This document lays out some extra problems, for Math 240 students who want a special challenge. These problems are realistic, in that they connect to actual applications or actual higher math, and they use a variety of math to get there. Because they draw on so much math and require students to absorb so many new concepts, they are not suitable for Math 240 exams, or even as study questions for Math 240 exams. Anyway, let me know if you solve any of these problems, or if you want to discuss them.

1 Quantum Computing

Computers can manipulate many kinds of data: numbers, text, audio, etc. At a fundamental level, all of these data are just strings of bits, with each bit being 0 or 1. A n-bit string can be regarded as a vector $\vec{v}$ of dimension n over the integers modulo 2. For example, here is the sum of the 8-bit strings 01110100 and 11011110:

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

Scalar multiplication is really simple, because there are only two scalars: $0\vec{v} = \vec{0}$, and $1\vec{v} = \vec{v}$. And here is an example of the dot product:

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} = 3 = 1.$$

When $\vec{v} \cdot \vec{w} = 0$, we say that $\vec{v}$ and $\vec{w}$ are *orthogonal*. This is all quite similar to linear algebra over the real numbers, but orthogonality requires a warning: Two orthogonal vectors are not necessarily independent. In fact, the first 8-bit string above is orthogonal to itself.

For the rest of this problem, fix a positive integer n. Imagine that it's large. We work with

n-bit strings, or equivalently n-dimensional vectors over the integers modulo 2. Fix a vector $\vec{x}$, and let P be the set of vectors orthogonal to $\vec{x}$.

Exercise 1.1 *How many vectors of dimension n are there, overall? How many vectors are there in P? How many vectors are there, in a k-dimensional vector subspace of P?*

In a certain quantum computing problem, the vector $\vec{x}$ is unknown, and we're trying to discover it. We have a secret weapon: a quantum algorithm, which, each time it is invoked, produces a uniformly random vector $\vec{r}$ in P. By invoking the quantum algorithm many times, we eventually acquire $n-1$ such bit strings $\vec{r}_0, \ldots, \vec{r}_{n-2}$ that are independent. We place them into the rows of an $(n-1) \times n$ matrix R, and solve $R\vec{x} = \vec{0}$ to obtain the desired $\vec{x}$. Make sense?

The sticky point is the word "independent". We need $n-1$ independent $\vec{r}$s, but the quantum algorithm does not necessarily produce independent $\vec{r}$s. Some of the $\vec{r}$s, that it produces, are redundant with $\vec{r}$s that we already have. Using standard linear algebra calculations, we can easily detect these redundant $\vec{r}$s and get rid of them. But it does mean that some of our invocations of the quantum algorithm are wasted. We'd like to put an upper bound, on the number of invocations that we need, to discover $\vec{x}$. But in theory there is no upper bound. So let's upper-bound the *expected* number of invocations.

Exercise 1.2 *Let k be one of $0, 1, \ldots, n-2$. Suppose that we already have k independent vectors $\vec{r}_0, \vec{r}_1, \ldots, \vec{r}_{k-1}$. What is the probability that the next invocation of the algorithm produces an $\vec{r}$ that is independent of $\vec{r}_0, \vec{r}_1, \ldots, \vec{r}_{k-1}$? So what is the expected number of invocations needed to get the next usable $\vec{r}$, which is denoted $\vec{r}_k$?*

Exercise 1.3 *What is the expected number of invocations needed, in total, to discover $\vec{x}$? In the end, my answer consists of two terms: the minimum number of invocations possible, plus a summation that captures the excess, wasted invocations.*

Exercise 1.4 *Show that the expected number of excess, wasted invocations is less than 2.*

2 Structural Geology

This problem relates to a data visualization technique that's used in geology. For reasons that we won't go into, each data point is a point on the unit sphere. Given a data set of such points, it is natural to ask, "Did these data come from a uniform distribution on the sphere?" The answer is subtle, because most data sets exhibit some clumping, even if they come from a uniform distribution. So it's valuable to have a way of measuring non-uniformity.

Let $\mathbb{S}^2$ be the unit sphere centered on the origin in $\mathbb{R}^3$. In other words, $\mathbb{S}^2 = \{\vec{p} : |\vec{p}| = 1\}$. To be clear, when I talk of a point "on the sphere", I mean a point $\vec{p}$ such that $|\vec{p}| = 1$. For example, the origin $\vec{0}$ is not a point on the sphere.

When I talk of a distance "along the sphere", I'm talking about great-circle distance, which is usually greater than the straight-line distance in $\mathbb{R}^3$. For example, consider the points $\vec{p} = (1,0,0)$ and $\vec{q} = (0,1,0)$. The straight-line distance between them is $\sqrt{2}$, but the distance along the sphere is $\pi/2$.

