HOMEWORK 0 SOLS

Exercise 1 (Matrix Norms)**.** *Let A be an $n \times n$ matrix with singular values $\sigma_1 \geq \sigma_2 \geq \dots \sigma_n$. The* spectral *norm $\|A\|_2$ of A is the maximum of $\|Av\|_2$ over all vectors $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$. The* Frobenius *norm $\|A\|_F$ of A is $\sqrt{\sum A_{i,j}^2}$.*

1. *Prove that $\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2$. When are these two inequalities tight (give examples)?*

   Let $a_1, \dots, a_n$ be the rows of A. For all vectors v, $\|Av\|_2 = \sqrt{\sum_{i=1}^n \langle a_i, v\rangle^2} \leq \sqrt{\sum_{i=1}^n \|a_i\|_2^2} = \|A\|_F$, by Cauchy-Schwarz. Hence, $\|A\|_2 \leq \|A\|_F$. Let a_j be the row such that $\|a_j\|_2^2$ is maximized. Then $\|A\|_F = \sqrt{\sum_{i=1}^n \|a_i\|_2^2} \leq \sqrt{n\|a_j\|_2^2} \leq \sqrt{n}\sqrt{\sum_{i=1}^n \langle a_i, a_j\rangle^2} = \sqrt{n}\|Aa_j\|_2 \leq \sqrt{n}\|A\|_2$.

   If $A = e_1 e_1^\top$ then $\|A\|_2 = \|A\|_F = 1$, and if $A = \mathrm{Id}$ then $\|A\|_F = \sqrt{n}\|A\|_2 = \sqrt{n}$.

2. *Prove that for symmetric A (i.e., $A = A^\top$), $\|A\|_2 \leq \max_i \sum_j |A_{i,j}|$*[1]*.*

   When A is symmetric, $\|A\|_2$ is $\max|\lambda|$ over eigenvalues λ of A. Let v be an eigenvector so that $Av = \lambda v$, where $|\lambda| = \|A\|_2$. Rescale v so that the largest coordinate of v (WLOG assume this is v_1) is 1. Then $\lambda = (Av)_1 = \sum_j A_{1,j}v_j$ so that $\|A\|_2 = |\lambda| \leq |\sum_j A_{1,j}v_j| \leq \sum_j |A_{1,j}||v_j| \leq \sum_j |A_{1,j}|$.

3. *Prove that for the adjacency matrix A of a d-regular graph $\|A\|_2 = d$.*

   First, observe that $A\mathbf{1} = d \cdot \mathbf{1}$ since A is d-regular, which implies that $\|A\|_2 \geq d$. We have that $\sum_j |A_{i,j}| = d$ for every i, and so by the previous exercise we have $\|A\|_2 \leq d$, and therefore $\|A\|_2 = d$.

[1] **Hint:** You can do this via the following stronger inequality: for any (not necessarily symmetric) matrix A, $\|A\| \leq \sqrt{\alpha\beta}$ where $\alpha = \max_i \sum_j |A_{i,j}|$ and $\beta = \max_j \sum_i |A_{i,j}|$.

Exercise 2 (Trace Moment Method)**.** *Let A be a symmetric $n \times n$ matrix.*[2]

[2] Tao's blogpost explains this and other methods to prove sharp upper-bounds on the norms of random matrices

1. *Let $\mathrm{Tr}(A) = \sum A_{i,i}$. Prove that for every even k, $\|A\|_2 \leq \mathrm{Tr}(A^k)^{1/k} \leq n^{1/k}\|A\|_2$.*

   We first show that for all k, $\mathrm{Tr}(A^k) = \sum_i \lambda_i^k$, where the λ_i's are the eigenvalues of A. Since A is symmetric, we can write $A = \sum_i \lambda_i v_i v_i^\top$, where the v_i's are the orthonormal eigenvectors of A. By induction, we have $A^k = \sum_i \lambda_i^k v_i v_i^\top$, as the base case is obvious and $(\sum_i \lambda_i^{k-1} v_i v_i^\top)(\sum_i \lambda_i v_i v_i^\top) = \sum_i \sum_j \lambda_i^{k-1}\lambda_j v_i v_i^\top v_j v_j^\top = \sum_i \lambda_i^k v_i v_i^\top$ because the v_i's are orthonormal. We then have that $\mathrm{Tr}(A^k) = \sum_i \lambda_i^k \mathrm{Tr}(v_i v_i^\top)$. We have that $\mathrm{Tr}(v_i v_i^\top) = \sum_j v_{i,j}v_{i,j} = \|v\|_2^2 = 1$, and so $\mathrm{Tr}(A^k) = \sum_i \lambda_i^k$.

