Exponential Families on Abstract Affine Spaces

Charles J. Geyer

May 13, 2024

> The first advice the author [Michel Talagrand] received from his advisor Gustave Choquet was as follows: Always consider a problem under the minimum structure in which it makes sense. This advice will probably be as fruitful in the future as it has been in the past, and it has strongly influenced this work. By following it, one is naturally led to the study of problems with a kind of minimal and intrinsic structure. Besides the fact that it is much easier to find the crux of the matter in a simple structure than in a complicated one, there are not so many really basic structures, so one can hope that they will remain of interest for a very long time.
>
> Talagrand (2005, p. 5)

1 Introduction

My PhD thesis (Geyer, 1990) partially develops the theory of exponential families on abstract finite-dimensional affine spaces. Chapter 1 of the thesis discusses three "pictures" of exponential families. One can take the space where the canonical statistic takes values to be $\mathbb{R}^d$ for some d, an abstract finite-dimensional vector space, or an abstract finite-dimensional affine space.

The $\mathbb{R}^d$ picture has nothing to recommend it except that almost all of the literature uses it and people are familiar with it. There is a lot of structure that $\mathbb{R}^d$ has over and above its vector space structure. It has a preferred coordinate system, a preferred inner product, and a preferred metric, among other things. An abstract vector space has none of these things. One can introduce coordinates or an inner product or a norm or a metric, but there is no unique way to do so. An abstract affine space has even less structure.

It has no preferred point (no origin), so every point is just like every other point.

So what is the harm in using all of this familiar structure? There are two kinds of harm. Extra irrelevant structure invites obscuring the essence of a subject by relying on the irrelevant extra structure in proofs and even in definitions. But even worse, it invites producing results and procedures that do not generalize to the abstract situation and hence are of questionable interest. We need the abstract picture to tell us what is interesting and essential.

In the process we will see that we have to relearn a lot of math, transferring what we know about linear algebra, differentiation, integration, convex functions, and random vectors to abstract vector and affine spaces. And all of that is very useful in all of statistics, not just for exponential families.

2 Abstract Vector Spaces

2.1 Introduction

An abstract vector space is what is defined at the beginning of any linear algebra book. We will use Halmos (1974) for our linear algebra references, but any good abstract linear algebra book will do as well. Section 2 of Halmos (1974) has the vector space axioms. A vector space over a field consists of a set of objects called vectors that participate in two operations, vector addition and scalar multiplication (scalars being another name for elements of the field) that satisfy a list of axioms.

In exponential family theory we are only interested in finite-dimensional real vector spaces, meaning the field is $\mathbb{R}$, the real numbers, and there is a finite basis (Halmos, 1974, Section 8).

2.2 Isomorphism

A vector space isomorphism is an invertible linear transformation. The inverse function is automatically linear (Halmos, 1974, Section 36). Every finite-dimensional abstract vector space V is isomorphic to $\mathbb{R}^d$ for some d (Halmos, 1974, Theorem in Section 9), meaning there is an invertible linear transformation $V \to \mathbb{R}^d$. But there are many such isomorphisms and no favored isomorphism.

Every such isomorphism can be thought of as providing a coordinate system for V. But there is no preferred coordinate system.

2.3 Vectors, Matrices, and Functions

The unsophisticated view of linear algebra is that vectors are a special case of matrices which come in two kinds: row vectors and column vectors. The sophisticated view is the reverse: matrices are a special case of vectors; any things that can be added and multiplied by scalars are vectors.

Real-valued and vector-valued functions are also vectors. Vector addition and scalar multiplication are defined in the obvious way

$$(f + g)(x) = f(x) + g(x) \tag{1a}$$

$$(af)(x) = af(x) \tag{1b}$$

where f and g are scalar-valued or vector-valued functions having the same domain and codomain and a is a scalar. This is the reason that the study of infinite-dimensional topological vector spaces (which will be defined in the following section) is called "functional analysis."

In particular, if U and V are vector spaces, then the space of all linear functions $U \to V$ is another vector space (Halmos, 1974, Section 33).

The space of all functions $U \to V$ is also a vector space, but that is not important in what follows.

2.4 Topology

A *vector topology* for an abstract real vector space V is a topology satisfying the following axioms (Rudin, 1991, Section 1.6).

