REVIEW SHEET FOR MIDTERM 1: BASIC

MATH 196, SECTION 57 (VIPUL NAIK)

We will not be going over this sheet, but rather, we'll be going over the advanced review sheet in the session. Please review this sheet on your own time.

The summaries here are identical with the executive summaries you will find at the beginning of the respective lecture notes PDF files. The summaries are not intended to be exhaustive. Please review the original lecture notes as well, especially if any point in the summary is unclear. The section titles have the same names as the names of the corresponding lecture notes.

1. Linear functions: a primer

Words ...

(1) Linear models arise both because some natural and social phenomena are intrinsically linear, and because they are computationally tractable and hence desirable as approximations, either before or after logarithmic and related transformations.

(2) Linear functions (in the affine linear sense) can be characterized as functions for which all the second-order partial derivatives are zero. The second-order pure partial derivatives being zero signifies linearity in each variable holding the others constant (if this condition is true for each variable separately, we say the function is (affine) multilinear). The mixed partial derivatives being zero signifies *additive separability* in the relevant variables. If this is true for every pair of input variables, the function is completely additively separable in the variables.

(3) We can use logarithmic transformations to study multiplicatively separable functions using additively separable functions. For a few specific functional forms, we can make them linear as well.

(4) We can use the linear paradigm in the study of additively separable functions where the components are known in advance up to scalar multiples.

(5) If a function type is linear in the parameters (not necessarily in the input variables) we can use (input,output) pairs to obtain a system of linear equations in the parameters and determine the values of the parameters. Note that a linear function of the variables with no restrictions on the coefficients and intercepts must also be linear in the parameters (with the number of parameters being one more than the number of variables). However, there are many nonlinear functional forms, such as polynomials, that are linear in the parameters but not in the variables.

(6) Continuing the preceding point, the number of well-chosen (input,output) pairs that we need should be equal to the number of parameters. Here, the "well-chosen" signifies the absence of dependencies between the chosen inputs. However, choosing the bare minimum number does not provide any independent confirmation of our model. To obtain independent confirmation, we should collect additional (input,output) pairs. The possibility of modeling and measurement errors may require us to introduce error-tolerance into our model, but that is beyond the scope of the current discussion. We will return to it later.

Actions ...

(1) One major stumbling block for people is in writing the general functional form for a model that correctly includes parameters to describe the various degrees of freedom. Writing the correct functional form is half the battle. It's important to have a theoretically well-grounded choice of functional form and to make sure that the functional form as described algebraically correctly describes what we have in mind.

(2) It's particularly important to make sure to include a parameter for the *intercept* (or *constant term*) unless theoretical considerations require this to be zero.

(3) When dealing with polynomials in multiple variables, it is important to make sure that we have accounted for all possible monomials.

(4) When dealing with piecewise functional descriptions, we have separate functions for each piece interval. We have to determine the generic functional form for each piece. The total number of parameters is the sum of the number of parameters used for each of the functional forms. In particular, if the nature of the functional form is the same for each piece, the total number of parameters is (number of parameters for the functional form for each piece) × (number of pieces).

2. Equation-solving with a special focus on the linear case: a primer

Words ...

(1) We need to solve equations both for determining the parameters in a general functional description of a model, and for using existing models to estimate the values of real-world variables. Equations are also useful in solving max/min problems.

(2) When evaluating equation-solving algorithms for suitability, we should consider time, memory, parallelizability, precision, and elegance.

(3) The dimension of the solution space to a system of equations that has no redundancies or inconsistencies is expected to be (number of variables) - (number of equations). If there are fewer equations than variables, we expect that the system is *underdetermined*. If there are more equations than variables, we expect that the system is *overdetermined*. Generally, this means that either some of the equations are redundant (i.e., they do not provide additional information) or the equations are inconsistent.

(4) For diagonal systems of equations, i.e., systems of equations where each equation involves only one variable, we can solve the equations separately, then string together the coordinates of the solution. The process is parallelizable. The solution set satisfies a *rectangular completion property* (note: this isn't a standard term, it's just a shorthand for something that's hard to describe without full symbolism; you might prefer to think of it as a *coordinate interchange property* if that name connects better with the concept for you).

(5) For triangular systems of equations, we need to solve the equations in a particular order, and if we do so, we have to solve a series of equations in one variable. The solution set need not satisfy the rectangular completion property. In case of multiple solutions, we need to make cases and branch each case. Parallelization is possible between the branches but not between the equations themselves.

(6) The same remarks apply for triangular systems of linear equations, except that there is no branching to worry about.

(7) When we manipulate equations (by adding, subtracting, and multiplying) to obtain new equations, we must remember to keep the old equations around to protect against loss of information. However, keeping *all* the old equations around can result in a memory overload. The general idea is to keep enough of the old equations around that the other old equations that we are discarding can be recovered from the new system. In other words, we want our transformations to be *reversible*.

(8) One manipulation technique is to use one of the equations to eliminate one of the variables by expressing it in terms of the other variables. If we are able to do this systematically all the way through, we would have reduced the original system to a triangular system.

(9) Another manipulation technique is to add and subtract multiples of one equation to another equation. We generally keep the equation that is being multiplied before the addition, and discard the equation that is not being multiplied by anything, because that guarantees reversibility. If the value we are multiplying with is a nonzero constant, we can also keep the other equation instead.

(10) We typically add and subtract equations with the explicit goal of eliminating variables or eliminating difficult expressions in terms of the variables.

Actions ...

(1) Please be sure not to throw out information when manipulating the system. When in doubt, store more rather than less.

(2) Please remember that, as soon as you discover an inconsistency, you can abandon the equation-solving attempt (or the branch you are in). All previous solutions that you have found don't amount to

anything. In a diagonal system, if any one equation in the system has no solution for its corresponding variable, the system as a whole has no solution. In a triangular system, as soon as we discover an equation that has no solution, discard the branch (from the most recent branching point onward).

3. Gauss-Jordan elimination: a method to solve linear systems

Words and actions combined ...

(1) A system of linear equations can be stored using an *augmented matrix*. The part of the matrix that does not involve the constant terms is termed the *coefficient matrix*.
(2) Variables correspond to columns and equations correspond to rows in the coefficient matrix. The augmented matrix has an extra column corresponding to the constant terms.
(3) In the paradigm where the system of linear equations arises from an attempte to determine the parameters of a model (that is linear in the parameters) using (input,output) pairs, the parameters of the model are the new variables, and therefore correspond to the columns. The (input,output) pairs correspond to the equations, and therefore to the rows. The input part controls the part of the row that is in the coeficient matrix, and the output part controls the augmenting entry.
(4) If the coefficient matrix of a system of simultaneous linear equations is in reduced row-echelon form, it is easy to read the solutions from the coefficient matrix. Specifically, the non-leading variables are the parameters, and the leading variables are expressible in terms of those.
(5) In reduced row-echelon form, the system is inconsistent if and only if there is a row of the coefficient matrix that is all zeros with the corresponding augmented entry nonzero. Note that if the system is *not* in reduced row-echelon form, it is *still* true that a zero row of the coefficient matrix with a nonzero augmenting entry implies that the system is inconsistent, but the converse does not hold: the system may well be inconsistent despite the absence of such rows.
(6) In reduced row-echelon form, if the system is consistent, the dimension of the solution space is the number of non-leading variables, which equals (number of variables) - (number of nontrivial equations). Note that the all zero rows give no information.
(7) Through an appropriately chosen sequence of row operations (all reversible), we can transform any linear system into a linear system where the coefficient matrix is in reduced row-echelon form. The process is called Gauss-Jordan elimination.
(8) The process of Gauss-Jordan elimination happens entirely based on the coefficient matrix. The final column of the augmented matrix is affected by the operations, but does not control any of the operations.
(9) There is a quicker version of Gauss-Jordan elimination called Gaussian elimination that converts to row-echelon form (which is triangular) but not reduced row-echelon form. It is quicker to arrive at, but there is more work getting from there to the actual solution. Gaussian elimination is, however, sufficient for determining which of the variables are leading variables and which are non-leading variables, and therefore for computing the dimension of the solution space and other related quantities.
(10) The arithmetic complexity of Gauss-Jordan elimination, in both space and time terms, is polynomial in n. To get a nicely bounded bit complexity (i.e., taking into account the sizes of the numbers blowing up) we need to modify the algorithm somewhat, and we then get a complexity that is jointly poynomial in n and the maximum bit length.
(11) Depending on the tradeoff between arithmetic operations and using multiple processors, it is possible to reduce the arithmetic complexity of Gauss-Jordan elimination considerably.
(12) Gauss-Jordan elimination is not conceptually different from iterative substitution. Its main utility lies in the fact that it is easier to code since it involves only numerical operations and does not require the machine to understand equation manipulation. On the other hand, because the operations are purely mechanical, it may be easier to make careless errors because of the lack of “sense” regarding the operations. To get the best of both worlds, use the Gauss-Jordan elimination procedure, but keep the symbolic algebra at the back of your mind while doing the manipulations.
(13) The Gauss-Jordan elimination process is only one of many ways to get to reduced row-echelon form. For particular structures of coefficient matrices, it may be beneficial to tweak the algorithm a little

(for instance, swapping in an easier row to the top) and save on computational steps. That said, the existence of the Gauss-Jordan elimination process gives us a guarantee that we can reach our goal by providing one clear path to it. If we can find a better path, that's great.

(14) We can use the Euclidean algorithm/gcd algorithm/Bareiss algorithm. This is a variant of the Gauss-Jordan algorithm that aims for the same end result (a reduced row-echelon form) but chooses a different sequencing and choice of row operations. The idea is to keep subtracting multiples of rows from one another to get the entries down to small values (hopefully to 1), rather than having to divide by the leading entry. It's not necessary to master this algorithm, but you might find some of its ideas helpful for simplifying your calculations when solving linear systems.

4. Linear systems and matrix algebra

Words and actions combined ...

(1) The *rank* of a matrix is defined as the number of nonzero rows in its reduced row-echelon form, and is also equal to the number of leading variables. The rank of a matrix is less than or equal to the number of rows. It is also less than or equal to the number of columns.