   Now, when k is even, $\|A\|_2 = \max_i |\lambda_i| \leq (\sum_i |\lambda_i|^k)^{1/k} = \mathrm{Tr}(A^k)^{1/k}$, and $(\sum_i |\lambda_i|^k)^{1/k} \leq (n \cdot \max_i |\lambda_i|^k)^{1/k} = n^{1/k}\|A\|_2$.

2. *(Bonus) Let A be a symmetric matrix such that $A_{i,i} = 0$ for all i and $A_{i,j}$ is chosen to be a random value in $\{\pm 1\}$ independently of all others.* ***(a)*** *Prove that (for n sufficiently large) with probability at least 0.99, $\|A\|_2 \leq n^{0.9}$.* ***(b)*** *Prove that with probability at least 0.99, $\|A\|_2 \leq n^{0.51}$.*

   Part (a) can be shown by bounding $\mathrm{Tr}(A^4)$. We will prove part (b) by bounding $\mathrm{Tr}(A^k)$ for some even k. We have that $\mathrm{Tr}(A^k) = \sum_i \sum_{j_1,\ldots,j_{k-1}} A_{i,j_1} A_{j_1,j_2} \cdots A_{j_{k-1},i} = \sum_{i_1,\ldots,i_k} \prod_{j=1}^{k} A_{i_j,i_{j+1}}$, using the convention that $k+1$ is 1. Since $A_{i,i} = 0$ always, this is equivalent to summing over all length-k cycles in the complete graph (edges are allowed to repeat), where each cycle contributes the product of $A_{i,j}$ over the edges (i,j) in the cycle. Now, let us compute $\mathbb{E}[\mathrm{Tr}(A^k)]$. Observe that if a cycle contains an edge with odd multiplicity, then by independence of the entries of A it has expectation 0, and if a cycle contains every edge with even multiplicity, then its expectation is always 1. Hence, $\mathbb{E}[\mathrm{Tr}(A^k)]$ is simply the number of length-k cycles where every edge has even multiplicity. Therefore, each of these cycles has at most $k/2$ distinct edges, and therefore at most $k/2+1$ distinct vertices, as we have the initial vertex, and then every new edge adds at most one new distinct vertex. There are at most $n^\ell \ell^k$ cycles with ℓ distinct vertices, as first we choose a subset of size ℓ from $[n]$, and then every cycle is a sequence of length k of these vertices. Therefore, there are at most $k \cdot n^{k/2+1}(k/2+1)^k$ such cycles, and so $\mathbb{E}[\mathrm{Tr}(A^k)] \leq k \cdot n^{k/2+1}(k/2+1)^k$. By Markov's inequality, $\mathbb{P}[\|A\|_2 \geq \lambda] \leq \mathbb{P}[\mathrm{Tr}(A^k) \geq \lambda^k] \leq \lambda^{-k}\,\mathbb{E}[\mathrm{Tr}(A^k)] \leq \lambda^{-k} \cdot k \cdot n^{k/2+1}(k/2+1)^k$. Setting $\lambda = n^{1/2} \cdot 2k$, this is $\leq (2k)^{-k} \cdot nk^{k+1} \leq 2^{-k} \cdot nk \leq 0.01$ for $k = 2\log n$. So, with probability ≥ 0.99, we have $\|A\|_2 \leq O(\sqrt{n}\log n) \leq n^{0.51}$.

