NYU CS-GY 6763: Algorithmic Machine Learning and Data Science

Homework 4

Name: Solution key

Problem 1

(a). Let $V\Sigma V^T$ be the eigendecomposition of A. We have that $A^k = V\Sigma V^T V\Sigma V^T, \ldots, V\Sigma V^T$. All of the V^TV terms cancel to be identities, so in the end we have that $A^k = V\Sigma^k V^T$. At this point you can use cyclic property of the trace, or more directly, observe that $V\Sigma^k V^T$ is an eigendecomposition of A^k. So $\text{tr}(A^k) = \sum_{i=1}^n \Sigma_{ii}^k = \sum_{i=1}^n \lambda_i^k$.

(b) Since there are m terms in the sum, the claim follows if you can compute x^TBx for any vector x in $O(n^2k)$ time. This can be done by multiplying from right to left. I.e. to compute $A \cdot A \cdot \ldots \cdot A \cdot x$ first compute $A \cdot x$, then multiply this on the left by A, then repeat k times. Each matrix-vector multiplication takes $O(n^2)$ time for a total of $O(n^2k)$ time .

(c) By linearity, we only need to compute the expectation and variance of x^TBx for one x to compute the claim. For let's do expectation:

$$x^TBx = \sum_{i=1}^n \sum_{j=1}^n x_i x_j B_{ij}$$

so by linearly

$$\mathbb{E}[x^TBx] = \sum_{i=1}^n \sum_{j=1}^n B_{ij}\mathbb{E}[x_i x_j].$$

Since X_i and X_j are random ±1's, $x_ix_j = 1$ if $i = j$ and otherwise itself is a random ±1, and thus has expectation zero. So we have:

$$\mathbb{E}[x^TBx] = \sum_{i=1}^n \sum_{j=1}^n B_{ij}\mathbb{E}[x_i x_j] = \sum_{i=1}^n B_{ii} \cdot 1 = \text{tr}(B)$$

Next let's do variance. First note that, as above, we can write $x^TBx = \sum_{i=1}^n \sum_{j\neq i} x_i x_j B_{ij} + \sum_{i=1}^n B_{ii}$. The second part is a constant, so the variance of x^TBx is just equal to the variance of $\sum_{i=1}^n \sum_{j\neq i} x_i x_j B_{ij}$. We can *almost* evaluate this using linearity of variance because x_ix_j and $x_\ell x_k$ are independent as long as one of ℓ, k differ from i, j. However, the variables will not be indepdent if $\ell = j$ and $k = i$ – in fact in that case, x_ix_j and $x_\ell x_k$ are the same random variable. To deal with this issue, we regroup:

$$[x^TBx] = \left[\sum_{i=1}^n \sum_{j\neq i} x_i x_j B_{ij}\right] = \left[\sum_{i=1}^n \sum_{j>i} x_i x_j (B_{ij} + B_{ji})\right].$$

Now we can check that every term in the sum truly is independent, and so by linearity of variance we have:

$$[x^T Bx] = \sum_{i=1}^{n} \sum_{j>i} (B_{ij} + B_{ji})^2 [x_i x_j] \leq \sum_{i=1}^{n} \sum_{j>i} 2B_{ij}^2 + 2B_{ji}^2 = 2\|B\|_F^2$$

In the last inequality we used that $[x_i x_j] = 1$ and also the AM-GM inequality – ie that $(a+b)^2 = a^2 + 2ab + b^2 \leq a^2 + a^2 + b^2 + b^2$.

It follows that if we average m repeated trials of $x^T Bx$ we get variance $2\|B\|_F^2/m$.

(d) Can can apply Chebyshev's inequality.

(e) First we note that, when A is PSD, so is B. In particular, B's eigenvalues are equal to those of A raised to the k^{th} power, so are all positive.

So, we prove the claim for a generic PSD matrix B. Let $\lambda_1, \ldots, \lambda_n$ denote the matrix's eigenvalues. As discussed in class, $\|B\|_F^2 = \sum_{i=1}^n \lambda_i^2$ and $\text{tr}(B)^2 = (\sum_{i=1}^n \lambda_i)^2 = \sum_{i=1}^n \lambda_i^2 + \sum_{i \neq j} \lambda_i \lambda_j$. Since B is PSD, all λ_i, λ_j are positive, so $\sum_{i \neq j} \lambda_i \lambda_j \geq 0$ and thus we have that $\|B\|_F^2 \leq \text{tr}(B)^2$ as required.