(2) The rank of the coefficient matrix of a system of simultaneous linear equations describes the number of independent equational constraints in the system.

(3) How far the coefficient matrix is from having full column rank determines the dimension of the solution space if it exists.

(4) How far the coefficient matrix is from having full row rank determines the probability that the system is consistent, roughly speaking. If the coefficient matrix has full row rank, then the system is consistent for all outputs. Otherwise, it is consistent only for some outputs and inconsistent for others.

(5) For a consistent system, the dimension of the solution space equals (number of variables) - (rank).

(6) There is a concept of "expected dimension" which is (number of variables) - (number of equations). Note that if the system does not have full row rank, the expected dimension is less than the actual dimension (if consistent). The expected dimension can be thought of as averaging the actual dimensions over all cases, where inconsistent cases are assigned dimensions of $-\infty$. This is hard to formally develop, so we will leave this out.

(7) There are various terms commonly associated with matrices: zero matrix, square matrix, diagonal, diagonal matrix, scalar matrix, identity matrix, upper-triangular matrix, and lower-triangular matrix.

(8) A vector can be represented as a row vector or a column vector.

(9) We can define a dot product of two vectors and think of it in terms of a sliding snake.

(10) We can define a matrix-vector product: a product of a matrix with a column vector. The product is a column vector whose entries are dot products of the respective rows of the matrix (considered as vectors) with the column vector.

(11) Matrix-vector multiplication is linear in the vector.

(12) A linear system of equations can be expressed as saying that the coefficient matrix times the input vector column (this is the column of unknowns) equals the output vector column (this is the column that would be the last column in the augmented matrix).

5. Hypothesis testing, rank, and overdetermination

Words ...

(1) In order to test a hypothesis, we must conduct an experiment that could conceivably come up with an outcome that would falsify the hypothesis. This relates to Popper's notion of falsifiability.

(2) In the setting where we use a model with a functional form that is linear in the parameters, the situation where the coefficient matrix (dependent on the inputs) does *not* have full row rank is the situation where we can use consistency of the system to obtain additional confirmation of the hypothesis that the model is correct. If the coefficient matrix has full column rank, we can determine the parameters uniquely assuming consistency. The ideal situation would be that we choose inputs

such that the coefficient matrix has full column rank but does not have full row rank. In this situation, we can obtain verification of the hypothesis *and* find the parameter values.

(3) In order to test a hypothesis of a function of multiple variables being affine linear, we could choose three points that are collinear in the input space and see if the outputs behave as predicted. If they do, then this is evidence in favor of linearity, but it is not conclusive evidence. If they do not, this is conclusive evidence against linearity.

(4) If the goal is to find the coefficients rather than to test the hypothesis of linearity, we should try picking independent inputs (in general, as many inputs as the number of parameters, which, for an affine linear functional form, is one more than the number of variables). Thus, the choice of inputs differ for the two types of goals. If, however, we are allowed enough inputs, then we can both find all the coefficients *and* test for linearity.

6. Linear transformations

Words and actions combined...

(1) A $n \times m$ matrix defines a function from $\mathbb{R}^m$ to $\mathbb{R}^n$ via matrix-vector multiplication. A function arising in this manner is termed a *linear transformation*. Every linear transformation arises from a unique matrix, i.e., there is a bijection between the set of $n \times m$ matrices and the set of linear transformations from $\mathbb{R}^m$ to $\mathbb{R}^n$.

(2) A function (also called map) $f : A \to B$ of sets is termed *injective* if no two elements of A map to the same element of B. It is termed *surjective* if B equals the range of f. It is termed *bijective* if it is both injective and surjective. A bijective map has a unique inverse map.

(3) The standard basis vector $\vec{e}_i$ is the vector with a 1 in the i^{th} coordinate and 0s elsewhere. The image of $\vec{e}_i$ is precisely the i^{th} column of the matrix describing the linear tranformation.

(4) A linear transformation can alternatively be defined as a map that preserves addition and scalar multiplication. Explicitly, $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear transformation if and only if $T(\vec{x} + \vec{y}) = T(\vec{x}) + T(\vec{y})$ (additivity) and $T(a\vec{x}) = aT(\vec{x})$ (scalar multiples) for all $\vec{x}, \vec{y} \in \mathbb{R}^m$ and all $a \in \mathbb{R}$.

(5) A linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ is *injective* if the matrix of T has full column rank, which in this case means rank m, because the dimensions of the matrix are $n \times m$. Note that this in particular implies that $m \leq n$. The condition $m \leq n$ is a *necessary but not sufficient condition* for the linear transformation to be injective.

(6) A linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ is *surjective* if the matrix of T has full row rank, which in this case means rank n, because the dimensions of the matrix are $n \times m$. Note that this in particular implies that $n \leq m$. The condition $n \leq m$ is a *necessary but not sufficient condition* for the linear transformation to be surjective.

(7) A linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ is *bijective* if the matrix of T has full row rank and full column rank. Thus forces $m = n$, and forces the (now square) matrix to have full rank. As mentioned before, this is equivalent to invertibility.

(8) The inverse of a diagonal matrix with all diagonal entries nonzero is the matrix obtained by inverting each diagonal entry individually.

(9) A permutation matrix is a matrix where each row has one 1 and all other entries 0, and each column has one 1 and all other entries 0. A permutation matrix acts by permuting the standard basis vectors among themselves. It can be inverted by permuting the standard basis vectors among themselves in the reverse direction. Hence, the inverse is also a permutation matrix. If the permutation matrix just flips two basis vectors, it is self-inverse.

(10) The inverse of a shear operation is the shear operation obtained by using the negative of the shear. In particular:

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

The last two points from the summary of the lecture notes are not directly relevant to the midterm, but are included for completeness below.

(11) Inverting a matrix requires solving a system of simultaneous linear equations with that as coefficient matrix and with the augmenting column a *generic* output vector, then using the expression for the input in terms of the output to get the matrix of the transformation. It can alternatively be done by augmenting the matrix with the identity matrix, row-reducing the matrix to the identity matrix, and then looking at what the augmented side has become.

(12) We can use linear transformations for the purposes of coding, compression, and extracting relevant information. In many of these practical applications, we are working, not over the real numbers, but over the field of two elements, which is ideally suited for dealing with the world of bits (binary digits).

REVIEW SHEET FOR MIDTERM 1: ADVANCED

MATH 196, SECTION 57 (VIPUL NAIK)

Please bring a copy (print or readable electronic) of this sheet to the review session.

There is also a basic review sheet that contains executive summaries of the lecture notes. You should review that on your own time.

Many of the error-spotting exercises correspond to ideas that you have already seen in quiz questions. We have parenthetically indicated at the beginning of the question whether it corresponds to material seen in lecture (we use L for that) or in a quiz (we use Q for that). We preface with a $\sim$ if there is a considerable gap between the way the idea was presented in the earlier context and the way it is being tested now.

1. Linear functions: a primer

Error-spotting exercises ...

(1) (L, Q): f is a function of two variables x and y that is postulated to be *affine* linear, i.e., linear with a possibly nonzero intercept. Since there are two unknowns, knowing the value of f at 2 points will help us find f precisely (assuming it is linear) and also provide independent verification of the linearity of f.

(2) (L, Q): f is a function of one variable. It is believed to be a polynomial of degree at most n, where n is a known positive integer. In order to find the function f (assuming that it is indeed such a polynomial) we need to know the values of f at n points. If the goal is not merely to find f conditional to its being a polynomial of degree at most n but also to obtain independent confirmation of the hypothesis, we need to know the values of f at $n+1$ points.

(3) (Q): Suppose $x \mapsto mx + c$ is a linear function of one variable. We know the value of the function at two points, albeit there are uncorrelated measurement errors with a fixed error distribution at both points. Regardless of what two points we choose, we will end up with the same error range in our estimate of the function.

(4) A function is known to be of the form $t \mapsto A\sin(t + \varphi)$ where A and φ are constants. We need to use input-output pairs to find the function. We can do this by setting up a system of equations that are linear in terms of the parameters A and φ. Given enough input-output pairs, we will be able to determine A and φ uniquely.

2. Equation solving

Error-spotting exercises ...

(1) (L): Consider the system of simultaneous (not necessarily linear) equations:

$$\begin{aligned} f(x) &= 0 \\ g(y) &= 0 \end{aligned}$$

Suppose the set of solutions to the first equation viewed as an equation in x alone is $\{x_1, x_2, \dots, x_n\}$ and the set of solutions to the second equation viewed as an equation in y alone is $\{y_1, y_2, \dots, y_n\}$. Then, the set of solutions to the system as a whole, viewed as a system of equations in two variables, is the following set of points in the xy-plane: (x_1, y_1), (x_2, y_2), (x_3, y_3), $\dots$, (x_n, y_n).

(2) (L): Consider a triangular system of the form:

$$\begin{aligned} f(x) &= 0 \\ g(x,y) &= 0 \\ h(x,y,z) &= 0 \end{aligned}$$

Suppose the following are true:

- The first equation has 2 solutions for x.
- For each solution to the first equation, the second equation has 3 solutions for y.
- For each choice of x and y that solve the first two equations, the third equation has 4 choices for z.

Then, the total number of solutions for the system is $2 + 3 + 4 = 9$.

(3) (L): Consider the system:

$$\begin{aligned} x + e^{x^2 - x - y} &= 1 \\ y + xe^{x^2 - x - y} &= 0 \end{aligned}$$

Subtract x times the first equation from the second equation. We get:

$$y - x^2 = -x$$

Thus, $y = x^2 - x$. The solution set is thus the set of all points on the parabolic curve $y = x^2 - x$.