Note: *While $\|A\|_2$ can be computed in polynomial time, both $\max_i \sum_j |A_{i,j}|$ and $\|A\|_F$ give even simpler to compute upper bounds for $\|A\|_2$. However the examples in Exercise 1 show that they are not always tight. It is often easier to compute $\mathrm{Tr}(A^k)^{1/k}$ than trying to compute $\|A\|_2$ directly, and as k grows this yields a better and better estimate.*

Exercise 3 (Positive Semidefinite Matrices)**.** *Let A be an $n \times n$ symmetric matrix. Prove that the following are equivalent:*

1. *For every vector $v \in \mathbb{R}^n$, $v^\top A v \geq 0$ (note: $v^\top A v = \sum_{i,j} A_{i,j} v_i v_j$).*

2. *All eigenvalues of A are non-negative.*

3. *There are linear functions $L_1, \ldots, L_m$ such that the quadratic polynomial P_A defined as $P_A(x) = \sum A_{i,j} x_i x_j$ can be written as $P_A = \sum_i L_i^2$. In particular, the polynomial P_A is a* sum of squares*.*

4. *$A = B^\top B$ for some $r \times n$ matrix B*

Any of the above statements can be taken as a definition of positive semidefinite matrices (denoted by $A \succeq 0$.)

We show that $(1) \implies (2) \implies (3) \implies (1)$ *and that* $(2) \implies (4) \implies (1)$.

For any eigenvector v of A with eigenvalue λ we have $v^\top Av = \lambda\|v\|_2^2$. So, if (1) *holds then* (2) *must hold also.*

Since A is symmetric, by SVD we get that $P_A(x) = x^\top Ax = \sum_{i=1}^n \lambda_i \langle u_i, x\rangle^2$ where the u_i's are the eigenvectors of A with eigenvalue λ_i. If all λ_i are nonnegative then we can take $L_i(x) := \sqrt{\lambda_i}\langle u_i, x\rangle$, and then $P_A = \sum_{i=1}^n L_i^2$. This shows that $(2) \implies (3)$.

Suppose that (3) *holds. Then $v^\top Av = P_A(v) \geq 0$ for all $v \in \mathbb{R}^n$, and so* (1) *holds.*

Finally, we show that $(2) \implies (4) \implies (1)$. *By SVD, we can write $A = U^\top \Sigma U$, where U is orthogonal. If* (2) *holds then Σ is a diagonal matrix with nonnegative entries, and so $\Sigma = D^2$ where D is the diagonal matrix with $D_{i,i} = \sqrt{\Sigma_{i,i}}$. We then see that taking $B = DU$ we get that $B^\top B = A$, so* (4) *holds. If* (4) *holds then for any v, $v^\top Av = \|Bv\|_2^2 \geq 0$, and so* (1) *holds.*

Note: *The intention was for you to also show that $r \leq n$ in (4), but we did not deduct points if your proof didn't imply this as this wasn't specified in the question. Most proofs showing that* $(3) \implies (4)$ *typically did not imply that $r \leq n$.*

Exercise 4 (Gaussian Distributions)**.** *The standard Gaussian random variable g on $\mathbb{R}^n$ has the following probability density function: $\gamma(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp(-\frac{\|x\|_2^2}{2})$.*

1. *(Rotation Invariance) Let g be a a standard Gaussian random variable. Prove that for any matrix U such that $U^\top U = I = UU^\top$, Ug is also a standard Gaussian random variable*[3].

   Observe that the density function of Ug is $\gamma(U^{-1}x)|\det(U^{-1})|$. Since U is orthogonal, U^{-1} is also orthogonal, and $\det U^{-1} = \pm 1$. Hence, the density of Ug is $\gamma(U^{-1}x)$. We note that $\gamma(x)$ is a function of $\|x\|_2$ only, and $\|U^{-1}x\|_2 = \|x\|_2$ always since U^{-1} is orthogonal. Hence, Ug has density $\gamma(x)$, so it is distributed as a standard Gaussian random variable.

2. *(Orthogonality implies Independence) Prove*[4] *that for any orthogonal pair of vectors v, w, and standard Gaussian g, the random variables $\langle g, v\rangle$ and $\langle g, w\rangle$ are independent.*

   First, we observe that when $v = e_1$ and $w = e_2$, $\langle g, v\rangle$ and $\langle g, w\rangle$ are

[3] **Hint:** For any v, how is $\|Uv\|_2$ related to $\|v\|_2$?

