity of f, g and sandwich theorem. Thus, for every $\varepsilon > 0 \exists N \in \mathbb{N}$ such that $k \geq N \implies |y_k| < \varepsilon$. However

$$h(x_n, y_n) = (y_n - f(x_n))(y_n - g(x_n)) < 0$$

by choice of y_n. This proves that $(0, 0)$ cannot be a local minima of h, because $(x_n, y_n) \to (0, 0)$ but each $h(x_n, y_n) < 0$. ∎

Problem 3

1. Let $Q \in S^{n\times n}$ and assume $Q \succ 0$. Show that $f(x) = \sqrt{x^\top Q x}$ is a norm.
2. Show that Q^{-1} exists and is positive definite. Show that the dual norm of f is given by $f(x) = \sqrt{x^\top Q^{-1} x}$.
3. Let $A \in \mathbb{R}^{m\times n}$. Prove the following expression for its induced $2-$norm: $\|A\|_2 = \sqrt{\lambda_{\max}(A^\top A)}$.

Solution

Q is symmetric positive definite, whence $Q = UDU^\top$ with D being a diagonal matrix of positive diagonal entries (say $\lambda_1^2, \cdots, \lambda_n^2$ with $\lambda_i \geq 0$), and U a real orthogonal matrix. For a vector $x = (x_1, \cdots, x_n)$, we denote $\tilde{x} = U^\top x$ and its coordinates by $\tilde{x} = (\tilde{x}_1, \cdots, \tilde{x}_n)$.

1. Note that $x^\top Q x = \tilde{x}^\top D \tilde{x} = \sum_i \lambda_i^2 \tilde{x}_i^2$ for any $x \in \mathbb{R}^n$. So
   - (Positivity) $x^\top Q x \geq 0$ by definition of positive semi-definiteness and if $x \neq 0$ then $x^\top Q x > 0$ because $Q \succ 0$; so that if $x^\top Q x = 0$ then $x = 0$.
   - (Triangle inequality) Observe $f(x+y)^2 = \sum_i \lambda_i^2 (\tilde{x}_i + \tilde{y}_i)^2 = f(x)^2 + f(y)^2 + 2\sum_i \lambda_i^2 \tilde{x}_i \tilde{y}_i$. Thus

$$\begin{aligned}
\left(\sum_i \lambda_i^2 \tilde{x}_i^2\right)\left(\sum_j \lambda_j^2 \tilde{y}_j^2\right) &= \sum_{1\leq i,j\leq n} \lambda_i^2 \lambda_j^2 \left(\tilde{x}_i^2 \tilde{y}_j^2\right) \\
&= \sum_{1\leq i<j\leq n} \lambda_i^2 \lambda_j^2 \left(\tilde{x}_i^2 \tilde{y}_j^2 + \tilde{x}_j^2 \tilde{y}_i^2\right) \\
&\geq 2 \sum_{1\leq i<j\leq n} \lambda_i^2 \lambda_j^2 (\tilde{x}_i \tilde{y}_j \tilde{x}_j \tilde{y}_i) \\
&= \sum_{1\leq i,j\leq n} \lambda_i^2 \lambda_j^2 (\tilde{x}_i \tilde{y}_j \tilde{x}_j \tilde{y}_i) \\
&= \left(\sum_i \lambda_i^2 \tilde{x}_i \tilde{y}_i\right)\left(\sum_j \lambda_j^2 \tilde{x}_j \tilde{y}_j\right)
\end{aligned}$$

$$\begin{aligned}
\implies f(x)^2 f(y)^2 = \left(\sum_i \lambda_i^2 \tilde{x}_i^2\right)\left(\sum_j \lambda_j^2 \tilde{y}_j^2\right) &\leq \left(\sum_i \lambda_i^2 \tilde{x}_i \tilde{y}_i\right)^2 \\
\implies 2f(x)f(y) \leq 2\sum_i \lambda_i^2 \tilde{x}_i \tilde{y}_i &= f(x+y)^2 - f(x)^2 - f(y)^2 \\
\implies [f(x) + f(y)]^2 &\leq f(x+y)^2 \\
\implies f(x) + f(y) &\leq f(x+y).
\end{aligned}$$