(4) (Q, ∼L): If we are trying to find an existing function of one or more variables based on some input-output values with some measurement error, and we believe that the function is a polynomial function but we don't know the degree, it makes sense to try to fit using a polynomial function of as large a degree as we can computationally afford. This is because the larger the degree of the polynomial, the easier it is to get a good fit on a given collection of input-output pairs. For instance, given three input-output pairs for a function of one variable, we can always fit them perfectly (with zero measurement error) using a quadratic, but it may be difficult to fit them using a linear function. Larger degree polynomials are better because we have more parameters to work with and they offer us more flexibility. And more flexibility is always good.

3. Gauss-Jordan elimination

Error-spotting exercises ...

(1) (L, Q): Any process to solve an arbitrary linear system with n equations and n variables must take time of the order at least n^3, because Gauss-Jordan elimination is $\Theta(n^3)$. This is true even if we can pre-process the coefficient matrix of the linear system.

(2) (L, ∼Q): If Gauss-Jordan elimination gives us a row in the coefficient matrix that is all zeros, then the system of linear equations cannot have a solution.

(3) (L, Q): The dimension of the solution space to a system of simultaneous linear equations equals the number of leading variables in the system. This number can be effectively computed by converting the coefficient matrix to reduced row-echelon form and counting the number of pivotal 1s. The columns corresponding to these 1s give us the leading variables.

(4) (L): There is a shorter variant of Gauss-Jordan elimination called Gaussian elimination. The idea here is to skip the step of clearing out entries below the pivotal 1s, but concentrate on clearing out all entries above the pivotal 1s. This gets us to a form where it is easy to read the solutions and answer questions about the rank. The form of matrix we obtain at the end of this process is called the *row-echelon form*.

(5) (L, Q): A system of linear equations is inconsistent if and only if there is a row of the coefficient matrix that is all zeros such that the corresponding augmenting entry is nonzero.

4. Linear systems and matrix algebra

Error-spotting exercises ...

1. (L): If the rank of the coefficient matrix of a linear system equals the number of rows, then the system is consistent and has a unique solution. Otherwise, it may or may not be consistent, and if it is consistent, it has infinitely many solutions.
2. (∼L): If the number of columns in the coefficient matrix of a linear system is less than the number of rows, then the system has either no solution or a unique solution. It cannot have infinitely many solutions.
3. (∼L): If the number of rows in the coefficient matrix of a linear system is less than the number of columns, then the system has either a unique solution or infinitely many solutions. In other words, it must be consistent.
4. (∼L): A matrix has full row rank if and only if none of its rows is a zero row *and* no row is a multiple of another row. This is because, if both these conditions are satisfied, no two rows are dependent on each other. From the equational perspective, it means that every pair of equations is independent, so the equations are all independent of each other.
5. (Q): Let m, n, and k be natural numbers with $m \geq 3$. We are given a bunch of numbers $x_0 < x_1 < x_2 < \dots < x_m$ and another bunch of numbers $y_0, y_1, y_2, \dots, y_m$. We want to find a continuous function f on $[x_0, x_m]$, such that $f(x_i) = y_i$ for all $0 \leq i \leq m$, and such that the restriction of f to any interval of the form $[x_i, x_{i+1}]$ (for $0 \leq i \leq m-1$) is a polynomial of degree $\leq n$. Further, we want f to be at least k times differentiable on the open interval (x_0, x_m). Let's try to see how many equations we can get from the constraints.

   We have $m+1$ equations of the form $f(x_i) = y_i$. In addition, at each point of transition, we know that the first k derivatives have to match up. There are $m-1$ transition points and k derivatives to compare, so we get $k(m-1)$ equations that way. In total we thus have $k(m-1)+m+1$ equations. The number of parameters is mn, because we have m pieces, and polynomials of degree n in each piece, so n parameters for the coefficients in each piece.

   Thus, we have $k(m-1)+m+1$ equations in mn variables that we need to solve.

5. Hypothesis testing, rank, and overdetermination

Nothing here (but some of these ideas appear in the error-spotting exercises for earlier sections).

6. Linear transformations

Error-spotting exercises ...

1. (L, Q): Suppose $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear transformation. If $m < n$, T is injective but not surjective. If $m > n$, T is surjective but not injective. If $m = n$, T is bijective, and hence invertible.
2. (L, Q): Let A be a $m \times n$ matrix that defines a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ by $T(\vec{x}) = A\vec{x}$. Denote by $\vec{e}_i$ the vector in $\mathbb{R}^n$ with a 1 in the i^{th} coordinate and 0s elsewhere. Then, $T(\vec{e}_i)$ is the i^{th} row of A.
3. A matrix A is termed *self-inverse* if $A = A^{-1}$. The only self-inverse 2×2 matrices are the diagonal matrices where each diagonal entry is either 1 or -1.

PLEASE TURN OVER FOR THE PRACTICE WORKSHEET.

7. Practice worksheet: guiding questions

(1) I have an unknown function f of one variable. I know $f(-1)$, $f(0)$, $f(1)$, and $f(3)$ (but I haven't told you these values). If I assumed that f was a polynomial of degree $\leq d$, I could use input-output pairs to construct a linear system. The coefficient matrix depends only on the inputs.

Suppose $d = 2$. Can you enumerate the parameters, then construct the coefficient matrix? Row reduce the coefficient matrix, and store all your steps. I will then provide you with an augmenting column, and you should be able to quickly determine f (if it exists).

Employ the same procedure for $d = 3$.

(2) I have a function f that is continuous on $[0, 3]$, differentiable on $(0, 3)$, and piecewise quadratic, with the pieces on which it is quadratic being the $[0, 1]$, $[1, 2]$, and $[2, 3]$. I'm going to give you the values $f(0)$, $f(1)$, $f(2)$, and $f(3)$. I will also give you the right hand derivative at 0. How would you use this information to find f? I'll give actual numerical examples in the session (or perhaps you can give each other actual numerical examples in the session).

(3) What are the properties of $n \times m$ matrices with randomly chosen entries in terms of whether the system is consistent, and what the dimension of the solution space is? Start by noting that the rank is almost certainly $\min\{m, n\}$. A general rule of thumb is that anything whose truth requires two independent random values to coincide will almost certainly not happen.

REVIEW SHEET FOR MIDTERM 2: BASIC

MATH 196, SECTION 57 (VIPUL NAIK)

We will not be going over this sheet, but rather, we'll be going over the advanced review sheet in the session. Please review this sheet on your own time.

The summaries here are identical with the executive summaries you will find at the beginning of the respective lecture notes PDF files. The summaries are not intended to be exhaustive. Please review the original lecture notes as well, especially if any point in the summary is unclear.

1. Matrix multiplication and inversion

Note: The summary does not include some material from the lecture notes that is not important for present purposes, or that was intended only for the sake of illustration.

(1) *Recall*: A $n \times m$ matrix A encodes a linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ given by $T(\vec{x}) = A\vec{x}$.

(2) We can add together two $n \times m$ matrices entry-wise. Matrix addition corresponds to addition of the associated linear transformations.

(3) We can multiply a scalar with a $n \times m$ matrix. Scalar multiplication of matrices corresponds to scalar multiplication of linear transformations.

(4) If $A = (a_{ij})$ is a $m \times n$ matrix and $B = (b_{jk})$ is a $n \times p$ matrix, then AB is defined and is a $m \times p$ matrix. The $(ik)^{th}$ entry of AB is the sum $\sum_{j=1}^{n} a_{ij}b_{jk}$. Equivalently, it is the dot product of the i^{th} row of A and the k^{th} column of B.

(5) Matrix multiplication corresponds to composition of the associated linear transformations. Explicitly, with notation as above, $T_{AB} = T_A \circ T_B$. Note that $T_B : \mathbb{R}^p \to \mathbb{R}^n$, $T_A : \mathbb{R}^n \to \mathbb{R}^m$, and $T_{AB} : \mathbb{R}^p \to \mathbb{R}^m$.

(6) Matrix multiplication makes sense *only if* the number of columns of the matrix on the left equals the number of rows of the matrix on the right. This comports with its interpretation in terms of composing linear transformations.

(7) Matrix multiplication is associative. This follows from the interpretation of matrix multiplication in terms of composing linear transformations and the fact that function composition is associative. It can also be verified directly in terms of the algebraic definition of matrix multiplication.

(8) Some special cases of matrix multiplication: multiplying a row with a column (the inner product or dot product), multiplying a column with a row (the outer product or Hadamard product), and multiplying two $n \times n$ diagonal matrices.

(9) The $n \times n$ identity matrix is an identity (both left and right) for matrix multiplication wherever matrix multiplication makes sense.

(10) Suppose n and r are positive integers. For a $n \times n$ matrix A, we can define A^r as the matrix obtained by multiplying A with itself repeatedly, with A appearing a total of r times.

(11) For a $n \times n$ matrix A, we define A^{-1} as the unique matrix such that $AA^{-1} = I_n$. It also follows that $A^{-1}A = I_n$.

(12) For a $n \times n$ invertible matrix A, we can define A^r for all integers r (positive, zero, or negative). A^0 is the identity matrix. $A^{-r} = (A^{-1})^r = (A^r)^{-1}$.

(13) Suppose A and B are matrices. The question of whether $AB = BA$ (i.e., of whether A and B commute) makes sense only if A and B are both square matrices of the same size, i.e., they are both $n \times n$ matrices for some n. However, $n \times n$ matrices need not always commute. An example of a situation where matrices commute is when both matrices are powers of a given matrix. Also, diagonal matrices commute with each other, and scalar matrices commute with all matrices.

(14) Consider a system of simultaneous linear equations with m variables and n equations. Let A be the coefficient matrix. Then, A is a $n \times m$ matrix. If $\vec{y}$ is the output (i.e., the augmenting column) we

can think of this as solving the vector equation $A\vec{x} = \vec{y}$ for $\vec{x}$. If $m = n$ and A is invertible, we can write this as $\vec{x} = A^{-1}\vec{y}$.

(15) There are a number of algebraic identities relating matrix multiplication, addtion, and inversion. These include distributivity (relating multiplication and addition) and the involutive nature or reversal law (namely, $(AB)^{-1} = B^{-1}A^{-1}$). See the "Algebraic rules governing matrix multiplication and inversion" section in the lecture notes for more information.

Computational techniques-related ...