[4] First deal with the special case when $v = e_1$ and $w = e_2$, the standard basis vectors.

independent. This is because $\gamma(x) = \prod_{i=1}^{n} f(\langle e_i, x\rangle)$ where f is the density of a standard univariate Gaussian. Now, choose an orthogonal matrix U so that the first entry of Ug is $\langle v, g\rangle / \|v\|$ and the second entry is $\langle w, g\rangle / \|w\|$. By part (1), Ug is a standard Gaussian random variable, and so by the above $\langle e_1, Ug\rangle$ and $\langle e_2, Ug\rangle$ are independent. Hence, $\langle v, g\rangle / \|v\|$ and $\langle w, g\rangle / \|w\|$ are independent, and so $\langle g, v\rangle$ and $\langle g, w\rangle$ are also.

***Note:** A few people only showed that $\mathbb{E}[\langle g, v\rangle\langle g, w\rangle] = 0$. This does* not *imply that they are independent. Note that if X and Y are random variables that (marginally) are Gaussian and are uncorrelated, then it is not necessarily the case that they are independent. If X and Y are jointly Gaussian and uncorrelated, then they are independent, but one should prove this.*

Exercise 5 (Covariance Matrices)**.** *Let D be a probability distribution on $\mathbb{R}^n$. The* covariance matrix *of D is defined (assuming it exists) as $\mathbb{E}_D(x - \mathbb{E}_D x)(x - \mathbb{E}_D x)^\top$.*

1. *Prove that for any vectors $v_1, v_2, \dots, v_k \in \mathbb{R}^n$, the matrix $M = \sum_{i \le k} v_i v_i^\top$ is positive semidefinite.*

   Observe that for any $x \in \mathbb{R}^n$, $x^\top M x = \sum_{i \le k} x^\top v_i v_i^\top x = \sum_{i \le k} \langle x, v_i\rangle^2 \ge 0$, so $M \succeq 0$.

2. *Prove that the covariance matrix of any distribution D is positive semidefinite.*

   Let M be the covariance matrix. For any $v \in \mathbb{R}^n$, we have $v^\top M v = v^\top(\mathbb{E}_D(x - \mathbb{E}_D x)(x - \mathbb{E}_D x)^\top)v = \mathbb{E}_D v^\top(x - \mathbb{E}_D x)(x - \mathbb{E}_D x)^\top v = \mathbb{E}_D\langle x - \mathbb{E}_D x, v\rangle^2 \ge 0$.

3. *Prove that for any positive semidefinite matrix $M \in \mathbb{R}^{n \times n}$, there exists a probability distribution D on $\mathbb{R}^n$ with covariance M (and mean $\mathbf{0}$).*

   Let $v_1, \dots, v_n$ and $\lambda_1, \dots, \lambda_n \ge 0$ be the eigenvectors and eigenvalues of M. Note that $M = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$. Let $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ be independent Gaussian random variables, and let $x = \sum_{i=1}^{n} \sqrt{\lambda_i}\alpha_i v_i$. Note that x is well-defined since $\lambda_i \ge 0$ for all i, and that $\mathbb{E}\, x = \mathbf{0}$ since $\mathbb{E}\, \alpha_i = 0$ for all i.

   The covariance matrix of x is $\mathbb{E}(x - \mathbb{E}\, x)(x - \mathbb{E}\, x)^\top = \mathbb{E}\, xx^\top = \mathbb{E} \sum_{i=1}^{n} \sum_{j=1}^{n} \sqrt{\lambda_i \lambda_j}\alpha_i \alpha_j v_i v_j^\top = \mathbb{E} \sum_{i=1}^{n} \lambda_i \alpha_i^2 v_i v_i^\top = \sum_{i=1}^{n} \lambda_i\, \mathbb{E}[\alpha_i^2] v_i v_i^\top = \sum_{i=1}^{n} \lambda_i v_i v_i^\top = M$.