Problem 2

1. Run gradient descent for q steps and collect all intermediate results $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(q)}$. Note that this takes $O(ndq)$ time – $O(nd)$ time for each gradient step to multiply a vector by $\mathbf{A}^T\mathbf{A}$, and we run q steps. Now, given a q degree polynomial p with coefficients $c_0, c_1, \ldots, c_q$, form the vector:

$$\mathbf{y} = c_0 \mathbf{x}^{(0)} + c_1 \mathbf{x}^{(1)} + \ldots + c_q \mathbf{x}^{(q)}.$$

By the fact the $\mathbf{x}^{(i)} = (\mathbf{I} - 2\eta \mathbf{A}^T\mathbf{A})(\mathbf{x}^{(i-1)} - \mathbf{x}^*) + \mathbf{x}^*$, we can check that $\mathbf{y}$ satisfies:

$$\mathbf{y} = p\left(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{A}^T\mathbf{A}\right)(\mathbf{x}^{(0)} - \mathbf{x}^*) + (c_0 + c_1 + \ldots + c_q)\mathbf{x}^*.$$

And since we assume the coefficients summed to 1, we have that, as desired,

$$\mathbf{y} - \mathbf{x}^* = p\left(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{A}^T\mathbf{A}\right)(\mathbf{x}^{(0)} - \mathbf{x}^*).$$

2. We use the polynomial p from Claim 4 of the Lanczos notes. $p(1) = 1$ for that polynomial, and since $1^q = 1$ for all q, $p(1) = c_0 + \ldots + c_q$, so this polynomial satisfies the coefficient requirement. Let $\lambda_1 \geq \ldots, \geq \lambda_d \geq 0$ be the eigenvalues of the PSD matrix $\mathbf{A}^T\mathbf{A}$. I.e. the diagonal entries of $\mathbf{\Lambda}$ in the eigendecomposition $\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. Since $\mathbf{I} = \mathbf{V}\mathbf{I}\mathbf{V}^T$, we have that:

$$\mathbf{I} - \frac{1}{\lambda_1}\mathbf{A}^T\mathbf{A} = \mathbf{V}(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{\Lambda})\mathbf{V}^T$$

and

$$p(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{A}^T\mathbf{A}) = \mathbf{V}p(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{\Lambda})\mathbf{V}^T$$

The entries of the diagonal matrix $\mathbf{I} - \frac{1}{\lambda_1}\mathbf{\Lambda}$ lie between 0 at the smallest and and $1 - \lambda_d/\lambda_1$ at the largest. So, by Claim 4, as long as p is chosen to have degree $O(\sqrt{\lambda_1/\lambda_d})$, the values in $p(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{\Lambda})$ are all less than ϵ. $p(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{A}^T\mathbf{A}) = \mathbf{V}p(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{\Lambda})\mathbf{V}^T$ is thus a matrix with eigenvalues bounded by ϵ in absolute value.

3. By part 2, we have $\|\mathbf{y} - \mathbf{x}^*\|_2 = \|p\left(\mathbf{I} - \frac{1}{\lambda_1}\mathbf{A}^T\mathbf{A}\right)(\mathbf{x}^{(0)} - \mathbf{x}^*)\|_2 \leq \epsilon\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2$ as long as we use degree $q = O(\sqrt{\lambda_1/\lambda_d})$ (i.e. run for q iterations).

Problem 3

1. Assume without loss of generality that $S = \{1, \ldots, k\}$. $E[A]$ has its top left $k \times k$ block equal to all ones, and all other entries equal to p.

2. Let $\bar{A}$ denote $\mathbb{E}[A]$. The hint must hold because the first k rows of $\bar{A}$ are identical. So for any eigenvector v, it must be that the first k entries of $\bar{A}v$ are all identical. But we have $\bar{A}v = \lambda v$, so this implies that the first k entries of v itself are identical. The same goes for the remaining $n-k$ entries since the bottom $n-k$ rows of A are all identical. Since scaling won't impact whether or not v is an eigenvector, we can always rescale so that the last $n-k$ entries are equal to 1.

Given the form of the eigenvector above, we note that $\bar{A}v$ can be written as

$$\begin{bmatrix} 1 & \ldots & 1 & p & \ldots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \ldots & 1 & p & \ldots & p \\ p & \ldots & p & p & \ldots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \ldots & p & p & \ldots & p \end{bmatrix} \begin{bmatrix} \alpha \\ \vdots \\ \alpha \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} k\alpha + (n-k)p \\ \vdots \\ k\alpha + (n-k)p \\ kp\alpha + (n-k)p \\ \vdots \\ kp\alpha + (n-k)p \end{bmatrix} \tag{1}$$

For v to be an eigenvector, we need that $Av = \lambda v$ for some constant λ, and thus that:

$$k\alpha + (n-k)p = \lambda\alpha$$
$$kp\alpha + (n-k)p = \lambda$$

Now we solve for α. Multiplying the bottom equation by α and subtracting off the first we get:

$$[kp]\,\alpha^2 + [(n-k)p - k]\,\alpha - [(n-k)p] = 0.$$

This is a quadratic equation, so use the quadratic formula to find that:

$$\alpha = \frac{-\left[(n-k)p - k\right] \pm \sqrt{\left[(n-k)p - k\right]^2 + 4\left[(n-k)p^2k\right]}}{2kp}$$

These two solutions for α immediately give our two eigenvectors, and make it clear there are only 2!