     (One can alternately prove Cauchy-Schwarz inequality for real inner products, which it is in this case due to positive definiteness, from which the above inequality follows — without the eigendecomposition)
   - (Homogeneity) $f(\lambda x) = \sqrt{\lambda^2 x^\top Q x} = |\lambda| \sqrt{x^\top Q x} = |\lambda| f(x)$ for any $\lambda \in \mathbb{R}, x \in \mathbb{R}^n$.
2. Note that D is invertible, its inverse is simply given by the diagonal matrix with diagonal entries $\frac{1}{\lambda_i^2} > 0$. Taking $P = UD^{-1}U^\top$ gives $PQ = UD^{-1}U^\top UDU^\top = UD^{-1}DU^\top = UU^\top = \mathbf{1}_n$ and this proves that P is the inverse of Q because Q is a square matrix.

Denote $\sqrt{Q} \coloneqq U\sqrt{D}U^\top$ where $\sqrt{D} = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ and observe that it is symmetric, invertible and satisfies $\sqrt{Q}^2 = Q$, because $\sqrt{D}$ is invertible (positive eigenvalues) and $\sqrt{D}^2 = D$. Recall that the dual norm is given by $(\|x\|_* =) \|x\|_{*,f} = \sup\left\{x^\top y : f(y) \le 1\right\}$. We start with the observation that $f(y) = \sqrt{y^\top Q y} = \sqrt{y^\top \sqrt{Q}\sqrt{Q} y} = \sqrt{y^\top \sqrt{Q}^\top \sqrt{Q} y} = \left\|\sqrt{Q}y\right\|_2$. Note that $y = \frac{Q^{-1}x}{\sqrt{x^\top Q^{-1}x}}$ satisfies $f(y)^2 = y^\top Q y = \frac{x^\top Q^{-1} Q Q^{-1} x}{x^\top Q^{-1} x} = \frac{x^\top Q^{-1} x}{x^\top Q^{-1} x} = 1$ and $y^\top x = \frac{x^\top Q^{-1} x}{\sqrt{x^\top Q^{-1}x}} = \sqrt{x^\top Q^{-1} x}$. This shows that $\|x\|_* \ge \sqrt{x^\top Q^{-1} x}$. But Cauchy Schwarz inequality gives

$$x^\top y = \left(\left(\sqrt{Q}\right)^{-1} x\right)^\top \left(\sqrt{Q} y\right) \le \left\|\left(\sqrt{Q}\right)^{-1} x\right\|_2 \left\|\sqrt{Q} y\right\|_2 \le \left\|\left(\sqrt{Q}\right)^{-1} x\right\|_2 = \sqrt{x^\top Q^{-1} x}.$$

This proves that $\|x\|_* = \sqrt{x^\top Q^{-1} x}$.

3. $A \in \mathbb{R}^{n\times n}$. $\|A\|_2^2 = \max_{\|x\|_2 = 1} \|Ax\|_2^2 = \max_{\|x\|_2 = 1} x^\top (A^\top A) x$. Say μ_i are the eigenvalues of $A^\top A$ with $\mu_1 \ge \cdots \ge \mu_n$. By the previous problem, $\mu_i \ge 0$. And by the calculation shown above, $x^\top A^\top A x = \sum_i \mu_i \tilde{x}_i^2$ (Here $\tilde{x} = U^\top x$ where U is determined by the orthogonal diagonalization of $A^\top A$). Taking $\tilde{x} = (1, 0, \cdots, 0) =: \tilde{x}_0$, thus $x_0 = U\tilde{x}_0$, gives $x_0^\top A^\top A x_0 = \mu_1$, so $\boxed{\|A\|_2^2 \ge \mu_1}$ because $\|\tilde{x}_0\|_2 = 1$. Note that we can independently choose $\tilde{x}$ because $x, \tilde{x}$ are different only by unitary transformation (both invertible and preserves norm: $\|Ux\|^2 = x^\top U^\top U x = x^\top x = \|x\|^2$).

   The observation that $\max\{\mu_i\} = \mu_1$ gives $\|Ax\|_2^2 = \sum_i \mu_i \tilde{x}_i^2 \le \sum_i \mu_1 \tilde{x}_i^2 = \mu_1$ whence $\boxed{\|A\|_2^2 \le \mu_1}$.