(1) The arithmetic complexity of matrix addition for two $n \times m$ matrices is $\Theta(mn)$. More precisely, we need to do mn additions.
(2) Matrix addition can be completely parallelized, since all the entry computations are independent. With such parallelization, the arithmetic complexity becomes $\Theta(1)$.
(3) The arithmetic complexity for multiplying a generic $m \times n$ matrix and a generic $n \times p$ matrix (to output a $m \times p$ matrix) using naive matrix multiplication is $\Theta(mnp)$. Explicitly, the operation requires mnp multiplications and $m(n-1)p$ additions. More explicitly, computing each entry as a dot product requires n multiplications and $(n-1)$ additions, and there is a total of mp entries.
(4) Matrix multiplication can be massively but not completely parallelized. All the entries of the product matrix can be computed separately, already reducing the arithmetic complexity to $\Theta(n)$. However, we can parallelized further the computation of the dot product by parallelizing addition. This can bring the arithmetic complexity (in the sense of the depth of the computational tree) down to $\Theta(\log_2 n)$.
(5) We can compute powers of a matrix quickly by using repeated squaring. Using repeated squaring, computing A^r for a positive integer r requires $\Theta(\log_2 r)$ matrix multiplications. An explicit description of the minimum number of matrix multiplications needed relies on writing r in base 2 and counting the number of 1s that appear.
(6) To assess the invertibility and compute the inverse of a matrix, augment with the identity matrix, then row reduce the matrix to the identity matrix (note that if its rref is not the identity matrix, it is not invertible). Now, see what the augmented side has turned to. This takes time (in the arithmetic complexity sense) $\Theta(n^3)$ because that's the time taken by Gauss-Jordan elimination (about n^2 row operations and each row operation requires $O(n)$ arithmetic operations).
(7) We can think of pre-processing the row reduction for solving a system of simultaneous linear equations as being equivalent to computing the inverse matrix first.

Material that you can read in the lecture notes, but not covered in the summary.

(1) Real-world example(s) to illustrate matrix multiplication and its associativity (Sections 3.4 and 6.3).
(2) The idea of fast matrix multiplication (Section 4.2).
(3) One-sided invertibility (Section 8).
(4) Noncommuting matrix examples and finite state automata (Section 10).

2. Geometry of linear transformations

(1) There is a concept of *isomorphism* as something that preserves essential structure or feature, where the concept of isomorphism depends on what feature is being preserved.
(2) There is a concept of *automorphism* as an isomorphism from a structure to itself. We can think of automrohpisms of a structure as *symmetries* of that structure.
(3) Linear transformations have already been defined. An *affine linear transformation* is something that preserves lines and ratios of lengths within lines. Any affine linear transformation is of the form $\vec{x} \mapsto A\vec{x} + \vec{b}$. For the transformation to be linear, we need $\vec{b}$ to be the zero vector, i.e., the transformation must send the origin to the origin. If A is the identity matrix, then the affine linear transformation is termed a *translation*.
(4) A linear *isomorphism* is an invertible linear transformation. For a linear isomorphism to exist from $\mathbb{R}^m$ to $\mathbb{R}^n$, we must have $m = n$. An affine linear isomorphism is an invertible affine linear transformation.
(5) A linear automorphism is a linear isomorphism from $\mathbb{R}^n$ to itself. An affine linear automorphism is an affine linear isomorphism from $\mathbb{R}^n$ to itself.

(6) A self-isometry of $\mathbb{R}^n$ is an invertible function from $\mathbb{R}^n$ to itself that preserves Euclidean distance. Any self-isometry of $\mathbb{R}^n$ must be an affine linear automorphism of $\mathbb{R}^n$.

(7) A self-homothety of $\mathbb{R}^n$ is an invertible function from $\mathbb{R}^n$ to itself that scales all Euclidean distances by a factor of λ, where λ is the factor of homothety. We can think of self-isometries precisely as the self-homotheties by a factor of 1. Any self-homothety of $\mathbb{R}^n$ must be an affine linear automorphism of $\mathbb{R}^n$.

(8) Each of these forms a group: the affine linear automorphisms of $\mathbb{R}^n$, the linear automorphisms of $\mathbb{R}^n$, the self-isometries of $\mathbb{R}^n$, the self-homotheties of $\mathbb{R}^n$.

(9) For a linear transformation, we can consider something called the *determinant.* For a 2×2 linear transformation with matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the determinant is $ad - bc$.

We can also consider the *trace*, defined as $a + d$ (the sum of the diagonal entries).

(10) The trace generalizes to $n \times n$ matrices: it is the sum of the diagonal entries. The determinant also generalizes, but the formula becomes more complicated.

(11) The determinant for an affine linear automorphism can be defined as the determinant for its linear part (the matrix).

(12) The sign of the determinant being positive means the transformation is orientation-preserving. The sign of the determinant being negative means the transformation is orientation-reversing.

(13) The magnitude of the determinant gives the factor by which volumes are scaled. In the case $n = 2$, it is the factor by which areas are scaled.

(14) The determinant of a self-homothety with factor of homothety λ is $\pm\lambda^n$, with the sign depending on whether it is orientation-preserving or orientation-reversing.

(15) Any self-isometry is volume-preserving, so it has determinant ± 1, with the sign depending on whether it is orientation-preserving or orientation-reversing.

(16) For $n = 2$, the orientation-preserving self-isometries are precisely the translations and rotations. The ones fixing the origin are precisely rotations centered at the origin. These form groups.

(17) For $n = 2$, the orientation-reversing self-isometries are precisely the reflections and glide reflections. The ones fixing the origin are precisely reflections about lines passing through the origin.

(18) For $n = 3$, the orientation-preserving self-isometries fixing the origin are precisely the rotations about axes through the origin. The overall classification is more complicated.

3. Image and kernel of a linear transformation

(1) For a function $f : A \to B$, we call A the domain, B the co-domain, $f(A)$ the range, and $f^{-1}(b)$, for any $b \in B$, the fiber (or inverse image or pre-image) of b. For a subset S of B, $f^{-1}(B) = \bigcup_{b \in S} f^{-1}(b)$.

(2) The sizes of fibers can be used to characterize injectivity (each fiber has size at most one), surjectivity (each fiber is non-empty), and bijectivity (each fiber has size exactly one).

(3) Composition rules: composite of injective is injective, composite of surjective is surjective, composite of bijective is bijective.

(4) If $g \circ f$ is injective, then f must be injective.

(5) If $g \circ f$ is surjective, then g must be surjective.

(6) If $g \circ f$ is bijective, then f must be injective and g must be surjective.

(7) Finding the fibers for a function of one variable can be interpreted geometrically (intersect graph with a horizontal line) or algebraically (solve an equation).

(8) For continuous functions of one variable defined on all of $\mathbb{R}$, being injective is equivalent to being increasing throughout or decreasing throughout. More in the lecture notes, sections 2.2-2.5.

(9) A vector $\vec{v}$ is termed a *linear combination* of the vectors $\vec{v_1}, \vec{v_2}, \dots, \vec{v_r}$ if there exist real numbers $a_1, a_2, \dots, a_r \in \mathbb{R}$ such that $\vec{v} = a_1\vec{v_1} + a_2\vec{v_2} + \dots + a_r\vec{v_r}$. We use the term *nontrivial* if the coefficients are not all zero.

(10) A subspace of $\mathbb{R}^n$ is a subset that contains the zero vector and is closed under addition and scalar multiplication.

(11) The span of a set of vectors is defined as the set of all vectors that can be written as linear combinations of vectors in that set. The span of any set of vectors is a subspace.
(12) A spanning set for a subspace is defined as a subset of the subspace whose span is the subspace.
(13) Adding more vectors either preserves or increases the span. If the new vectors are in the span of the previous vectors, it preserves the span, otherwise, it increases it.
(14) The kernel and image (i.e., range) of a linear transformation are respectively subspaces of the domain and co-domain. The kernel is defined as the inverse image of the zero vector.
(15) The column vectors of the matrix of a linear transformation form a spanning set for the image of that linear transformation.
(16) To find a spanning set for the kernel, we convert to rref, then find the solutions parametrically (with zero as the augmenting column) then determine the vectors whose linear combinations are being discussed. The parameters serve as the coefficients for the linear combination. There is a shortening of this method. (See the lecture notes, Section 4.4, for a simple example done the long way and the short way).
(17) The fibers for a linear transformation are translates of the kernel. Explicitly, the inverse image of a vector is either empty or is of the form (particular vector) + (arbitrary element of the kernel).
(18) The dimension of a subspace of $\mathbb{R}^n$ is defined as the minimum possible size of a spanning set for that subspace.
(19) For a linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ with $n \times m$ matrix having rank r, the dimension of the kernel is $m - r$ and the dimension of the image is r. Full row rank $r = n$ means surjective (image is all of $\mathbb{R}^n$) and full column rank $r = m$ means injective (kernel is zero subspace).
(20) We can define the intersection and sum of subspaces of $\mathbb{R}^n$.
(21) The kernel of $T_1 + T_2$ contains the intersection of the kernels of T_1 and T_2. More is true (see the lecture notes).
(22) The image of $T_1 + T_2$ is contained in the sum of the images of T_1 and T_2. More is true (see the lecture notes).
(23) The dimension of the inverse image $T^{-1}(X)$ of any subspace X of $\mathbb{R}^n$ under a linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ satisfies:

$$\dim(\mathrm{Ker}(T)) \leq \dim(T^{-1}(X)) \leq \dim(\mathrm{Ker}(T)) + \dim(X)$$

The upper bound holds if X lies inside the image of T.

(24) Please read through the lecture notes thoroughly, since the summary here is very brief and inadequate.

REVIEW SHEET FOR MIDTERM 2: ADVANCED

MATH 196, SECTION 57 (VIPUL NAIK)

Please bring a copy (print or readable electronic) of this sheet to the review session.

There is also a basic review sheet that contains executive summaries of the lecture notes. You should review that on your own time.

