MATH 547 Homework 1

Qilin Ye

January 28, 2022

Problem 1

Let $x^{(1)}, \ldots, x^{(m)}$ be m vectors with $\|x^{(i)}\| = 1$ for all $1 \leqslant i \leqslant m$. Let $\epsilon > 0$. Assume that $m > (1 + 2/\epsilon)^n$. Show that there exists $i, j \in \{1, \ldots, m\}$ such that $\|x^{(i)} - x^{(j)}\| < \epsilon$. Consequently, the vectors $x^{(i)}, x^{(j)}$ are highly correlated so that $\left\langle x^{(i)}, x^{(j)} \right\rangle > 1 - \epsilon^2/2$.

Proof. Let $\{x^{(i)}\}_{i=1}^m \subset B(0,1)$ be given with $m > (1 + 2/\epsilon)^n$. Suppose that for all $i \neq j$, $\|x^{(i)} - x^{(j)}\| > \epsilon$. Then, the balls $B(x^{(i)}, \epsilon/2)$ are pairwise disjoint and are all contained in $B(0, 1 + \epsilon/2)$. Therefore their disjoint union is also contained in $b(0, 1 + \epsilon/2)$. Computing volumes,

$$(1 + 2/\epsilon)^n (\epsilon/2)^n < m(\epsilon/2)^n \leqslant (1 + \epsilon/2)^n$$

whereas the first term is nothing but $(1 + \epsilon/2)^n$. This clearly gives a contradiction, so there must exist $i \neq j$ with $\|x^{(i)} - x^{(j)}\| < \epsilon$. □

Problem 2

Let $A_{m \times n}$ be given with $m \geqslant n$. Show that A has rank n if and only if $A^T A$ is positive definite.

Proof. Note that $A^T A$ is always PSD since for all $x \in \mathbb{R}^n$,

$$x^T A^T A x = (Ax)^T (Ax) = \|Ax\|^2 \geqslant 0.$$

Now suppose A has rank n; that is, if $0 \neq x \in \mathbb{R}^n$ then $Ax \neq 0 \in \mathbb{R}^m$, so $x^T A^T A x = \|Ax\|^2 > 0$.
Conversely suppose $A^T A$ is PD. Then, for all $x \neq 0$, $\|Ax\|^2 = x^T A^T A x \neq 0$, so $Ax \neq 0$. This is precisely the characterization of A having rank n. □

Problem 3

Let $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$. Let $y \in \mathbb{R}^n$. Show that

$$\sum_{j=1}^m \left\| x^{(j)} - \frac{1}{m} \sum_{k=1}^m x^{(k)} \right\|^2 \leqslant \sum_{j=1}^m \|x^{(j)} - y\|^2.$$

That is, the barycenter is the point in $\mathbb{R}^n$ that minimizes the sum of squared distances.

Proof. Define $f : \mathbb{R}^m \to [0, \infty)$ by

$$f(y) = f((y_1, ..., y_n)) = \sum_{j=1}^{m} \|x^{(j)} - y\|^2.$$

The gradient is then given by $\nabla f = \sum_{j=1}^{m} 2(x^{(j)} - s)$, and a critical point exists at

$$\tilde{y} := \frac{1}{m} \sum_{k=1}^{m} x^{(k)}.$$

To show that $\tilde{y}$ is the unique minimizer of f, note that

$$\begin{aligned} f((y_1, ..., y_n)) &= \sum_{j=1}^{m} \sum_{k=1}^{n} (x_k^{(j)} - y_k)^2 \\ &= \sum (x_k^{(j)})^2 + \sum y_k^2 - 2 \sum \sum x_k^{(j)} y_k \\ &= \text{constant} + \text{convex} - \text{linear} = \text{ convex}, \end{aligned}$$

from which we conclude that f attains its unique minimum at $\tilde{y}$. □

Problem 4

Let $n \geqslant 2$ and let S^{n-1} be the boundary of the n-dimensional ball. Let $x \in S^{n-1}$ be fixed and let v be a random vector uniformly distributed in S^{n-1}. Prove that

$$\mathbb{E}|\langle x, v\rangle| \geqslant \frac{1}{10\sqrt{n}}.^1$$

Proof. First we reduce the claim to a much simpler case. Since the uniform distribution on S^{n-1} is invariant under rotations about the origin, and since inner product is also preserved under rotations, i.e., $\langle a, b\rangle = \langle Ra, Rb\rangle$, we have, for any rotation $R : S^{n-1} \to S^{n-1}$,

$$\mathbb{E}|\langle x, v\rangle| = \mathbb{E}|\langle x, Rv\rangle| = \mathbb{E}|\langle R^{-1}x, R^{-1}Rv\rangle| = \mathbb{E}|\langle R^{-1}x, v\rangle|.$$