Problem 4

Prove or disprove the following statements. All matrices are *symmetric, $n \times n$, and with real entries*.

(a) Suppose $A \succeq 0$. Then the largest entry in absolute value of A must be on the diagonal.

(b) If $A \succeq 0$ and $\mathrm{Tr}(A) = 0$, then $A = 0$.

(c) If $A \succeq 0, B \succeq 0$, and $A + B = 0$, then $A = B = 0$.

(d) If $A \succeq 0, B \succeq 0$, and $AB = 0$, then $A = 0$ or $B = 0$.

(e) Suppose $x^\top Ax \geq 0$ for all $x \geq 0$, then either $A \geq 0$ or $A \succeq 0$.

Solution

(a) **True**.
For the sake of contradiction, assume there are indices $i < j$ such that $|A_{ji}| = |A_{ij}| > |A_{kk}| \forall 1 \leq k \leq n$. First we see that each $e_k^\top Ae_k = A_{kk} \geq 0$ by positive semi-definiteness. Next $0 \leq (e_i + e_j)^\top A(e_i + e_j) = A_{ii} + 2A_{ij} + A_{jj} \implies A_{ii} + A_{jj} \geq -2A_{ij}$. But, $A_{ii}, A_{jj} < |A_{ij}| \implies A_{ii} + A_{jj} < 2|A_{ij}|$. Putting these together, we have $2|A_{ij}| > A_{ii} + A_{jj} \geq -2A_{ij}$ which suggests that $A_{ij} \geq 0$ and thus $|A_{ij}| = A_{ij}$. Now $(e_i - e_j)^\top A(e_i - e_j) = A_{ii} + A_{jj} - 2A_{ij} = A_{ii} + A_{jj} - 2|A_{ij}| < 0$, contradicting that A is positive semi-definite.

(b) **True**.
$A \succeq 0$ implies that all eigenvalues are non-negative. Recall that trace is simply the sum of eigenvalues. But if their sum (of non-negative numbers) is zero, the only way that's possible is that all of them are zero. A is a conjugate of the zero matrix (because $A = UDU^\top$ for orthogonal U and diagonal D of eigenvalues), which is the 0 matrix in this case.

(c) **True**.
$A = -B \implies A, B$ are simultaneously diagonalizable (because each of them are diagonalizable and they commute). So $UAU^\top = -UBU^\top$ is diagonal. This means that if λ_i's are the eigenvalues of A, the eigenvalues of B are $-\lambda_i$'s. Positive semi-definiteness implies that $\lambda_i \geq 0, -\lambda_i \geq 0 \forall i$ and thus $\lambda_i = 0 \forall i$. So $UAU^\top = -UBU^\top$ is the 0 matrix. This means $A = B = 0$.

(d) **False**.
For a concrete example, take $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ are both symmetric positive semi-definite satisfying $AB = 0$, but neither of A, B is 0. One can find such an example for each dimension by taking $A = \mathrm{diag}(1, 0, \cdots, 0, 0), B = \mathrm{diag}(0, 0, \cdots, 0, 1)$.

(e) **False**.
Consider $A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. First observe that $A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ where the first summand is entry-wise non-negative and the second summand is symmetric positive semi-definite (with eigenvalues $0, 2$). This shows that $x^\top Ax \geq 0 \forall x \geq 0$. But neither does A have all non-negative entries, nor is A positive semi-definite (because $(e_2 - e_3)^\top A(e_2 - e_3) = -1$).

Image Compression

[1]:
```
#Nilava Metya
#nm8188@princeton.edu
#ORF 523
#Python 3
```

[2]:
```
import numpy as np
from PIL import Image as im
import matplotlib.pyplot as plt
```

[3]:
```
image = im.open('conway.jpg')
A = np.array(image, dtype=float) / 255.0
A = np.dot(A, [0.2989, 0.5870, 0.1140])
m,n = A.shape
```

[4]:
```
U, S, Vt = np.linalg.svd(A)                                    #do the SVD
K = [40,80,120,160]                                            #storing values of k
l = len(K)
S = np.diag(S)                                                 #original S is only a linear␣
 ↪array, need to convert to matrix
```