I've kept the error-spotting exercises brief, because I intend to concentrate more on reviewing some of the techniques covered in the quizzes.

1. Matrix multiplication and inversion

Error-spotting exercises ...

(1) Suppose A and B are $n \times n$ matrices, with B invertible. Suppose r is a positive integer. Then, $(BAB^{-1})^r = B^r A^r (B^{-1})^r = B^r A^r B^{-r}$. Note that since A and B do not in general commute, we must write the terms in precisely this order.

(2) Suppose A and B are $n \times n$ matrices and r is a positive integer such that $(AB)^r = 0$. Then, we can conclude that $(BA)^r = 0$ as follows: we can write $(BA)^r = (BA)^r BB^{-1} = BABA \dots BABB^{-1} = B(AB)^r B^{-1} = B(0)B^{-1} = 0$.

(3) Suppose A and B are $n \times n$ matrices. Then, AB is nilpotent if and only if at least one of the matrices A or B is nilpotent. To see this, suppose r is a positive integer such that $(AB)^r = 0$. Then, we know that $(AB)^r = A^r B^r$, so $A^r B^r = 0$, forcing that either $A^r = 0$ or $B^r = 0$. The argument also works in reverse: if either of the matrices is nilpotent, there exists r such that one of the matrices A^r and B^r is 0. Thus, $A^r B^r = 0$, so $(AB)^r = 0$, so AB is nilpotent.

(4) Suppose A and B are invertible $n \times n$ matrices. Then, the sum $A + B$ is also an invertible $n \times n$ matrix, and $(A + B)^{-1} = A^{-1} + B^{-1}$.

(5) Suppose A and B are invertible $n \times n$ matrices. Then, the product AB is also an invertible $n \times n$ matrix, and $(AB)^{-1} = A^{-1}B^{-1}$.

(6) Suppose A and B are matrices with real entries, with A a single row matrix and B a single column matrix. Then, AB makes sense if and only if BA makes sense, and if so, we must have that $AB = BA$.

(7) Suppose A is a $m \times n$ matrix and B is a $n \times p$ matrix. The product $C = AB$ is a $m \times p$ matrix. For $1 \le i \le m$ and $1 \le k \le p$, the value c_{ik} is the product $a_{ij}b_{jk}$, with $1 \le j \le n$.

2. Geometry of linear transformations

Error-spotting exercises ... (this section is not too important, so we will probably do it last)

(1) Suppose D is a 2×2 diagonal matrix and T is the linear transformation corresponding to D. Let's say the two diagonal entries of D are a and d. The necessary and sufficient condition for T to be area-preserving is that the total effect on the x and y directions add up to 1 (the ratio of change of areas). Thus, T is area-preserving if and only if $a + d = 1$.

(2) The composite of the reflection maps about two lines through the origin that make an angle of θ with each other is the rotation map by the angle of θ. Moreover, the order of composition does not matter, i.e., the composite for both orders of composition is the same.

(3) The composite of two rotations in $\mathbb{R}^2$ is always a rotation, even if the centers of rotation differ.

(4) A bijective function $f : \mathbb{R}^n \to \mathbb{R}^n$ is an affine linear automorphism of $\mathbb{R}^n$ if and only if it sends lines to lines.

3. Image and kernel

3.1. Injectivity, surjectivity, and bijectivity.

Error-spotting exercises ...

(1) Suppose $f_1, f_2, f_3 : A \to A$ are set maps. Suppose the composite $f_1 \circ f_2 \circ f_3$ is bijective. Then, f_1 must be injective (because it's the one done first, so it cannot create any collision), f_2 must be bijective, and f_3 must be surjective (because it's the one done last, so it must hit everything).

(2) Suppose $f : \mathbb{R} \to \mathbb{R}$ is a polynomial of degree equal to the natural number $n \geq 3$. If n is even, f is surjective but not injective (e.g., $f(x) = x^4$). If n is odd, f is injective but not surjective (e.g., $f(x) = x^3$).

(3) Suppose f is a function from $\mathbb{R}$ to $\mathbb{R}$. Suppose that the restriction of f to $\mathbb{Z}$ maps $\mathbb{Z}$ to inside $\mathbb{Z}$ (i.e., f takes integer values at integer inputs). Let $g : \mathbb{Z} \to \mathbb{Z}$ be the function obtained by restricting f to $\mathbb{Z}$. Then:
  (a) f is injective if and only if g is injective.
  (b) f is surjective if and only if g is surjective.
  (c) f is bijective if and only if g is bijective.

3.2. **Linear transformation and rank.** Error-spotting exercises ...

(1) If T_1 and T_2 are linear transformations from $\mathbb{R}^2$ to $\mathbb{R}^2$, then the kernel of $T_1 + T_2$ equals the intersection of the kernels of T_1 and T_2. Here's a proof. Suppose a vector $\vec{u}$ is in the kernel of T_1 as well as the kernel of T_2. Then, $T_1(\vec{u}) = T_2(\vec{u}) = 0$. Thus, $(T_1 + T_2)(\vec{u}) = T_1(\vec{u}) + T_2(\vec{u}) = 0 + 0 = 0$.

In particular, this means that if both T_1 and T_2 are invertible, then $T_1 + T_2$ is invertible.

(2) Suppose $T_1 : \mathbb{R}^a \to \mathbb{R}^b$ and $T_2 : \mathbb{R}^b \to \mathbb{R}^c$ are linear transformations. Then, the composite $T_1 \circ T_2$ is a linear transformation from $\mathbb{R}^a$ to $\mathbb{R}^c$. In terms of matrices, the matrix for T_1 is an $a \times b$ matrix and the matrix for T_2 is a $b \times c$ matrix. So the matrix for $T_1 \circ T_2$ is an $a \times c$ matrix, and is given by the matrix product of those two matrices.

Suppose the kernel of T_1 has dimension m and the kernel of T_2 has dimension n. A vector is in the kernel of $T_1 \circ T_2$ if and only if it is in the kernel of either T_1 or T_2. Thus, the dimension of the kernel of $T_1 \circ T_2$ is the maximum of the dimensions of the kernels of T_1 and T_2, which is $\max\{m, n\}$.

(3) Suppose $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear transformation with matrix A. Then A is a $m \times n$ matrix and the following are true:
  (a) The rows of A form a spanning set for the image of T.
  (b) The columns of A form a spanning set for the kernel of T.

(4) Consider the linear transformation:

$$\nu = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \mapsto \begin{bmatrix} (y+z)/2 \\ (z+x)/2 \\ (x+y)/2 \end{bmatrix}$$

The kernel of T is precisely those vectors where $x = -y = z$, i.e., each coordinate is the negative of the next one. The image of T is the set where $x = y = z$.

3.3. **The linear operation of differentiation.** Error-spotting exercises ...

We will get to this only if we have enough time, since we will not see these topics outside the MCQs on the test.

(1) Denote by $C(\mathbb{R})$ the vector space of all continuous functions with the usual addition and scalar multiplication of functions. Denote by $C^1(\mathbb{R})$ the vector space of all continuously differentiable functions with the usual addition and scalar multiplication of functions. Differentiation defines a linear transformation from $C(\mathbb{R})$ to $C^1(\mathbb{R})$. The image of this linear transformation is precisely the set of constant functions.

(2) For every positive integer k, denote by $C^k(\mathbb{R})$ the subspace of $C(\mathbb{R})$ comprising those polynomials that are at least k times continuously differentiable. Then, $C^k(\mathbb{R}) \subseteq C^{k+1}(\mathbb{R})$ and the union of all the spaces $C^k(\mathbb{R})$ for k varying over the positive integers is the space $C^\infty(\mathbb{R})$ of infinitely differentiable functions.

(3) The set of polynomials of degree at most k form a vector subspace of $C^k(\mathbb{R})$ but not of $C^{k+1}(\mathbb{R})$.

REVIEW SHEET FOR FINAL: BASIC

MATH 196, SECTION 57 (VIPUL NAIK)

We will not be going over this sheet, but rather, we'll be going over the advanced review sheet in the session. Please review this sheet on your own time.

The summaries here are identical with the executive summaries you will find at the beginning of the respective lecture notes PDF files. The summaries are not intended to be exhaustive. Please review the original lecture notes as well, especially if any point in the summary is unclear.

1. Linear dependence, bases, and subspaces

(1) A *linear relation* between a set of vectors is defined as a linear combination of these vectors that is zero. The *trivial* linear relation refers to the trivial linear combination being zero. A nontrivial linear relation is any linear relation other than the trivial one.

(2) The trivial linear relation exists between any set of vectors.

(3) A set of vectors is termed *linearly dependent* if there exists a nontrivial linear relation between them, and *linearly independent* otherwise.

(4) Any set of vectors containing a linearly dependent subset is also linearly dependent. Any subset of a linearly independent set of vectors is a linearly independent set of vectors.

(5) The following can be said of sets of small size:
- The empty set (the only possible set of size zero) is considered linearly independent.
- A set of size one is linearly dependent if the vector is the zero vector, and linearly independent if the vector is a nonzero vector.
- A set of size two is linearly dependent if either one of the vectors is the zero vector or the two vectors are scalar multiples of each other. It is linearly independent if both vectors are nonzero and they are not scalar multiples of one another.
- For sets of size three or more, a *necessary* condition for linear independence is that no vector be the zero vector and no two vectors be scalar multiplies of each other. However, this condition is not sufficient, because we also have to be on the lookout for other kinds of linear relations.

(6) Given a nontrivial linear relation between a set of vectors, we can use the linear relation to write one of the vectors (any vector with a nonzero coefficient in the linear relation) as a linear combination of the other vectors.

(7) We can use the above to prune a spanning set as follows: given a set of vectors, if there exists a nontrivial linear relation between the vectors, we can use that to write one vector as a linear combination of the others, and then remove it from the set *without affecting the span*. The vector thus removed is termed a *redundant vector*.

(8) A *basis* for a subspace of $\mathbb{R}^n$ is a linearly independent spanning set for that subspace. Any finite spanning set can be pruned down (by repeatedly identifying linear relations and removing vectors) to reach a basis.