For any $x \in S^{n-1}$, letting R be such that $R^{-1}x = u := (1, 0, \cdots, 0)$, we have

$$\mathbb{E}|\langle x, v\rangle| = \mathbb{E}|\langle u, v\rangle| = \frac{1}{\text{Area}(S^{n-1})} \int_{S^{n-1}} |v_1| \, \mathrm{d}V. \tag{Q9.1}$$

Note that, under spherical coordinates with parameters $r, \varphi_1, \varphi_2, ..., \varphi_{n-1}$, the first component v_1 can be expressed as $r \cos \varphi_1$, and the Jacobian is

$$r^{n-1} \prod_{i=1}^{n-2} \sin^{n-1-i}(\varphi_i) = r^{n-1} \sin^{n-2}(\varphi - 1) \sin^{n-3}(\varphi_2) \cdots \sin(\varphi_{n-2}).$$

In this case $r \equiv 1$ on S^{n-1} so we get two simpler $(n-1)$-fold integrals:

$$\int_{S^{n-1}} |v_1| \, \mathrm{d}V = \int_{\varphi_{n-1}=0}^{2\pi} \int_{\varphi_{n-2}=0}^{\pi} \cdots \int_{\varphi_1=0}^{\pi} |\cos \varphi_1| \prod_{i=1}^{n-2} \sin^{n-1-i}(\varphi_i) \, \mathrm{d}\varphi_1 \cdots \mathrm{d}\varphi_{n-2} \, \mathrm{d}\varphi_{n-1} \tag{Q9.2}$$

and

$$\text{Area}(S^{n-1}) = \int_{S^{n-1}} 1 \, \mathrm{d}V = \int_{\varphi_{n-1}=0}^{2\pi} \int_{\varphi_{n-2}=0}^{\pi} \cdots \int_{\varphi_1=0}^{\pi} \prod_{i=1}^{n-2} \sin^{n-1-i}(\varphi_i) \, \mathrm{d}\varphi_1 \cdots \mathrm{d}\varphi_{n-2} \, \mathrm{d}\varphi_{n-1}. \tag{Q9.3}$$

[1] The solution is copied from 541a HW1.

Division gives $\mathbb{E}|\langle x, v\rangle| = (\text{Q9.2})/(\text{Q9.3}) = \int_0^\pi |\cos\varphi| \sin^{n-2}\varphi \, \mathrm{d}\varphi \Big/ \int_0^\pi \sin^{n-2}\varphi \, \mathrm{d}\varphi$. Since both integrals satisfy $\int_0^\pi = 2\int_0^{\pi/2}$, the ratio further equals $\int_0^{\pi/2} \cos\varphi \sin^{n-2}\varphi \, \mathrm{d}\varphi \Big/ \int_0^{\pi/2} \sin^{n-2}\varphi \, \mathrm{d}\varphi$. The numerator is $1/(n-1)$ by a simple u-substitution with $u := \sin\varphi$, and for $n \geqslant 3$, the denominator is bounded by 0 and $\sqrt{\pi/2(n-2)}$ since $\cos x \leqslant \exp(-x^2/2)$ on $[0, \pi/2]$ and

$$\begin{aligned}\int_0^{\pi/2} \sin^{n-2}\varphi \, \mathrm{d}\varphi = \int_0^{\pi/2} \cos^{n-2}\varphi \, \mathrm{d}\varphi &\leqslant \int_0^{\pi/2} \exp(-(n-2)x^2/2) \, \mathrm{d}x \\ &< \int_0^{\infty} \exp(-(n-2)x^2/2) \, \mathrm{d}x = \frac{1}{2} \cdot \sqrt{2\pi/(n-2)}.\end{aligned}$$

For $n = 2$, it is immediate that

$$\int_0^{\pi/2} \sin^2\varphi \, \mathrm{d}\varphi = \frac{1}{2}\int_0^{\pi/2} \sin^0\varphi \, \mathrm{d}\varphi = \pi/4$$

using the well-known reduction formula

$$\int_0^{\pi/2} \sin^k\varphi \, \mathrm{d}\varphi = \frac{k-1}{k}\int_0^{\pi/2} \sin^{k-2}\varphi \, \mathrm{d}\varphi.$$