3. First we observe that $\mathbb{E}[A]$ is positive semidefinite since it can be written as $p \cdot \vec{1}\vec{1}^T + (1-p) \cdot \vec{1}_k\vec{1}_k$, where $\vec{1}$ denotes the all ones vector and $\vec{1}_k$ denotes a vector that is 1 in it's first k entries and

zeros elsewhere. So, both of its non-zero eigenvalues are positive. Accordingly, to find the largest magnitude eigenvalue, we just need to find the most positive eigenvalue of A. To do so, note that $\lambda = kp\alpha + (n-k)p$ is always more positive for large values of α. So, the more positive eigenvalue corresponds to taking the "+" in the quadratic equation. Once we do, it's clear that $\alpha > 0$ because the term inside the square root is greater than $[(n-k)p - k]^2$, and thus the top of the fraction evaluates to a positive number.

Once we know that $\alpha > 0$, we have that $\lambda\alpha = k\alpha + (n-k)p > kp\alpha + (n-k)p = \lambda \cdot 1$ since $p < 1$ and $k\alpha$ is positive. Since $\lambda > 0$ if $\lambda\alpha \geq \lambda$ then $\alpha > 1$.

Problem 4 (15 pts)

We follow the steps laid out in the hint.

Without loss of generality, assume $\|x\|_2 = 1$. Then $x^T Rx/x^T x$ equals:

$$x^T Rx = \sum_{i=1}^{n} \sum_{j=i+1}^{n} 2R_{ij}x_i x_j + \sum_{i=1}^{n} R_{ii}x_i^2.$$

Note that $\mathbb{E}[x^T Rx] = 0$, so by a Hoeffding bound, we have that:

$$\Pr[|x^T Rx| \geq t] \leq 2e^{\frac{-2t^2}{\sum_{i=1}^n \sum_{j=i+1}^n 16x_i^2 x_j^2 + \sum_{i=1}^n 4x_i^4}}$$

Note that $\sum_{i=1}^n \sum_{j=i+1}^n 16x_i^2 x_j^2 + \sum_{i=1}^n 4x_i^4 \leq 8(x_1^2 + \ldots + x_n^2)^2 = 8\|x\|_2^4$. So if we set $t = c\sqrt{q \log n}$ for sufficiently large c, we have that

$$\Pr[|x^T Rx| \geq t] \leq 1/n^q$$

for any constant c_1 we desire.

Next we do the ϵ-net argument. A sticking point for students is they often try to follow the argument from class a bit too closely, but the argument here can be simpler.

Construct an ϵ-net $\mathcal{N}_\epsilon$ over the unit ball so that, for any unit vector $x \in \mathbb{R}^n$, there is a vector $w \in \mathcal{N}_\epsilon$ such that $\|x - w\|_2 \leq \epsilon$. We will choose ϵ to be *really small* – on the order of $1/n^c$, noting that this doesn't hurt out net-size by much. Let $x - w = e$. Consider $|x^T Rx|$:

$$\begin{aligned} |x^T Rx| = |w^T Rw + e^T Re + 2w^T Re| &\leq |w^T Rw| + (2w + e)^T Re \\ &\leq |w^T Rw| + \|2w + e\|_2 \|Re\|_2 \\ &\leq |w^T Rw| + 3\|Re\|_2. \end{aligned}$$

The second to last step follows from Cauchy-Schwarz, and the last from triangle inequality assuming $\epsilon \leq 1$. We have that $\|Re\|_2 \leq \epsilon\|R\|_2$ based on the first definition of the spectral norm given in the problem, and trivially that $\|R\|_2 \leq n$. So, if we set $\epsilon \leq 1/n$, we have that:

$$|x^T Rx| \leq |w^T Rw| + O(1).$$

Note that $\mathcal{N}_\epsilon$ has size $(cn)^n$, so if we set $q = O(n)$ above, then we have that, with high probability, $|w^T Rw| \leq c\sqrt{q \log n}$ for all $w \in \mathcal{N}_\epsilon$. Combine with the above reduction for all x, this gives the result.