(9) The size of a basis for a subspace of $\mathbb{R}^n$ depends only on the choice of subspace and is *independent* of the choice of basis. This size is termed the *dimension* of the subspace.

(10) Given an ordered list of vectors, we call a vector in the list *redundant* if it is redundant relative to the preceding vectors, i.e., if it is in the span of the preceding vectors, and *irredundant* otherwise. The irredundant vectors in any given list of vectors form a basis for the subspace spanned by those vectors.

(11) Which vectors we identify as redundant and irredundant depends on how the original list was ordered. However, the *number* of irredundant vectors, insofar as it equals the dimension of the span, does not depend on the ordering.

(12) If we write a matrix whose column vectors are a given list of vectors, the linear relations between the vectors correspond to vectors in the kernel of the matrix. Injectivity of the linear transformation given by the matrix is equivalent to linear independence of the vectors.

(13) Redundant vector columns correspond to non-leading variables and irredundant vector columns correspond to leading variables if we think of the matrix as a coefficient matrix. We can row-reduce to find which variables are leading and non-leading, then look at the irredundant vector columns in the *original* matrix.

(14) *Rank-nullity theorem*: The nullity of a linear transformation is defined as the dimension of the kernel. The nullity is the number of non-leading variables. The rank is the number of leading variables. So, the sum of the rank and the nullity is the number of columns in the matrix for the linear transformation, aka the dimension of the domain. See Section 3.7 of the notes for more details.

(15) The problem of finding all the vectors orthogonal to a given set of vectors can be converted to solving a linear system where the rows of the coefficient matrix are the given vectors.

2. Coordinates (includes discussion of similarity of linear transformations)

(1) Given a basis $\mathcal{B} = (\vec{v}_1, \vec{v}_2, \dots, \vec{v}_m)$ for a subspace $V \subseteq \mathbb{R}^n$ (note that this forces $m \le n$), every vector $\vec{x} \in V$ can be written in a unique manner as a linear combination of the basis vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_m$. The fact that there exists a way of writing it as a linear combination follows from the fact that $\mathcal{B}$ spans V. The uniqueness follows from the fact that $\mathcal{B}$ is linearly independent. The coefficients for the linear combination are called the *coordinates* of $\vec{x}$ in the basis $\mathcal{B}$.

(2) Continuing notation from point (1), finding the coordinates amounts to solving the linear system with coefficient matrix columns given by the basis vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_m$ and the augmenting column given by the vector $\vec{x}$. The linear transformation of the matrix is injective, because the vectors are linearly independent. The matrix, a $n \times m$ matrix, has full column rank m. The system is consistent if and only if $\vec{x}$ is actually in the span, and injectivity gives us uniqueness of the coordinates.

(3) A canonical example of a basis is the *standard* basis, which is the basis comprising the standard basis vectors, and where the coordinates are the usual coordinates.

(4) Continuing notation from point(1), in the special case that $m = n$, $V = \mathbb{R}^n$. So the basis is $\mathcal{B} = (\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n)$ and it is an alternative basis for all of $\mathbb{R}^n$ (here, alternative is being used to contrast with the standard basis; we will also use "old basis" to refer to the standard basis and "new basis" to refer to the alternative basis). In this case, the matrix S whose columns are the basis vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ is a $n \times n$ square matrix and is invertible. We will denote this matrix by S (following the book).

(5) Continuing notation from point (4), if we denote by $[\vec{x}]_{\mathcal{B}}$ the coordinates of $\vec{x}$ in the new basis, then $[\vec{x}]_{\mathcal{B}} = S^{-1}\vec{x}$ and $\vec{x} = S[\vec{x}]_{\mathcal{B}}$.

(6) For a linear transformation T with matrix A in the standard basis and matrix B in the new basis, then $B = S^{-1}AS$ or equivalently $A = SBS^{-1}$. The S on the right involves first converting from the new basis to the old basis, then we do the middle operation A on the old basis, and then we do S^{-1} to re-convert to the new basis.

(7) If A and B are $n \times n$ matrices such that there exists an invertible $n \times n$ matrix S satisfying $B = S^{-1}AS$, we say that A and B are *similar* matrices. Similar matrices have the same trace, determinant, and behavior with respect to invertibility and nilpotency. Similarity is an equivalence relation, i.e., it is reflexive, symmetric, and transitive.

(8) Suppose S is an invertible $n \times n$ matrix. The conjugation operation $X \mapsto SXS^{-1}$ from $\mathbb{R}^{n \times n}$ to $\mathbb{R}^{n \times n}$ preserves addition, scalar multiplication, multiplication, and inverses.

3. Abstract vector spaces and the concept of isomorphism

General stuff ...

(1) There is an abstract definition of real vector space that involves a set with a binary operation playing the role of addition and another operation playing the role of scalar multiplication, satisfying a bunch of axioms. The goal is to axiomatize the key aspects of vector spaces.

(2) A subspace of an abstract vector space is a subset that contains the zero vector and is closed under addition and scalar multiplication.

(3) A linear transformation is a set map between two vector spaces that preserves addition and preserves scalar multiplication. It also sends zero to zero, but this follows from its preserving scalar multiplication.

(4) The *kernel* of a linear transformation is the subset of the domain comprising the vectors that map to zero. The kernel of a linear transformation is always a subspace.

(5) The *image* of a linear transformation is its range as a set map. The image is a subspace of the co-domain.

(6) The *dimension* of a vector space is defined as the size of any basis for it. The dimension provides an upper bound on the size of any linearly independent set in the vector space, with the upper bound attained (in the finite case) only if the linearly independent set is a basis. The dimension also provides a lower bound on the size of any spanning subset of the vector space, with the lower bound being attained (in the finite case) only if the spanning set is a basis.

(7) Every vector space has a particular subspace of interest: the zero subspace.

(8) The *rank* of a linear transformation is defined as the dimension of the image. The rank is the answer to the question: "how much survives the linear transformation?"

(9) The *nullity* of a linear transformation is defined as the dimension of the kernel. The nullity is the answer to the question: "how much gets killed under the linear transformation?"

(10) The sum of the rank and the nullity of a linear transformation equals the dimension of the domain. This fact is termed the *rank-nullity theorem*.

(11) We can define the *intersection* and *sum* of subspaces. These are again subspaces. The intersection of two subspaces is defined as the set of vectors that are present in both subspaces. The sum of two subspaces is defined as the set of vectors expressible as a sum of vectors, one in each subspace. The sum of two subspaces also equals the subspace spanned by their union.

(12) A linear transformation is *injective* if and only if its kernel is the zero subspace of the domain.

(13) A linear transformation is *surjective* if and only if its image is the whole co-domain.

(14) A *linear isomorphism* is a linear transformation that is *bijective*: it is both injective and surjective. In other words, its kernel is the zero subspace of the domain and its image is the whole co-domain.

(15) The dimension is an isomorphism-invariant. It is in fact a *complete isomorphism-invariant*: two real vector spaces are isomorphic if and only if they have the same dimension. Explicitly, we can use a bijection between a basis for one space and a basis for another. In particular, any n-dimensional space is isomorphic to $\mathbb{R}^n$. Thus, by studying the vector spaces $\mathbb{R}^n$, we have effectively studied all finite-dimensional vector spaces up to isomorphism.

Function spaces ...

(1) For any set S, consider the set $F(S, \mathbb{R})$ of *all* functions from S to $\mathbb{R}$. With pointwise addition and scalar multiplication of functions, this set is a vector space over $\mathbb{R}$. If S is finite (*not* our main case of interest) this space has dimension $|S|$ and is indexed by a basis of S. We are usually interested in *subspaces* of this space.

(2) We can define vector spaces such as $\mathbb{R}[x]$ (the vector space of all polynomials in one variable with real coefficients) and $\mathbb{R}(x)$ (the vector space of all rational functions in one variable with real coefficients). These are both infinite-dimensional spaces. We can study various finite-dimensional subspaces of these. For instance, we can define P_n as the vector space of all polynomials of degree less than or equal to n. This is a vector space of dimension $n+1$ with basis given by the monomials $1, x, x^2, \dots, x^n$.

(3) There is a natural injective linear transformation $\mathbb{R}[x] \to F(\mathbb{R}, \mathbb{R})$.

(4) Denote by $C(\mathbb{R})$ or $C^0(\mathbb{R})$ the subspace of $F(\mathbb{R}, \mathbb{R})$ comprising the functions that are continuous everywhere. For k a positive integer, denote by $C^k(\mathbb{R})$ the subspace of $C(\mathbb{R})$ comrpising those functions that are at least k times continuously differentiable, and denote by $C^\infty(\mathbb{R})$ the subspace of $C(\mathbb{R})$ comprising all the functions that are *infinitely* differentiable. We have a descending chain of subspaces:

$$C^0(\mathbb{R}) \supseteq C^1(\mathbb{R}) \supseteq C^2(\mathbb{R}) \supseteq \dots$$

The image of $\mathbb{R}[x]$ inside $F(\mathbb{R}, \mathbb{R})$ lands inside $C^\infty(\mathbb{R})$.

(5) We can view differentiation as a linear transformation $C^1(\mathbb{R}) \to C(\mathbb{R})$. It sends each $C^k(\mathbb{R})$ to $C^{k-1}(\mathbb{R})$. It is surjective from $C^\infty(\mathbb{R})$ to $C^\infty(\mathbb{R})$. The kernel is constant functions, and the kernel of k-fold iteration is P_{k-1}. Differentiation sends $\mathbb{R}[x]$ to $\mathbb{R}[x]$ and is surjective to $\mathbb{R}[x]$.

(6) We can also define a formal differentiation operator $\mathbb{R}(x) \to \mathbb{R}(x)$. This is not surjective.

(7) Partial fractions theory can be formulated in terms of saying that some particular rational functions form a basis for certain finite-dimensional subspaces of the space of rational functions, and exhibiting a method to find the "coordinates" of a rational function in terms of this basis. The advantage of expressing in this basis is that the basis functions are particularly easy to integrate.