Let $\vec{p}$ be a point on the sphere, and let $r > 0$ be a number. Consider all points $\vec{q}$ on the sphere, whose distance from $\vec{p}$ along the sphere is less than or equal to r. The set of such points $\vec{q}$ forms a "curved disk", like a bowl or a contact lens or a kippah. The number r is the "radius" of the curved disk.

Exercise 2.1 *What is the area of a flat disk of radius $r = \pi/2$? What is the area of a curved disk of radius $r = \pi/2$? Which one is greater? (Do not use the next exercise. Use a simpler kind of reasoning.)*

Exercise 2.2 *Show that the area of a curved disk of radius r is $2\pi(1-\cos r)$. (My solution uses a double integral in spherical coordinates.)*

Now we start doing probability. Fix a point $\vec{p}$ and a radius r about $\vec{p}$. Suppose that we have n data points $\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_n$ on the sphere. Let X be the number of these data points that lie in the curved disk of radius r about $\vec{p}$. This X is a random variable.

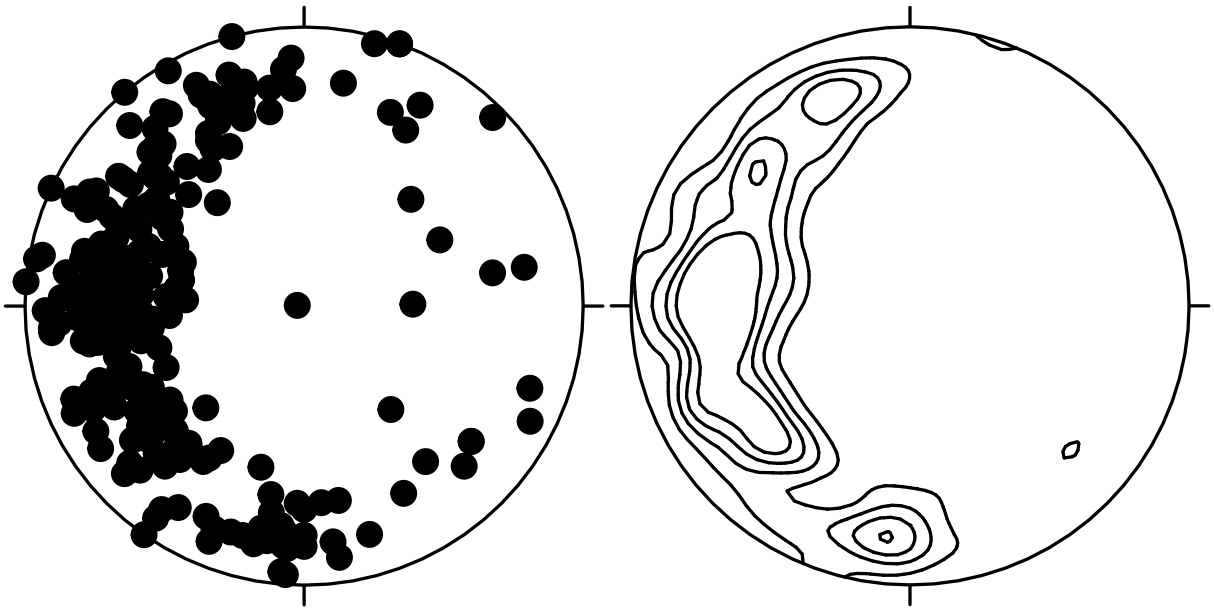
Exercise 2.3 *What is the distribution of X, under the assumption that the n data points are uniformly random on the sphere? Your answer should be in terms of n and r.*

Let $k = 3$. We'll work in terms of k, instead of simply 3, in case we change our minds about what value k should have. But we'll always keep $k^2 < n$. (If $n \leq k^2$, then the data set is too small for this technique to be applied.) Anyway, we're going to use this k to choose r.

Exercise 2.4 *Still under the assumption that the data are uniformly random, what should r be, if we want to make it so that $E(X) = kSD(X)$? Your answer should be in terms of n and k.*

Now that we have r nailed down, we can define the "density" of any point $\vec{p}$: The density is the number d such that there are $d \cdot SD(X)$ data points in the curved disk of radius r about $\vec{p}$.

Now that we have a notion of density, we can program all of this into a computer. For example, we can make a contour plot of the density. Glossing over some details of how we visualize a sphere, an example is shown below. The left plot shows $n = 257$ data. The right plot shows the $d = 3, 6, 9, 12$ contours.