Exercise 6 (Basic Convexity and Hyperplane Separation)**.** *Let $K \subseteq$*

$\mathbb{R}^n$. *K is said to be* convex *if for every* $x, y \in K$ *and any* $\alpha \in [0,1]$, $\alpha x + (1-\alpha)y \in K$. *Let* $B = \{x \mid \sum_i x_i^2 \leq 1\}$ *be the unit ball in* ℓ_2 *norm. K is said to be* bounded *if* $K \subseteq cB$ *for some finite c. A point x is said to be a* limit point *of K if x can be approximated arbitrarily well by points in K. That is, for every* $\epsilon > 0$, *there exists a* $y \in K$ *such that* $\|x - y\|_2 \leq \epsilon$. *K is said to be* closed *if every limit point of K is contained in K. The* halfspace *defined by a vector h and a threshold* τ *is the set:* $H_{h,\tau} = \{x \in \mathbb{R}^n \mid \langle h, x\rangle \geq \tau\}$.

In the following, we will prove the following fundamental theorem of convex analysis: For any convex set K that is closed, bounded[5] *and convex and* $v \notin K$, *there exists a halfspaces* $H_{h,\tau}$ *such that* $K \subseteq H_{h,\tau}$ *while* $v \notin H_{h,\tau}$. *The hyperplane* $\{x \mid \langle h, x\rangle = \tau\}$ *is said to be a* separating *hyperplane for K and v.*

[5] Both these assumptions are not necessary. But the proof is slightly more technical.

1. *(Projection) Let K be a closed, bounded, convex set. Let* $v \notin K$. *Let* $\text{dist} : K \to \mathbb{R}$ *be the function defined by* $\text{dist}(x) = \|v - x\|$. *Verify that* dist *is continuous over K. Apply the extreme value theorem*[6] *to conclude that there exists a points* $m \in K$ *such that* $0 < \text{dist}(m) \leq \text{dist}(x)$ *for every* $x \in K$.

   Fix $\varepsilon > 0$, *and let* $\delta = \varepsilon$. *For any* $x, y \in K$ *with* $\|x - y\| \leq \delta$ *it holds that* $|\text{dist}(x) - \text{dist}(y)| \leq \text{dist}(x, y) \leq \varepsilon$ *by triangle inequality. Hence,* dist *is continuous. Since K is a closed, bounded subset of* $\mathbb{R}^n$, *by the extreme value theorem there exists* $m \in K$ *such that* $\text{dist}(m) = \inf_{x \in K} \text{dist}(x) \leq \text{dist}(y)$ *for all* $y \in K$. *Since* $m \in K$ *and* $v \notin K$, *we must have* $\text{dist}(m) > 0$.

[6] The extreme value theorem says that a continuous function on a closed, bounded subset of $\mathbb{R}^n$ attains its supremum and infimum.

2. *Let* $f_v : K \to \mathbb{R}$ *be defined as* $f_v(x) = \langle m - v, x\rangle - \frac{\|m\|_2^2 - \|v\|_2^2}{2}$. *Our goal is to prove that* $\{x \mid f_v(x) = 0\}$ *is a good separating hyperplane. Towards this, verify that* $f_v((m+v)/2) = 0$. *Prove that* $f_v(v) < 0$. *Have you used convexity of K yet?*

   We have $f_v((m+v)/2) = \langle m - v, m + v\rangle/2 - \frac{\|m\|_2^2 - \|v\|_2^2}{2} = \frac{1}{2}(\|m\|_2^2 - \|v\|_2^2) - \frac{\|m\|_2^2 - \|v\|_2^2}{2} = 0$. *We have that* $f_v(v) = \langle m - v, v\rangle - \frac{\|m\|_2^2 - \|v\|_2^2}{2} = \frac{1}{2}(2\langle m, v\rangle - \|v\|_2^2 - \|m\|_2^2) = -\frac{1}{2}(\|v - m\|_2^2) < 0$ *since* $\text{dist}(m) > 0$. *We have not used convexity yet!*

3. *We will now prove that* $f_v(x) > 0$ *for every* $x \in K$. *Suppose towards a contradiction that there exists a* $y \in K$ *such that* $f_v(y) \leq 0$. *Prove that* $\nabla \text{dist}^2(m)^\top (y - m) < 0$. *That is, "moving" in the direction of y decreases* dist^2.

