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 $X^T X\vec{\beta} = X^T\vec{y}$. Note that this process does not involve finding the error vector $\vec{\varepsilon}$ directly. The matrix $X^T X$ 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), $X^T X$ 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.