(8) We can define a vector space of sequences. This is a special type of function space where the domain is $\mathbb{N}$. In other words, it is the function space $F(\mathbb{N}, \mathbb{R})$.

(9) We can define a vector space of formal power series. The Taylor series operator and series summation operator are back-and-forth operators between this vector space (or an appropriate subspace therefore) and $C^\infty(\mathbb{R})$.

(10) Formal differentiation is a linear transformation $\mathbb{R}[[x]] \to \mathbb{R}[[x]]$. It is surjective but not injective. The kernel is the one-dimensional space of formal power series.

(11) We can consider linear differential operators from $C^\infty(\mathbb{R})$ to $C^\infty(\mathbb{R})$. These are obtained by combining the usual differentiation operator and multiplication operators using addition, multiplication (composition) and scalar multiplication. Finding the kernel of a linear differential operator is equivalent to solving a homogeneous linear differential equation. Finding the inverse image of a particular function under a linear differential operator amounts to solving a non-homogeneous linear differential equation, and the solution set here is a translate of the kernel (the corresponding solution in the homogeneous case, also called the *auxilliary solution*) by a particular solution. The first-order case is particularly illuminative because we have an explicit formula for the fibers.

4. Ordinary least squares regression

Words ...

(1) Consider a model where the general functional form is linear in the parameters. Input-output pairs give a system of simultaneous linear equations in terms of the parameters. Each row of the coefficient matrix corresponds to a particular choice of input, and each corresponding entry of the augmenting column is the corresponding output. In the "no-error" case, what we would ideally like is that the coefficient matrix has full column rank (i.e., we get unique solutions for the parameters assuming consistency) and does *not* have full row rank (i.e., we have some extra input-output pairs so that consistency can be used as evidence in favor of the hypothesis that the given functional form is correct). If the model is correct, the system will be consistent (despite potential for inconsistency) and we will be able to deduce the values of the parameters.

(2) Once we introduce measurement error, we can no longer find the parameters with certainty. However, what we *can* hope for is to provide a "best guess" for the parameter values based on the given data points (input-output pairs).

(3) In the case where we have measurement error, we still aim to choose a large number of inputs so that the coefficient matrix has full column rank but does not have full row rank. Now, however, even if the model is correct, the system will probably be inconsistent. What we need to do is to replace the existin output vector (i.e., the existing augmenting column) by the vector closest to it that is in the image of the linear transformation given by the coefficient matrix. Explicitly, if $\vec{\beta}$ is the parameter vector that we are trying to find, X is the coefficient matrix (also called the design matrix), and $\vec{y}$ is the output vector, the system $X\vec{\beta} = \vec{y}$ may not be consistent. We try to find a vector $\vec{\varepsilon}$ of minimum length subject to the constraint that $\vec{y} - \vec{\varepsilon}$ is in the image of the linear transformation given by X, so that the system $X\vec{\beta} = \vec{y} - \vec{\varepsilon}$ is consistent. Because in our setup (if we chose it well), X had full column rank, this gives the unique "best"choice of $\vec{\beta}$. Note also that "length" here refers to Euclidean length (square root of sum of squares of coordinates) when we are doing an *ordinary least squares regression* (the default type of regression) but we could use alternative notions of length in other types of regressions.

(4) In the particular case that the system $X\vec{\beta} = \vec{y}$ is consistent, we choose $\vec{\varepsilon} = \vec{0}$. However, this does not mean that ths is the actual correct parameter vector. It is still only a guess.

(5) In general, the more data points (i.e., input-output pairs) that we have, the better our guess becomes. However, this is true only in a probabilistic sense. It may well happen that a particular data point worsens our guess because that data point has a larger error than the others.

(6) The corresponding geometric operation to finding the vector $\vec{\varepsilon}$ is orthogonal projection. Explicitly, the image of X is a subspace of the vector space $\mathbb{R}^n$ (where n is the number of input-output pairs). If there are m parameters (i.e., $\vec{\beta} \in \mathbb{R}^m$). and we chose X wisely, the image of X would be a m-dimensional subspace of $\mathbb{R}^n$. In the no-error case, the vector $\vec{y}$ would be in this subspace, and we would be able to find $\vec{\beta}$ uniquely and correctly. In the presence of error, $\vec{y}$ may be outside the subspace. The vector $\vec{y} - \vec{\varepsilon}$ that we are looking for is the orthogonal projection of $\vec{y}$ onto this subspace. The error vector $\vec{\varepsilon}$ is the component of $\vec{y}$ that is perpendicular to the subspace.

(7) A fit is impressive in the case that m is much smaller than n and yet the error vector $\vec{\varepsilon}$ is small. This is philosophically for the same reason that consistency becomes more impressive the greater the excess of input-output pairs over parameters. Now, the rigid notion of consistency has been replaced by the more loose notion of "small error vector."

Actions ...

(1) Solving $X\vec{\beta} = \vec{y} - \vec{\varepsilon}$ with $\vec{\varepsilon}$ (unknown) as the vector of minimum possible length is equivalent to solving $X^TX\vec{\beta} = X^T\vec{y}$. Note that this process does not involve finding the error vector $\vec{\varepsilon}$ directly. The matrix X^TX is a square matrix that is symmetric.

(2) In the case that X has full column rank (i.e., we have unique solutions *if* consistent), X^TX also has full rank (both full row rank and full column rank – it is a square matrix), and we get a unique "best fit" solution.

REVIEW SHEET FOR FINAL: ADVANCED

MATH 196, SECTION 57 (VIPUL NAIK)

To maximize efficiency, please bring a copy (print or readable electronic) of this review sheet to the Saturday review sesssion.

Please come to the session *only if* you know the meanings of these:

- Subspace
- Linear transformation
- Linear combination
- Linear relation
- Span
- Spanning set
- Linear dependence
- Linear independence
- Basis
- Dimension
- Kernel
- Image
- Rank
- Nullity
- Injectivity
- Surjectivity
- Bijectivity
- Transpose of a matrix

Please go through the basic review sheet, book, or lecture notes, if any of these terms trip you up.

1. Linear dependence, bases and subspaces

Error-spotting exercises ...

(1) *Half-truth*: Consider $\mathbb{R}$ as a vector space. Then, $\mathbb{Z}$, the set of integers, is a subspace of $\mathbb{R}$ because it is closed under addition and contains the zero vector.

(2) *Something doesn't add up*: Consider $\mathbb{R}^2$ as a vector space. Then, the set comprising the vectors $\vec{e}_1, \vec{e}_2, \vec{e}_1 + \vec{e}_2$ and their scalar multiples is a subspace because it contains the zero vector, is closed under addition (after all, $\vec{e}_1 + \vec{e}_2 = \vec{e}_1 + \vec{e}_2$) and is closed under scalar multiplication (by assumption).

(3) *Too non-slanted, too uncrooked*: Suppose U is a vector subspace of $\mathbb{R}^n$. Then, we can obtain a basis for U as follows: start with the standard basis for $\mathbb{R}^n$. Now, pick the vectors from this that are also inside U. These form a basis for U. For instance, if U is the span of $\vec{e}_2$ and $\vec{e}_3$ in $\mathbb{R}^4$, this procedure works.

(4) *Telepathic basis*: Suppose U_1 and U_2 are vector subspaces of $\mathbb{R}^n$. Suppose we are given a basis S_1 for U_1 and a basis S_2 for U_2. Then, $S_1 \cap S_2$ is a basis for the vector space $U_1 \cap U_2$ and $S_1 \cup S_2$ is a basis for the vector space $U_1 \cup U_2$.

(5) *Too trivial to be true*: If a collection of vectors in $\mathbb{R}^n$ satisfies the trivial linear relation, it is termed linearly independent. If, however, it does not satisfy the trivial linear relation, it is termed linearly dependent.

(6) *Throwing out the baby with the bathwater*: Suppose S is a set of vectors in $\mathbb{R}^n$ that spans a subspace V of $\mathbb{R}^n$, and there is a nontrivial linear relation between the vectors of S that uses four of the vectors in S nontrivially (i.e., it has nonzero coefficients for four of the vectors in S). This means

that we can throw out any of those four vectors and still span V. Thus, we can reduce the size of S by 4 and still get a spanning set for V.

2. Coordinates (in focus: similarity of linear transformations)

Error-spotting exercises ...

(1) *Same similar*: Suppose A_1 and B_1 are similar $n \times n$ matrices, and A_2 and B_2 are also similar $n \times n$ matrices. Then, A_1A_2 and B_1B_2 are similar $n \times n$ matrices. Here is the proof: since A_1 and B_1 are similar, there exists a matrix S such that $A_1 = SB_1S^{-1}$. Since A_2 and B_2 are similar, there exists a matrix S such that $A_2 = SB_2S^{-1}$. Then, $A_1A_2 = (SB_1S^{-1})(SB_2S^{-1}) = S(B_1B_2)S^{-1}$. Thus, A_1A_2 and B_1B_2 are similar.

(2) *Shallow roots*: Suppose A and B are $n \times n$ matrices and r is a positive integer. Is it the case that A^r being similar to B^r implies that A is similar to B? Well, this depends on whether r is odd or even. If r is even, then A and B need not be similar. We can get counterexamples even using 1×1 matrices: $[1]$ and $[-1]$ have the same square, but are different.

On the other hand, if r is odd, then A^r and B^r being similar implies that A and B are similar. Here is the proof. If A^r and B^r are similar, this implies that there exists an invertible $n \times n$ matrix S such that $A^r = SB^rS^{-1} = (SBS^{-1})^r$. So, $A^r = (SBS^{-1})^r$. Since r is odd, we can cancel it from the exponent (note that we do not have the ± issue that we have with even r) and we get that $A = SBS^{-1}$, so that A and B are similar.

(3) *One-sided scaling*: Any two scalar matrices are similar because they represent the same linear transformation viewed at different scalings.

(4) *One-sided relabeling*: Interchanging the roles of the standard basis vectors $\vec{e}_1$ and $\vec{e}_2$ shows that the matrices:

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$$

are similar to one another.