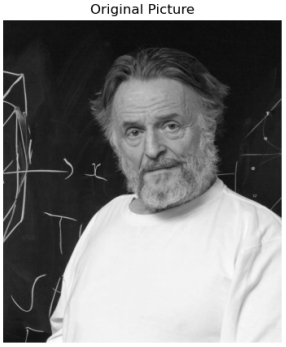
[5]:
```
approx = []
for i in range(l):
    approx.append(U[:, :K[i]] @ S[:K[i], :K[i]] @ Vt[:K[i], :])           ␣
 ↪#taking approximations
```

[6]:
```
#obtaining and printing a table for Frobeius norm of differences
error = [0 for _ in range(l)]
print("k\t|\t Frobenius norm")
print("--------+--------------------------")
for i in range(len(K)):
    error[i] = np.linalg.norm(A-approx[i])
    print(str(K[i])+"\t|\t"+str(error[i]))
```

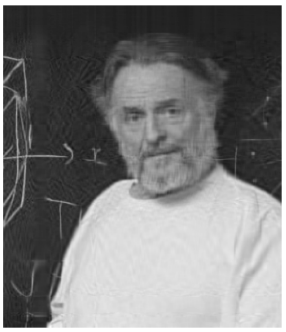
```
k       |        Frobenius norm
--------+--------------------------
40      |       31.733288485394482
80      |       19.016427727768185
120     |       13.055854194249326
160     |       9.511015173570117
```

[7]:
```
fig, ax = plt.subplots(len(K)+1, 1, figsize = (10, 30))

ax[0].set_title("Original Picture")
ax[0].imshow(A, cmap = 'gray')
ax[0].axis('off')

for i in range(l):
    ax[i+1].imshow(approx[i], cmap = 'gray')
    ax[i+1].set_title("k = " + str(K[i]))
    ax[i+1].axis('off')

plt.show()
```