Therefore, for $n = 2$, $\mathbb{E}|\langle x, v\rangle| = 4/\pi > 1/(10\sqrt{2})$ and for $n \geqslant 3$,

$$10\sqrt{n} \cdot \mathbb{E}|\langle x, v\rangle| \geqslant \frac{20}{\sqrt{\pi}} \cdot \left(\frac{n^2-2n}{n^2-2n+1}\right)^{1/2} \geqslant \frac{20\sqrt{3/4}}{\sqrt{\pi}} > 1.$$

(Note that $(n^2-2n)/(n^2-2n+1)$ is monotone on $[3, \infty)$ and equals $3/4$ at 3.) This proves the claim. □

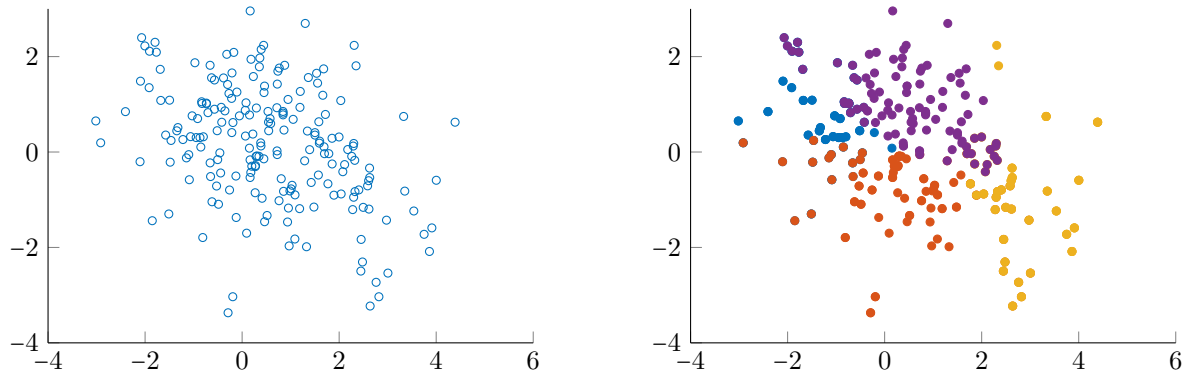
Problem 6

Run a k-means clustering algorithm on a planted data set in $\mathbb{R}^2$ consisting of samples from different Gaussian distributions. Also run a k-means algorithm on Airline Safety Information.

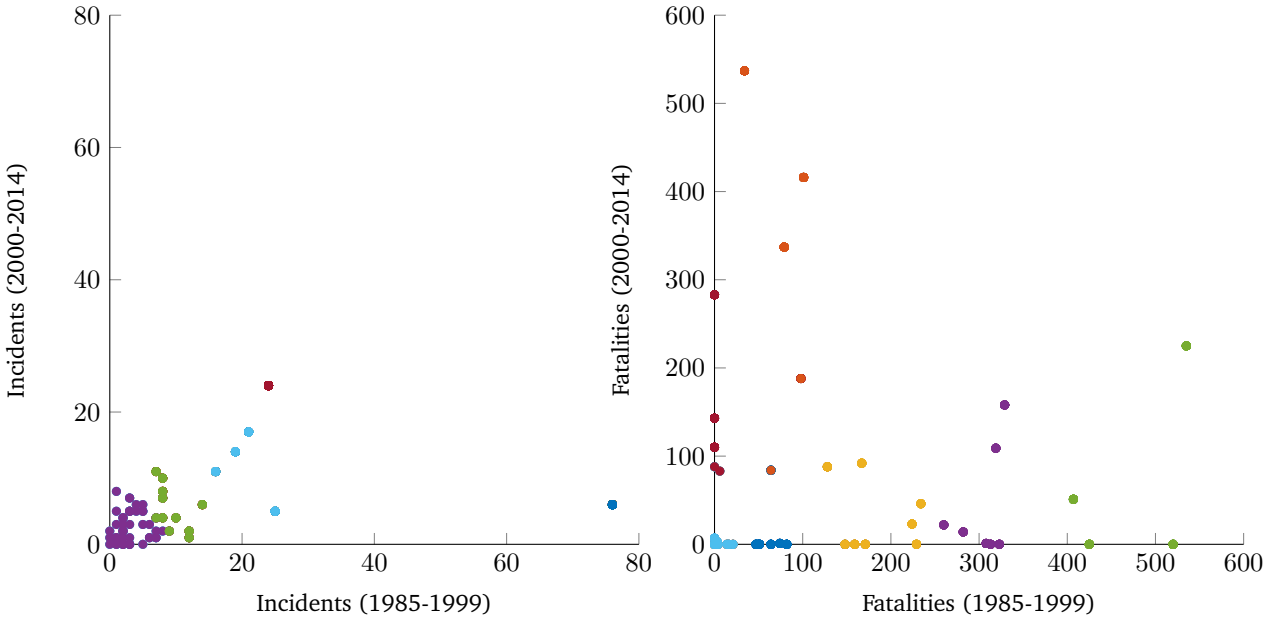
Solution. To make things look a bit nicer, I used four Gaussian pairs, all with variance 1 but centered at $(0,0)$, $(1,1)$, $(-1,1)$, and $(2,-1)$, respectively. I sampled each Gaussian 50 times.

```
% (code to generate Gaussians)
dot_list = [X1,X2,X3,X4; Y1,Y2,Y3,Y4];
dot_index = linspace(1,200,200);

subplot(1,2,1);
scatter([X1,X2,X3,X4],[Y1,Y2,Y3,Y4]);

T = zeros(4,100);
y = randn(2,4);

for p = 1:5
    for j = 1:length(dot_list)
        tempdiff = vecnorm(y - repmat(dot_list(:,j),1,4));
        [tempdiff, min_val] = min(tempdiff);
        loc = min(find(~T(min_val,:)));
        T(min_val, loc) = j;
    end

