Homework 2: SP&LA Study Group

Guillermo Angeris

December 2023

A general note

None of the solutions to these exercises should be more than a few lines long. If you find yourself writing more than half a page or so for a solution, there's probably a simpler way! (Of course, *finding* the simpler way might not, itself, be so simple.) Problem parts beginning with an **(H)** are harder parts. The problems in this homework correspond to §2-3 of the paper 'Succinct Proofs and Linear Algebra' by Evans and Angeris; you are encouraged to read through this section, and watch session 2 of the SPLA Study Group sessions before starting this homework.

Solutions. The solutions (if available) can be found at the following Github repository:

`https://github.com/angeris/spla-repo`

If you find any errors or would like to write a solution, please raise an issue (if there are errors) or open a pull request (if you have a fix for the errors, or if you would like to write a solution).

1 How randomness helps

In this problem, we'll explore how randomness can be used to greatly reduce the number of queries needed to verify a certain task to a reasonable precision.

In this problem, we are given a vector $x \in \mathbf{F}^n$ which we would like to ensure has no more than q nonzero entries. (In other words, we would like to ensure that at least $n-q$ entries are equal to zero, or that the vector is *sparse* since many entries are zero, with high probability.) For the remainder of the problem, assume that q is given as part of the formulation.

Part 1. If we uniformly randomly sample a single entry of x, say x_r, and we find that it is zero, how likely is it that the vector x has more than q nonzero entries? (To "uniformly randomly sample an entry of x" means that we, uniformly randomly, pick an index $r = 1, \ldots, n$ and 'look at' entry x_r.) You should give a simple bound in terms of q and n. Additionally, write this statement, and its resulting probability bound, as a probabilistic implication.

Part 2. Let's say we repeat the above procedure (uniformly sample an entry of x, check if it is zero) ℓ times; assume that the randomness between each attempt is independent. If every attempt does indeed 'pass' (*i.e.*, every time we draw r, we find that $x_r = 0$) then what's the probability that the vector x has more than q nonzero entries? The bound should be relatively simple and should depend on q, n, and ℓ. Write this statement as a probabilistic implication.

Hint. For fun, there's two ways of solving this part: one uses a direct approach, one uses the 'conjunctions' from §1.4 of the paper. Both ultimately reduce to the same thing, in this case, but it might be a useful exercise to see why!

Part 3. From before, we know the size of the vector n and we want to ensure that it has no more than q nonzero entries with probability at least $1 - p$. How many times do we have to repeat the check from part 1 to have this assurance? (In the notation of part 2, how large does ℓ have to be?) Picking a concrete set of numbers, if $q = n/10$ (*i.e.*, we want to ensure that at least 90% of the entries are zero), the number of entries is $n = 2^{20}$, and the probability of error is no more than $p = 2^{-100}$, then how many times do we need to repeat the check from part 1? Compare this with the result of the paper in §2.1.1.

Part 4. Finally, we'll see the importance of randomness for these checks! A *deterministic check* looks at a fixed set of ℓ entries and verifies that each is zero, instead of uniformly randomly sampling ℓ entries to check. This check succeeds if all entries are zero and fails if any are not. Argue why, for any proposed deterministic check, there exists some vector x which always causes it to incorrectly succeed, unless $\ell \geq n - q$. (In other words, argue why ℓ, the number of entries any deterministic check needs to look at, is at least the number of entries we wanted to check are zero, $n - q$.) Using the same q, n, and p as part 3, compare the number of entries we need to look at in the deterministic check.

2 Checking equality

In this problem, we'll use some of the definitions in the paper and show how the zero check and the sparsity check can be used in a variety of different settings. Similar to the paper, assume that we are given a vector $x \in \mathbf{F}^n$ and a matrix $G \in \mathbf{F}^{m \times n}$ with distance $d > 0$.

Part 1. If $x \neq 0$ then show that the probability that $(Gx)_r \neq 0$, for a uniformly random $r = 1, \ldots, m$, is at least

$$\frac{d}{m},$$

using the definition of distance. (See homework 1, problem 4, or §1.2.2 of the paper.) Argue why this is the same as the following probabilistic implication:

$$(Gx)_r = 0 \quad \underset{p}{\Longrightarrow} \quad x = 0,$$

with $p \leq 1 - d/m$.

Part 2. Argue that this bound is *tight*: that the probability p given above cannot be improved. To do this, show that there is some $x \in \mathbf{F}^n$ such that the probability the check fails is exactly $1 - d/m$.

Hint. Take a second look at the definition of distance!

Part 3. One way of understanding the zero check above is that it is a way of checking that $x = 0$ with high probability by checking, instead, that $(Gx)_r = 0$ for randomly sampled $r = 1, \ldots, m$. An *equality check* instead seeks to show that, given another vector $y \in \mathbf{F}^n$, we have $x = y$ with high probability. Show that the equality check can be easily performed using a zero check.

Hint. Use the fact that matrix multiplication is a linear operation.

Part 4. The analogue of the zero check in the direct access model is the *sparsity check*, which we discussed in problem 1. (If you have not done problem 1, see §3.2.1 in the paper.) In this case, we would instead like to check that two vectors are close in norm, *i.e.*,

$$\|x - y\|_0 \leq q,$$

for some q. Argue how we can use the sparsity check to verify the above statement with high probability.

3 A tale of two models

As referenced in lecture, there are some things that are possible in the coding model that are not useful to attempt in the direct access model. In this problem, we'll compare and contrast the two models in a variety of ways.