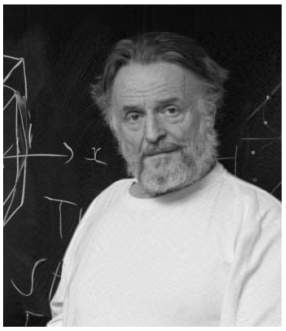
Original Picture



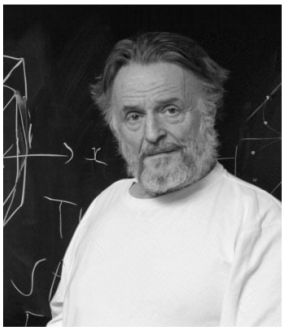
k = 40



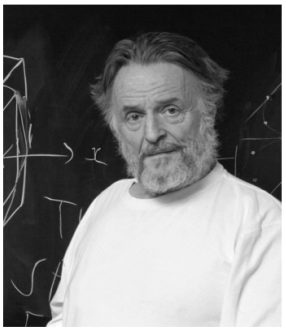
k = 80



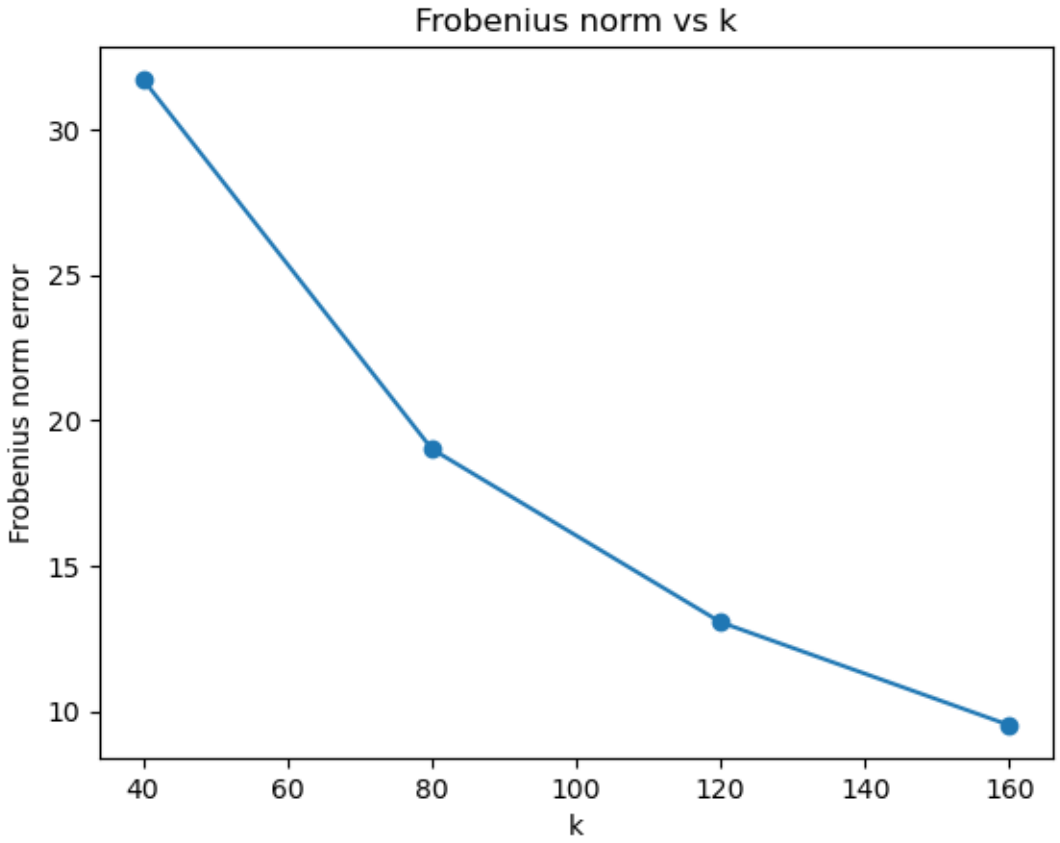
k = 120



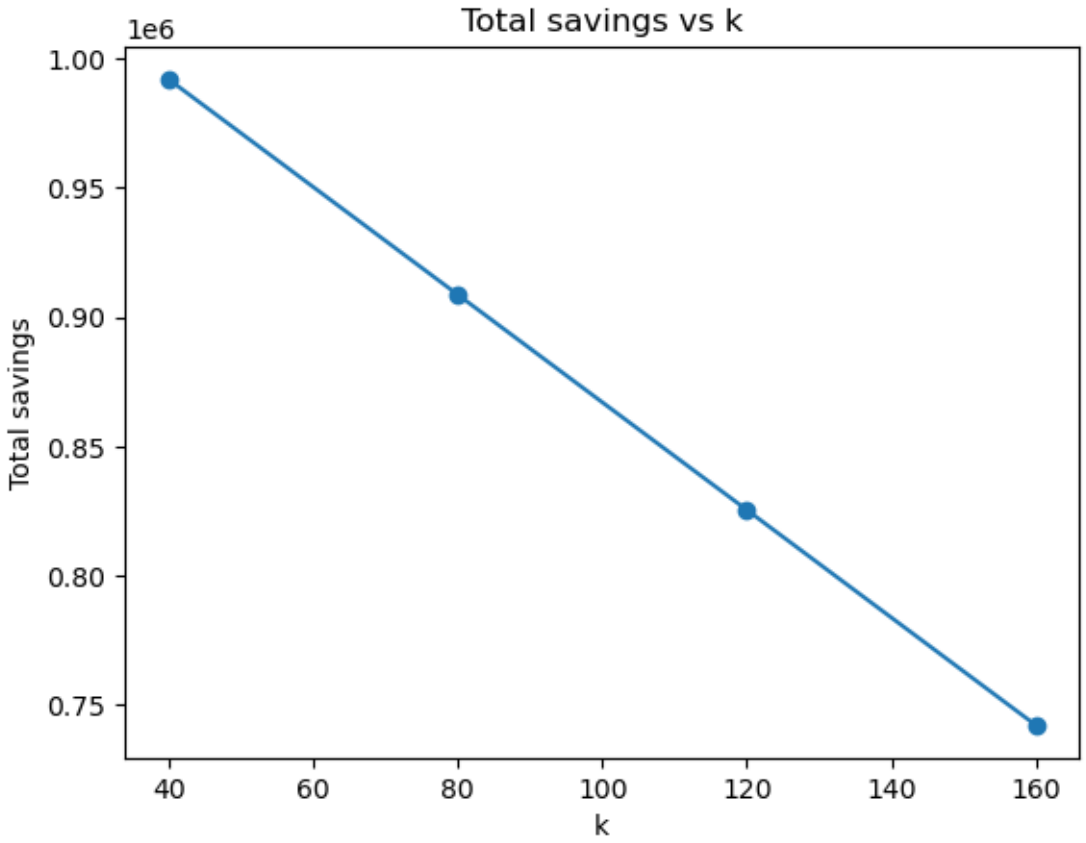
k = 160



[8]:
```
plt.plot(K, error, marker = 'o')
plt.xlabel('k')
plt.ylabel('Frobenius norm error')