The foregoing problem is a slight simplification of what geologists actually use. Here's the real thing. For any point $\vec{p}$ on the sphere, its antipode $-\vec{p}$ is also on the sphere. Let's declare an equivalence relation, that $\vec{p} \sim -\vec{p}$ for all $\vec{p}$. The space of equivalence classes is called the *real projective plane* and denoted $\mathbb{RP}^2$. Locally, on a small scale, $\mathbb{RP}^2$ looks a lot like $\mathbb{S}^2$, but globally, viewed as a whole, $\mathbb{RP}^2$ is "half as big as" $\mathbb{S}^2$ and has a different *topology*. (The plots above are of $\mathbb{RP}^2$ rather than $\mathbb{S}^2$, actually. Sorry for misleading you earlier.)

Exercise 2.5 *Modify the foregoing problem and its solution, so that it uses* $\mathbb{RP}^2$ *instead of* $\mathbb{S}^2$*.*

3 Functional Analysis

For any $\sigma > 0$, let f_σ be the PDF of the normal distribution with mean 0 and standard deviation σ:

$$f_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}.$$

The graph of f_σ is the (in)famous bell curve. Its inflection points are at $x = \pm\sigma$.

Exercise 3.1 *Describe how the shape of the bell curve changes, as* σ *decreases toward* 0*. And how does its integral change? So what happens in the limit, as* $\sigma \to 0$*? (My answer is a short paragraph of mostly plain English rather than mathematical notation.)*

Let D be the set of all continuous functions $f : \mathbb{R} \to \mathbb{R}$. Because you can add and scale functions, D is a vector space. Its dimension is infinite. For what follows, we need to work in some subspace of D — namely, the subspace consisting of all functions that make the integrals

below well-defined. But please ignore this technicality. Just imagine that D is a vector space of nice functions, that have all of the nice properties that we need.

We're going to study linear transformations $T : D \to \mathbb{R}$. One way to make such a linear transformation is to start with an $f \in D$, and define $T_f : D \to \mathbb{R}$ by

$$T_f(g) = \int_{-\infty}^{\infty} f(x)g(x)\,dx.$$

Exercise 3.2 *Check that T_f is a linear transformation, for all $f \in D$.*

Now, for any $g \in D$, we can approximate the function $y = g(x)$ by the constant function $y = g(0)$. This is a crude approximation, but it turns out to be good enough.

Exercise 3.3 *Use the approximation to get an approximation for $T_{f_\sigma}(g)$, and argue that it's a good approximation for small values of σ. (If you can't figure out what this problem is asking, then you might want to look ahead to the next exercise. In any event, you are not expected to rigorously prove anything about how good the approximation is.)*

The moral of this story is: Although there is no function f_0 such that $f_\sigma \to f_0$, there is a linear transformation T_{f_0} such that $T_{f_\sigma} \to T_{f_0}$. In fact, f_0 is a famous function that doesn't actually exist, and T_{f_0} is a famous way to make it sort-of exist.

Exercise 3.4 *What is T_{f_0}, explicitly? Once you've defined it, verify that it's a linear transformation.*

4 Statistical Mechanics, Information Theory

Let X be a continuous random variable with PDF f. For simplicity, we assume that the support of f is the entire real line $(-\infty, \infty)$. Define the *entropy* $H(X)$ to be

$$H(X) = -\int_{-\infty}^{\infty} f(x) \log f(x)\,dx,$$

where the logarithm has base e. Entropy is a concept from thermodynamics, statistical mechanics, and information theory. High-entropy random variables are of interest to physicists, because many physical systems evolve over time into high-entropy configurations. They are also of interest to statisticians, because in a Bayesian statistical analysis you often want your prior distribution to contain minimal information, which (in one view, and if I understand correctly) corresponds to maximum entropy.

Exercise 4.1 *Show that $H(X + c) = H(X)$ for any constant c.*

Fix a number $\sigma^2 > 0$ for the rest of this problem. The preceding exercise says that, in computing the entropy of $X \sim \text{Norm}(\mu, \sigma^2)$, we might as well assume that $\mu = 0$.

Exercise 4.2 *Compute the entropy of* $X \sim \text{Norm}(0, \sigma^2)$.

It can be proved that, among all distributions with support $(-\infty, \infty)$ and variance σ^2, the one with greatest entropy is the normal distribution. We now do much, but not all, of the proof. We use a style of argument called *calculus of variations*. Let f be the PDF of $X \sim \text{Norm}(0, \sigma^2)$. Let g be any other PDF of support $(-\infty, \infty)$, mean 0, and variance σ^2.