3. Abstract vector spaces (in focus: function spaces)

Error-spotting exercises ...

(1) *Too big to compare*: Consider differentiation. We can think of this as a linear transformation from $C^1(\mathbb{R})$ (the vector space of all continously differentiable functions on all of $\mathbb{R}$) to $C(\mathbb{R})$ (the vector space of all continuous functions on $\mathbb{R}$). Here is an explanation for why the map is surjective: we know that the kernel of this linear transformation is the vector space of constant functions, which is 1-dimensional. By the rank-nullity theorem, we know that (rank) + (nullity) = dimension of domain. Since the nullity is 1, and the domain is infinite-dimensional, the rank is infinite. This equals the dimension of the co-domain. Since the rank equals the dimension of the co-domain, that means that the image is the whole co-domain, so the differentiation linear transformation is surjective.

(2) *Answer not in the answer key*: We can view differentiation as a linear transformation from $\mathbb{R}(x)$, the vector space of all rational functions in one variable, to $\mathbb{R}(x)$. This linear transformation is not injective, because its kernel is the space of constant functions, which is one-dimensional. The linear transformation is surjective, because we know how to integrate any rational function.

(3) (Can't think of a witty name): Consider the vector space of all rational functions that can be written in the form $r(x)/p(x)$ where p is a fixed polynomial of degree n. This vector space is n-dimensional. A basis for this is given by the rational functions of the form $1/(x - \alpha)$ for all roots α of p, as well as rational functions of the form $1/q(x)$ for all irreducible quadratic factors of p.

(4) *Off by one errors*: Consider differentiation as a linear transformation from P_n to P_n, where P_n is the vector space of all polynomials of degree less than or equal to n. P_n is a n-dimensional space. The linear transformation we obtain is bijective from P_n to P_n. This is because the derivative of any such polynomial is such a polynomial, and every such polynomial is the derivative of a polynomial. Explicitly, if we write a matrix for the linear transformation of differentiation, the matrix is a square matrix and has full rank n.

4. Ordinary least squares regression

Error-spotting exercises ...

(1) *Explaining the past versus predicting the future, aka it's hard to make predictions, especially about the future*: Suppose we are trying to fit a linear function of one variable using some data points (input-output pairs). If we use only two data points, then we can get a unique line through it, with an error vector of zero. In other words, we get a *perfect fit* with zero error.

Suppose that instead we use three data points. Due to measurement error, it is likely that there will be no line that perfectly fits all the three data points. We can still get a fit that minimizes error. However, notice that the magnitude of the error vector is now bigger: the error vector was earlier a zero vector, but now it is (probably) a nonzero vector.

A similar argument can be used to show that *the more data points we have, the larger the magnitude of the error vector for our best fit.* In fact, this is true even when we make an adjustment for the number of coordinates (the error vector with four data points will not just have bigger length in expectation than the error vector with three data points, but the ratio of lengths will be expected to be more than $\sqrt{4}/\sqrt{3}$, i.e., *each coordinate* of the error vector is getting bigger in expectation.

So, this means that, for a given functional form, the greater the number of data points we decide to use, the worse the fit we will obtain. Therefore, to obtain a good fit, we should choose as few data points as possible, though still enough to uniquely determine the function. In the linear case, this ideal number is 2. Less is too little. More is too confusing, because of the possible inconsistencies that arise.

(2) *Off by one errors*: Consider trying to fit a function of one variable, with n data points (i.e., n input-output pairs) where we attempt to fit it using a polynomial of degree (at most) m. Then, the design matrix for the regression (i.e., the coefficient matrix of the linear system) is a $m \times n$ matrix, because there are m parameters and n input-output pairs.

(3) *The best route to success is to avoid listening to negative feedback*: When choosing the design matrix of a linear regression, i.e., choosing the inputs of the input-output pairs, we should attempt to make the design matrix a square matrix of full rank. This is because we want full column rank in order to uniquely determine the parameters, and we need full row rank in order to make sure that a solution *exists*.

(4) *Portrait versus landscape*: Suppose we are trying to find the parameter vector $\vec{\beta}$ given the design matrix X. In other words, we are trying to solve the equation below, with $\vec{\varepsilon}$ chosen to be the vector of minimum length for which the system is consistent:

$$X\vec{\beta} = \vec{y} - \vec{\varepsilon}$$

We know that the vector $\vec{\varepsilon}$ is orthogonal to the image of X. Therefore, it is orthogonal to all the rows of the matrix X. In other words, $X\vec{\varepsilon} = \vec{0}$.

Thus, if we multiply both equations on the left by X, we obtain:

$$X^2\vec{\beta} = X\vec{y}$$

We can solve this system to find the best fit parameter vector $\vec{\beta}$.

5. Extra topic covered in the quizzes: Linear dynamical systems

Error-spotting exercises ...

(1) *Get unreal!*: Suppose A is a $n \times n$ matrix and $\vec{x}$ is a nonzero vector in $\mathbb{R}^n$. Suppose there exists a positive integer r such that $A^r\vec{x}$ is the zero vector in $\mathbb{R}^n$. Since $\vec{x}$ is a *non*zero vector, this forces A^r to be the zero matrix. Hence, A is nilpotent.

Conversely, if A is nilpotent, then $A^r = 0$. Thus, there exists a nonzero vector $\vec{x}$ such that $A^r\vec{x}$ is the zero vector.

The upshot: a $n \times n$ matrix A is nilpotent if and only if there exists a nonzero vector $\vec{x} \in \mathbb{R}^n$ and a positive integer r such that $A^r\vec{x}$ is the zero vector.

(2) *Just because you can return doesn't mean you will*: Suppose A is a $n \times n$ matrix and $\vec{x}$ is a nonzero vector in $\mathbb{R}^n$. Suppose there exists a positive integer r such that $A^r\vec{x} = \vec{x}$. Since A^r sends a nonzero vector to itself, it must be the identity matrix. Thus, $A^r = I_n$. So, $A(A^{r-1}) = (A^{r-1})A = I_n$. Thus, A^{r-1} equals A^{-1}, so in particular, A is invertible.

Conversely, consider the case that A is an invertible $n \times n$ matrix. This means that we can recover the vector $\vec{x}$ from knowledge of the vector $A\vec{x}$. This means that if we apply A enough times to $A\vec{x}$, we get $\vec{x}$. So, there exists s such that $A^s(A\vec{x}) = A^{s+1}(\vec{x}) = \vec{x}$. Set $r = s + 1$, and we have that $A^r\vec{x} = \vec{x}$.

The upshot: a $n \times n$ matrix A is invertible if and only if it has the property that there exists a nonzero vector $\vec{x} \in \mathbb{R}^n$ and a positive integer r such that $A^r\vec{x} = \vec{x}$.

(3) *You can't see all the wonders of the world in a short life*: Let T be the rotation about the origin in $\mathbb{R}^2$ by a fixed angle θ. Starting with any nonzero vector $\vec{x}$, consider the sequence:

$$\vec{x}, T(\vec{x}), T(T(\vec{x})), \ldots$$

When we rotate a vector, we preserve its length. Thus, the range of this sequence is the circle centered at the origin of radius equal to the length of $\vec{x}$.

For the following error-spotting exercises, use this (error-free) definition: Given a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$, a (possibly zero, possibly nonzero) real number λ, and a nonzero vector $\vec{x} \in \mathbb{R}^n$, we say that $\vec{x}$ is an eigenvector of T with eigenvalue λ if $T(\vec{x}) = \lambda\vec{x}$.

(4) *What we can't achieve alone, we can do together*: Consider the case $n = 2$ and define T to be the linear transformation with matrix:

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Note that T sends $\vec{e}_1$ to $\vec{e}_2$ and sends $\vec{e}_2$ to $\vec{e}_1$. In other words, T interchanges the two standard basis vectors. Since T does not preserve the lines of any of the standard basis vectors, neither of them is an eigenvector for T. Note that any vector would be a linear combination of the standard basis vectors, so T has no eigenvector.

(5) *Compatibility issues*: For any linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$, the set of eigenvectors of T, along with the zero vector, form a subspace of $\mathbb{R}^n$. Here's the proof. Note that:

- The zero vector is in the set by definition (although the zero vector is not considered an eigenvector, our definition here deliberately adds the zero vector in).
- Suppose vectors $\vec{u}$ and $\vec{v}$ are both eigenvectors for T. This means that there exists a real number λ such that $T(\vec{u}) = \lambda\vec{u}$ and $T(\vec{v}) = \lambda\vec{v}$, then $T(\vec{u} + \vec{v}) = \lambda\vec{u} + \lambda\vec{v}$ which becomes $\lambda(\vec{u} + \vec{v})$.
- Suppose $\vec{v}$ is an eigenvector for T with eigenvalue λ. Then, for any real number a, $T(a\vec{v}) = aT(\vec{v}) = a(\lambda\vec{v}) = \lambda(a\vec{v})$, so $a\vec{v}$ is also an eigenvector for T. Moreover, it has the same eigenvalue.

(6) *Don't be Procrustean!*: Consider the linear transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ with matrix:

$$\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$

The eigenvectors for this are $\vec{e}_1$ (with eigenvalue 1) and $\vec{e}_2$ (with eigenvalue 2). Note that there are no more eigenvectors. For instance, $\vec{e}_1 + \vec{e}_2$ is not an eigenvector because its image is $\vec{e}_1 + 2\vec{e}_2$, which is not a multiple of it.

(7) *A missed match*: Consider the linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ with matrix:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

The matrix is diagonal, so the eigenvectors for this linear transformation are precisely the standard basis vectors $\vec{e}_1$, $\vec{e}_2$, $\vec{e}_3$.

(8) *Zero's legit*: Consider the linear transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$ with matrix:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

This sends $\vec{e}_3$ to $\vec{e}_2$, sends $\vec{e}_2$ to $\vec{e}_1$, and sends $\vec{e}_1$ to the vector zero. Note that none of the standard basis vectors goes to itself, or for that matter, to a multiple of itself. In other words, T has no eigenvectors.