Part 1. Show that the coding model 'contains' the direct access model. More specifically, show that there is some matrix $G \in \mathbf{F}^{m \times n}$ such that randomly sampling elements of Gx is the same as randomly sampling elements of $x \in \mathbf{F}^n$. (What 'same' means here is that, for fixed x, the distributions of $(Gx)_r$, over r uniform from $1, \ldots, m$, and $x_{r'}$, with r' uniform over $1, \ldots, n$, are the same for any choice of x. There is at least one answer that is very simple and will not require you to argue anything about distributions or probability.) In other words, there is some matrix G such that the coding model behaves identically to the direct access model! As an aside, note that there are many such matrices: can you characterize some of them?

Part 2. The exact zero check turns out to be very informative in seeing just how different these models can be. Let's say we have a vector $x \in \mathbf{F}^n$ and a matrix $G \in \mathbf{F}^{m \times n}$ with distance $d > 0$ in the coding model. From the zero check (problem 2), we know that

$$(Gx)_r = 0 \quad \underset{p}{\Longrightarrow} \quad x = 0,$$

where $p \leq 1 - d/m$. If we take G to be a Reed–Solomon code matrix, then $d = m - n + 1$ (see §1.3.2 of the paper), so the probability of failure is

$$p \leq \frac{n+1}{m}. \tag{1}$$

On the other hand, let $G' \in \mathbf{F}^{m' \times n}$ be a matrix you found in part 1, such that we are 'emulating' the direct access model. Give a simple expression for the distance of G', written $d' > 0$, in terms of m' and n. Using that distance, compare the probability of failure of the zero check using this matrix, which is $1 - d'/m'$ from before, with the probability of failure given in (1).

Part 3. In many applications, n is somewhat large, say $n = 2^{20}$ and m is much larger, say $m = 2^{255}$. Using the expressions you derived for G' in part 3, argue why, in these applications, the exact zero check in the direct access model would very likely fail. Compare this with the probability of failure in the coding model, where G is a Reed–Solomon code matrix.

(H) Part 4. Show that even if the zero check, using a matrix G' with distance $d > 0$ from part 2, is repeated ℓ times, then ℓ needs to be around the order of n in order to fail with probability no more than $1/2$.

Hint. You are free to use the fact that $1 - 1/z \leq \log(z)$ whenever $z > 0$, where $\log(z)$ is the natural logarithm of z.

(H) Part 5. The 'bound' we derived in part 2 for the direct access model was for a specific choice of matrix G'. (Unless you're an overachiever.) Of course, we could imagine that there is some 'magical' choice of matrix G' such that (a) the coding model emulates the direct access model, as in part 1, yet (b) is better than the error probability you found in part 3. We will show that this is not the case for any finite field $\mathbf{F}$ with at least two elements.

Writing part 1 out, we said that the coding model, with matrix G', *emulates* the direct access model if, for any $x \in \mathbf{F}^n$, the distributions of $(G'x)_r$ and $x_{r'}$ are identical over uniform randomness r and r'. That is, for any $\alpha \in \mathbf{F}$, we have

$$\Pr((G'x)_r = \alpha) = \Pr(x_{r'} = \alpha),$$

where r is uniformly drawn from $1, \ldots, m'$, and r' is uniformly drawn from $1, \ldots, n$. Show that this implies that the distance d' of G' always satisfies

$$\frac{d'}{m'} \leq \frac{1}{n}.$$

In other words, the error bound is no better than the one for the matrix you found in part 1. This, combined with part 4, shows that we essentially always require on the order of n samples in the direct access model to check if a vector is exactly equal to zero, $x = 0$; *i.e.*, we might as well just check the whole vector! This also shows why checking for 'sparsity' is much more useful than attempting to check for exact equality in the direct access model: one can be done with far fewer queries.

4 Reduced matrix zero check

In lecture we saw the (vector) zero check and the matrix zero check, and an interesting suggestion: is it possible to combine the two? In this problem, we will show how to do so and one interesting extension. For this problem, assume that $x \in \mathbf{F}^n$ and $G \in \mathbf{F}^{m \times n}$, where the matrix G has distance $d > 0$. We will also assume that we are given a matrix $Y \in \mathbf{F}^{n \times k}$, with columns $y_1, \ldots, y_k$, that we wish to check is zero, along with a second matrix $G' \in \mathbf{F}^{m' \times k}$ with distance $d' > 0$.

Part 1. The (vector) zero check states that

$$(Gx)_r = 0 \quad \underset{p}{\Longrightarrow} \quad x = 0,$$

where $p \leq 1 - d/m$ and r is uniformly selected from $1, \ldots, n$. The matrix zero check states that

$$G'_{r'1} y_1 + G'_{r'2} y_2 + \cdots + G'_{r'k} y_k = 0 \quad \underset{p'}{\Longrightarrow} \quad Y = 0,$$

where $p' \leq 1 - d'/m'$ and r' is uniformly selected from $1, \ldots, m'$. Note that the quantity on the left hand side of the matrix zero check is a vector. Using probabilistic implications, show how one can apply the vector zero check to this quantity and give a simple bound on the error probability in terms of p and p'. This should give you a result that lets you check that a matrix Y is zero by checking, instead, that a single field element is equal to zero.

(H) Part 2. Let $m = m'$, $k = n$, and let G and G' now be Reed–Solomon code matrices over evaluation points $\alpha_1, \ldots, \alpha_m$. Give a simple description, in terms of polynomials over the field $\mathbf{F}$, of the check you found in part 1. In terms of the maximum degree of the polynomials, what is the probability of error you get from part 1? (Compare this with the bound from the Schwartz–Zippel lemma; see Wikipedia for a description.) In fact, this bound is not tight and it is possible to do better! See the discussion in §3.1.3 and appendix B in the paper for more.