- Vector addition is a continuous operation $V \times V \to V$.
- Scalar multiplication is a continuous operation $\mathbb{R} \times V \to V$.
- Points are closed sets.

An abstract vector space equipped with a vector topology is called a *topological vector space*. For an infinite-dimensional abstract vector space, there may be many vector topologies. It is commonplace in the study of infinite-dimensional topological vector spaces, which is called *functional analysis*, to use more than one vector topology for the same space, sometimes in the same argument.

A topological space isomorphism (also called homeomorphism) is an invertible function such that both the function and its inverse are continuous: they map open sets to open sets and convergent sequences to convergent sequences. A topological vector space isomorphism is a function that is

both a vector space isomorphism and a topological space isomorphism: an invertible function such that both the function and its inverse are linear and continuous.

A finite-dimensional real vector space V has exactly one vector topology, and any linear isomorphism $V \to \mathbb{R}^d$ is also a topological isomorphism (Rudin, 1991, Theorem 1.21, this theorem is stated for complex rather than real scalars, but the comment at the end of Rudin's Section 1.19 says this is also valid for real scalars, and indeed examination of the proofs of Rudin's Lemma 1.20 and Theorem 1.21 shows that they are valid when real scalars are substituted for complex scalars and $\mathbb{R}^n$ for $\mathbb{C}^n$).

The topology for $\mathbb{R}^d$ is the "usual" topology, which is the product topology (open sets are unions of open boxes) inherited from the "usual" topology for $\mathbb{R}$, which is the order topology (open sets are unions of open intervals).

Every linear function from one abstract finite-dimensional vector space to another is continuous (Rudin, 1991, Theorems 1.18 and 1.21). It follows that every invertible linear function between abstract finite-dimensional topological vector spaces is a topological vector space isomorphism.

Every vector subspace of a finite-dimensional topological vector space is a closed set (Rudin, 1991, Theorem 1.21).

2.5 Transfer

It follows from the previous section that any results that are purely linear-algebraic or purely topological can be transferred from what we call the "$\mathbb{R}^d$ picture" (finite-dimensional vectors are elements of $\mathbb{R}^d$ or perhaps $d \times 1$ matrices) to what we call the "abstract vector space picture" (abstract d-dimensional vector spaces are not $\mathbb{R}^d$, they are isomorphic to $\mathbb{R}^d$, but isomorphism is not equality).

The question is: what is purely linear-algebraic or topological? If there is more structure to a concept, then perhaps the additional structure is not transferred by linear isomorphism, and we will want to investigate transfer.

The reason for the interest in transfer is that most of the literature and most textbooks use the $\mathbb{R}^d$ picture so we need to transfer their results to the abstract picture in order not to have to rewrite the whole literature.

3 Abstract Affine Spaces

3.1 Introduction

An abstract affine space is what you learn about in high school geometry (Euclidean geometry) and then never see again (until now). Euclidean geometry is almost entirely replaced in modern (college level and above) mathematics by linear algebra.

Roughly speaking, an affine space is what you get when you start with a vector space and forget where the origin is. Conversely, a vector space is what you get when you start with an affine space and chose an arbitrary point to serve as the origin. In an affine space, every point is just like every other point. In a vector space, the origin is very special.

Affine spaces can be given an abstract axiomatic treatment (Godement, 1963; Bourbaki, 1970; Appendix A.2 of Geyer, 1990), but they are easier to understand (nowadays) if considered as subspaces of vector spaces. This is the approach taken in *Advanced Linear Algebra* (Roman, 2008, Chapter 16), *Convex Analysis* (Rockafellar, 1970, Section 1.1), and *Variational Analysis* (Rockafellar and Wets, 1998, Section 2.B).

Affine subspaces of vector spaces and affine functions appear here and there in applications but may escape notice because in many areas of applied mathematics affine functions are called "linear functions" in conflict with linear algebra and all of mathematics more advanced than that (including real and functional analysis and any subject with algebra or algebraic in its name) and because many affine subspaces are not named as such but given special names like point, line, plane, hyperplane. Thus many people are familiar with affine spaces and affine functions but call them something else.