plt.title("Frobenius norm vs k")
plt.show()
```



[9]:
```
plt.plot(K, [m*n-(m+n+1)*k for k in K], marker = 'o')
plt.xlabel('k')
plt.ylabel('Total savings')
plt.title("Total savings vs k")
plt.show()
```



[10]:
```
plt.imsave('A160.jpg', approx[3], cmap = 'gray')
plt.imsave('original.jpg', A, cmap = 'gray')
```

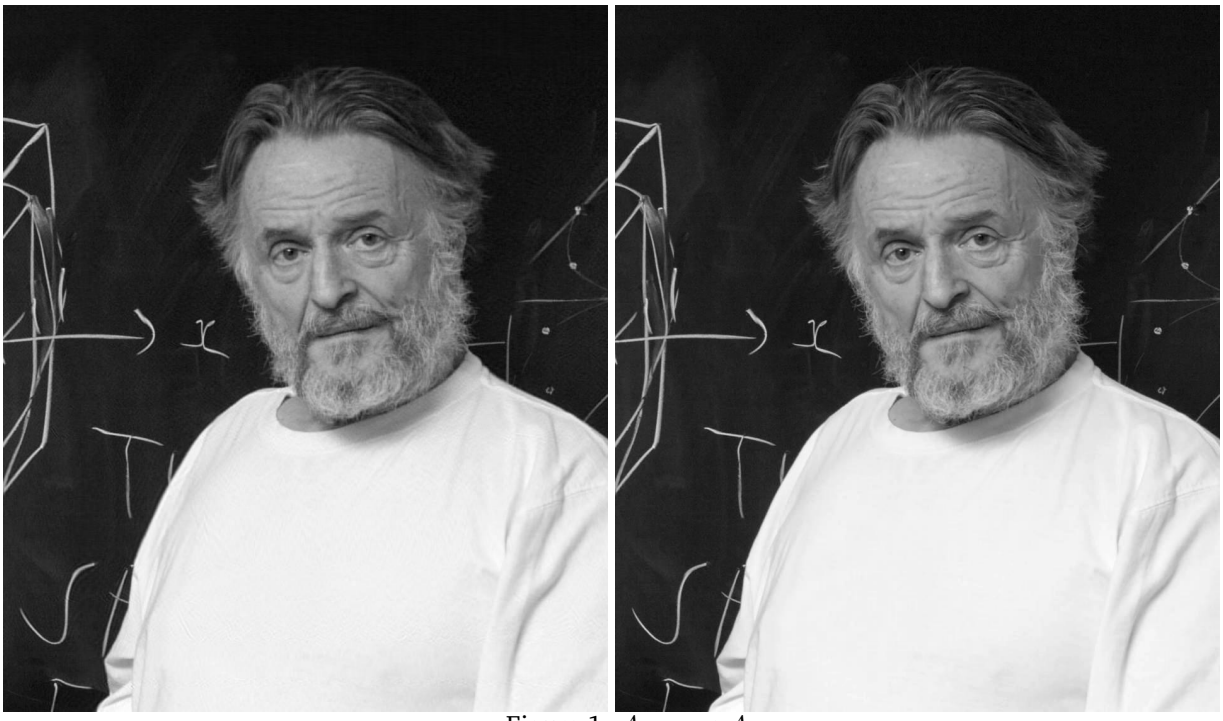


Figure 1: $A_{(160)}$ vs A