Exercise 4.3 *Show that, for any* $t \in [0, 1]$, *the convex combination* $(1-t)f + tg$ *is also a PDF of support* $(-\infty, \infty)$, *mean* 0, *and variance* σ^2.

Let $h = g - f$, so that $f + th = (1-t)f + tg$. In an abuse of notation, let's write $H(f + th)$ to mean the entropy of a random variable whose PDF is $f + th$.

Exercise 4.4 *Show that* $\frac{d}{dt} H(f(x) + th(x))$ *is zero at* $t = 0$.

Exercise 4.5 *Show that* $\frac{d}{dt}\frac{d}{dt} H(f(x) + th(x))$ *is negative at* $t = 0$.

Because these facts hold for all g, it follows that f is a local maximum for the entropy — well, almost. Where is the gap in the logic? And we certainly have not shown that f is a *global* maximum for entropy. But let's stop here.

5 Internet Search

A vector $\vec{p} \in \mathbb{R}^n$ is called a *probability vector* if the entries of $\vec{p}$ are non-negative and sum to 1. Such a $\vec{p}$ can be thought of as a PMF on the set $\{1, 2, \ldots, n\}$. For example, if $p_i = 1/n$ for all i, then $\vec{p}$ describes the uniform distribution.

Let M be an $n \times n$ real matrix, subject to one requirement: For any probability vector $\vec{p}$, the vector $M\vec{p}$ is also a probability vector.

Exercise 5.1 *Each column of* M *satisfies a rather strict condition. What is it?*

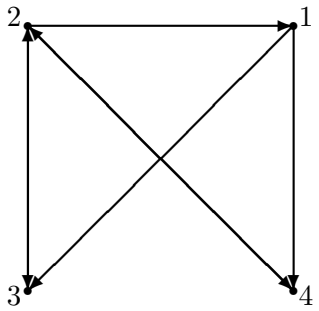
From any starting probability vector $\vec{p}$, we can consider the sequence of probability vectors

$$\vec{p}, \quad M\vec{p}, \quad M^2\vec{p}, \quad M^3\vec{p}, \quad \ldots.$$

In fancier language, we're talking about a *discrete-time dynamical system* with *transition function* M and *initial condition* $\vec{p}$. A probability vector $\vec{q}$ such that $M\vec{q} = \vec{q}$ is an *equilibrium* of the system.

Dynamical systems like this arise in statistical mechanics, but let's talk about Google instead. The founders of Google realized that these ideas could be applied to the World Wide Web. In this application, n is the number of Web pages, and p_i is the probability that a hypothetical random person is viewing the ith page at a specific time. When the person is done reading a page, they (uniformly randomly) click a link on that page and start reading the linked page. The link structure of the Web is encoded into M, so that if $\vec{p}$ captures the probability distribution at one specific time, then $M\vec{p}$ captures the probability distribution one click later in time.

Exercise 5.2 *The image below shows a toy example of a web with $n = 4$ pages and seven links among them. What is M in this example?*



At an equilibrium $\vec{q}$, the proportion of hypothetical readers who click out of each page is exactly balanced by the proportion of hypothetical readers who click into that page. The number q_i rates how popular the ith web page is, on a scale from 0 (nobody visits it ever) to 1 (everybody reads it all the time). This rating system is what made Google what it is.

So let's think about equilibria. And it doesn't help, mathematically at least, to focus on the Google version. Let's do it in general.

Exercise 5.3 *Prove that 1 is an eigenvalue of M. (Hint: Analyze $M^\top$ instead.)*

Because 1 is an eigenvalue of M, it must have an associated eigenvector $\vec{q}$, which satisfies $M\vec{q} = \vec{q}$. So that's an equilibrium, right? Oops, there's an issue. How do we know that $\vec{q}$ is a probability vector? If the entries of $\vec{q}$ are all non-negative or all non-positive, then we can scale $\vec{q}$ to get an eigenvector that's also a probability vector. But how do we know that the entries are all non-negative or all non-positive?

Probably there's an elementary proof that resolves this issue, but I haven't been able to find it. I know only a non-elementary proof, which uses a theorem from topology. In Carleton's Math 354, I prove low-dimensional versions of this theorem, but it also holds in higher dimensions, which we need here. So attempt the following exercise, only if you've taken such a course.

Exercise 5.4 *Prove that there exists a probability vector $\vec{q}$ such that $M\vec{q} = \vec{q}$. (Hint: Let $X \subseteq \mathbb{R}^n$ be the space of all probability vectors.)*