   We have that $\nabla \text{dist}^2(m) = 2(m - v)$. *So,* $\nabla \text{dist}^2(m)^\top (y - m) = 2(\langle m - v, y\rangle - \langle m - v, m\rangle) = 2f_v(y) + (\|m\|_2^2 - \|v\|_2^2) - 2\langle m - v, m\rangle \leq -\|v - m\|_2^2 < 0$. *Note that this inequality is strict as* $v \notin K$ *and so* $v \neq m$.

4. *Conclude that there is a positive $1 > \alpha > 0$ such that $\text{dist}(m + \alpha(y - m)) < \text{dist}(m)$. Derive a contradiction to finish the proof.*

Think of $\text{dist}^2(m + \alpha(y - m))$ as a univariate function $g(\alpha)$. It is differentiable at $\alpha = 0$ with derivative $g'(0) = \nabla\text{dist}^2(m)^\top(y - m) < 0$. Fix $\varepsilon = |g'(0)|/2$. Using the definition of the derivative, there exists $\delta > 0$ such that for any $0 < \alpha < \delta$, we have $|\frac{g(\alpha)-g(0)}{\alpha} - g'(0)| \le \varepsilon$, which implies that $g(\alpha) \le g(0) + \alpha(g'(0) + \varepsilon) \le g(0) - \alpha|g'(0)|/2 < g(0)$, as $g'(0) < 0$. Hence, $\text{dist}^2(m + \alpha(y - m)) = g(\alpha) < g(0) = \text{dist}^2(m)$. Taking square roots shows the claim. Therefore, $m + \alpha(y - m)$ is strictly closer to v than m, which is a contradiction as m is the closest point to v in K, which finishes the proof.

Exercise 7 (Separation Oracle for PSD Matrices)**.** *1. Prove that $\mathbb{S}_+ = \{M \mid M \succeq 0\} \subseteq \mathbb{R}^{n\times n}$ is a closed, convex set. Use the definition of a closed set given in the previous exercise, where the distance between two matrices M_1 and M_2 is $\|M_1 - M_2\|_F$.*

Convexity follows since for any $M_1, M_2 \in \mathbb{S}_+$ and $\alpha \in [0, 1]$, we have for all $x \in \mathbb{R}^n$ that $x^\top(\alpha M_1 + (1 - \alpha)M_2)x = \alpha(x^\top M_1 x) + (1 - \alpha)(x^\top M_2 x) \ge 0$, and so $\alpha M_1 + (1 - \alpha)M_2 \succeq 0$.

Suppose that N is a limit point of $\mathbb{S}_+$. Namely, there is a sequence $M_1, M_2, \ldots, M_k, \ldots$ such that $M_k \to N$ in Frobenius norm. Fix $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, and let $a_k = x^\top M_k x$ for all k. Let $b = x^\top N x$. Observe that $|a_k - b| = |x^\top(M_k - N)x^\top| \le \|M_k - N\|_2 \le \|M_k - N\|_F \to 0$, and so $a_k \to b$. Since a_k is a sequence of positive numbers that converges to b, it follows that $b \ge \inf_k a_k \ge 0$. Hence, $x^\top N x \ge 0$, and so $N \succeq 0$.

Note: choosing between spectral and Frobenius norm doesn't matter here, because if $M_k \to N$ in spectral norm then $\|M_k - N\|_F \le \sqrt{n}\|M_k - N\|_2 \to 0$ also, as n is fixed.

2. *Suppose $N \in \mathbb{R}^{n\times n}$ is not positive semidefinite. Show that there's a polynomial time algorithm to compute a matrix B such that $\sum_{i,j} B_{i,j}M_{i,j} \ge 0$ for every $M \in \mathbb{S}_+$ while $\sum_{i,j} B_{i,j}N_{i,j} < 0$.*

The algorithm is to run SVD on N and output $B = vv^\top$, where v is any eigenvector of N with negative eigenvalue.

We have that $\sum_{i,j} B_{i,j}N_{i,j} = \sum_{i,j} v_i v_j N_{i,j} = v^\top N v < 0$ since v has negative eigenvalue, and for every $M \in \mathbb{S}_+$ we have that $\sum_{i,j} B_{i,j}M_{i,j} = v^\top M v \ge 0$ since $M \in \mathbb{S}_+$.