```

```
    for j = 1: length(y)
        temp = T(j,:);
        temp = temp(temp~=0);
        y(:,j) = sum(dot_list(:,temp),2) / nnz(temp);
    end
end

subplot(1,2,2);
hold on
for j = 1:4
    temp = T(j,:);
    temp = temp(T(j,:)~=0);
    scatter(dot_list(1,temp), dot_list(2,temp),'filled');
end
matlab2tikz('gaussian_k_means.tex')
```



For the airline safety information, below are two plots — the one on left plots airlines' number of incidents during 2000 and 2014 (y-axis) against that during $1985-1999$ (x-axis); the second plot plots number of fatalities instead.



It seems like the fatality of an airline between 2000 and 2014 cannot be predicted by the data between 1985 and 1999, but the numbers of incidents seem to behave nicer, and they are potentially correlated. (I did not do regression though.) Assuming this is true, a possible explanation is that there are much more factors that influence the number of fatalities. For example, "big" airline disasters and "small" ones both count as one, whereas the fatalities are drastically different (e.g. PAL434 killed person, whereas JAL123 killed 520 and Tenerife killed 583 (two flights)).

Problem 7

Let n be a positive integer and let c_n be the number of boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ that are linear threshold functions. Use an inductive argument and prove the lower bound

$$c_n \geqslant 2^{n(n-1)/2}.$$

Proof. The base case $c_1 = 4 \geqslant 2^0$ is trivial. Since

$$2^{(n+1)n/2} = 2^{n(n-1)/2}2^n,$$

it suffices to prove that a k-variable linear threshold function $f : \{-1,1\}^k \to \{-1,1\}$ can be uniquely extended to 2^k linear threshold functions defined on $\{-1,1\}^{k+1}$.
Define $H_1 := \{-1,1\}^k \times \{-1\}$ and $H_2 := \{-1,1\}^k \times \{1\}$ so that H_1, H_2 are both k-dimensional hypercubes whereas they together form a $(k+1)$-dimensional hypercube.
For any f given above, let H_1 be its corresponding $(k-1)$-dimensional hyperplane P_{k-1}. For each point in $v \in H_2$, we can define a k-dimensional hyperplane, P_k, containing both P_{k-1} and the point v. We then shift the plane slightly so that it no longer contains v, and the corresponding linear threshold function f_v satisfies

(1) $f_v(x) = f(x)$ for all $x \in H_1$,

(2) $f_v(v) = 1$ (before the perturbation, $f_v(v) = 0$), and

(3) $f_v(x) = -1$ for all $x \in H_2 \backslash \{v\}$.

Note that every edge (i.e., a line segment connecting two points in H_2 that are distance 1 apart) of H_2 is parallel to H_1. Therefore, each $v \in H_2$ corresponds to a *unique* $(k+1)$-variable linear threshold function. That is, each k-variable linear threshold function can be extended in $|H_2| = 2^k$ ways. This proves the claim. □

Problem 8

Let $a > 0$. Let $X^{(1)}, \ldots, X^{(k)} \in \mathbb{R}^n$ be i.i.d. Gaussian random vectors with mean ae_1 (where $e_1 := (1, 0, \ldots, 0) \in \mathbb{R}^n$) and identity covariance matrix. Let $X^{(k+1)}, \ldots, X^{(2k)} \in \mathbb{R}^n$ be i.i.d. Gaussian random vectors with mean $-ae_1$. As in perceptron, define

$$\mathcal{B} := \max_{i \leqslant 2k} \|X^{(i)}\|$$

and

$$\Theta := \min \left\{ \|w\| : y_i \left\langle w, X^{(i)} \right\rangle \geqslant 1 \text{ for all } i \right\}.$$

(If such w does not exist, instead define $\Theta := \infty$.)
Define $y_1 = \ldots = y_k := 1$ and $y_{k+1} = \ldots = y_{2k} := -1$. Estimate $\mathbb{E}\mathcal{B}$ and $\mathbb{E}(1/\Theta)$ in terms of a.

Solution. Since $\mathcal{B}$ and Θ are both nonnegative, we use $\mathbb{E}\mathcal{B} = \int_0^\infty \mathbb{P}(\mathcal{B} > t)\,\mathrm{d}t$ and $\mathbb{E}\Theta = \int_0^\infty \mathbb{P}(\Theta > t)\,\mathrm{d}t$. Firstly,

$$\mathbb{P}(\mathcal{B} > t) \leqslant \sum_{i=1}^{2k} \mathbb{P}(|X^{(i)}| > t) = 2k\mathbb{P}(|X^{(1)}| > t) \leqslant 2k\mathbb{P}(|X_1^{(1)}| > t) = 2k\mathbb{P}(|\mathcal{N}(a,1)| > t),$$

so viewing $|\mathcal{N}(a,1)|$ as a "folded" Gaussian random variable and splitting $\mathbb{R}$ into $(-\infty, 0]$ and $(0,\infty)$, we have

$$\begin{aligned}
\mathbb{E}\mathcal{B} &\leqslant 2k\int_0^\infty \mathbb{P}(|\mathcal{N}(a,1)| > t)\,\mathrm{d}t \\
&= 2k\int_{-\infty}^0 \mathbb{P}(\mathcal{N}(a,1) < t)\,\mathrm{d}t + 2k\int_0^\infty \mathbb{P}(\mathcal{N}(a,1) > t)\,\mathrm{d}t \\
&= 2k\int_{-\infty}^0 |t| f_{\mathcal{N}(a,1)}(t)\,\mathrm{d}t + 2k\int_0^\infty t f_{\mathcal{N}(a,1)}(t)\,\mathrm{d}t \\
&\leqslant 2ka + 2ka = 4ka.
\end{aligned}$$

For an estimation of Θ, note that by independence

$$\begin{aligned}
\mathbb{P}(y_i\langle w, X^{(i)}\rangle \geqslant 1 \text{ for all } i) &= \left[\mathbb{P}(y_1\langle w, X^{(1)}\rangle \geqslant 1)\right]^{2k} = \left[\mathbb{P}(\langle w, X^{(1)}\rangle \geqslant 1)\right]^{2k} \\
&= \left[\mathbb{P}(\langle w/\|w\|, X^{(1)}\rangle \geqslant \|w\|^{-1})\right]^{2k} \leqslant \left[\mathbb{P}(\langle e_1, X^{(1)}\rangle \geqslant \|w\|^{-1})\right]^{2k} \\
&= \left[\mathbb{P}(\mathcal{N}(a,1) \geqslant \|w\|^{-1})\right]^{2k} = \left[\int_{\|w\|^{-1}}^\infty f_{\mathcal{N}(a,1)}(t)\,\mathrm{d}t\right]^{2k}.
\end{aligned}$$

If $\|w\|^{-1} \geqslant a$, then $\|w\| \leqslant 1/a$, and

$$\mathbb{P}(y_i\langle w, X^{(i)}\rangle \geqslant 1 \text{ for all } i) \leqslant \left[\int_{\|w\|^{-1}}^\infty f_{\mathcal{N}(a,1)}(t)\,\mathrm{d}t\right]^{2k} \leqslant \left[\int_a^\infty f_{\mathcal{N}(a,1)}(t)\,\mathrm{d}t\right]^{2k} = 2^{-2k}.$$

Thus $\mathbb{P}(\min\|w\| > 1/a) \leqslant 2^{-2k}$, so

$$\begin{aligned}
\mathbb{P}(1/\Theta \geqslant a) &= \mathbb{P}(\Theta \leqslant 1/a) = \mathbb{P}(\min\|w\| \leqslant 1/a) \\
&= 1 - \mathbb{P}(\min\|w\| > 1/a) \geqslant 1 - 2^{-2k}.
\end{aligned}$$

Using Markov's inequality we have

$$\mathbb{E}(1/\Theta) \geqslant a\mathbb{P}(1/\Theta \geqslant a) \geqslant a - a2^{-2k}.$$