Although affine subspaces can be defined in any vector space, we will only be interested in affine subspaces of real finite-dimensional vector spaces.

3.2 Definitions

3.2.1 Translates

For subsets A and B of a vector space, their *Minkowski sum* is

$$A + B = \{\, x + y : x \in A \text{ and } y \in B \,\}.$$

This operation is called *Minkowski addition* of sets.

When one of the sets is a singleton, we abuse notation by writing $x + A$ rather than $\{x\} + A$. This sum of a vector x and and a set of vectors A is called a *translate* of A.

3.2.2 Affine Spaces

A nonempty subset A of a vector space E is an *affine space* if and only if it is a translate of a vector subspace, that is, $A = x + V$, where V is a vector subspace of E. When we consider A as a substructure of E, we say it is an *affine subspace* of E.

In linear algebra, it is typical for the word "subspace" without qualification to refer to a vector subspace, but we will always say "vector subspace" or "affine subspace." If A and B are affine subspaces of a vector space and $A \subset B$ then we also say that A is an affine subspace of the affine space B.

The mathematical object V is called the *translation space* of A or (especially when discussing differentiation) the *tangent space* of A.

The empty set is also considered an affine space, but it does not have a translation space. It is an affine subspace of every affine space and every vector space.

The reason why the empty set is considered an affine subspace is to make valid Theorem 18 and Lemma 27 below. Without the empty set being considered an affine subspace, these theorems and others would need ugly conditions to rule out emptiness. Also the definition (25) of convex set is analogous to the characterization of affine set in Theorem 16 and both say the empty set satisfies the criterion (vacuously, because there is nothing to check unless the set has at least two points, so every empty set or singleton set is both convex and affine).

Every vector space is also an affine space because $x + V = V$ for any $x \in V$. And every vector space when considered as an affine space is its own translation space.

3.2.3 Types and Operations

The mathematics of affine spaces can be described two different ways: affine spaces can be embedded in vector spaces, so the mathematics of affine spaces is part of linear algebra, or we can consider affine spaces as objects in their own right, as the axiomatic definition of affine spaces does. Either way is mathematically equivalent to the other, so we are free to choose the one easiest to understand.

We will use parts of both. Formally, we have defined affine spaces as subsets of vector spaces. But we will use terminology and notation that applies to either approach.

In the mathematics of vector spaces (linear algebra) there are two types of objects, scalars and vectors, and two operations, vector addition and scalar multiplication. One cannot talk about vectors by themselves, only in conjunction with scalars (elements of the field the vector space is over).

When we consider affine spaces abstractly, as objects in their own right, we do not mention an enclosing vector space (because that is not part of the axiomatic definition). This is similar to the way we often do not mention a basis for a vector space or do not mention the probability space on which a random variable is defined. We consider just the affine space, its translation space, and the scalar field of its translation space. We do not mention any elements of the enclosing vector space that are not in the affine space under discussion or in its translation space.

Thus there are three types of objects: scalars, vectors, and points. The points are elements of the affine space (considered as a mathematical object in its own right), the vectors are elements of the translation space of the affine space, and the scalars are the elements of the field the translation space is over. There are two operations that involve points. The difference of points is a vector, and a point plus a vector gives a point. In symbols, if A is an affine space having translation space V, then $x - y \in V$ whenever $x, y \in A$, and $x + v \in A$ whenever $x \in A$ and $v \in V$. Moreover $x - y = v$ if and only if $x = y + v$.

When we are thinking abstractly like this, it makes no sense to multiply points by scalars (this would take us outside the affine space and its translation space into the enclosing vector space, which we do not want to mention) Thus if A is an affine space, and $x_1, \ldots, x_n$ are points in A and $a_1, \ldots, a_n$ are scalars, it makes no sense to write the general linear combination

$$a_1 x_1 + \cdots + a_n x_n$$

(this is, in general, not a point in A). What we can do is, if x_0 is another point in A, write the general affine combination

$$x_0 + a_1(x_1 - x_0) + \cdots + a_n(x_n - x_0) \tag{2}$$

(now $x_i - x_0$ is a vector, so $a_i(x_i - x_0)$ is a vector, so $\sum_i a_i(x_i - x_0)$ is a vector, so $x_0 + \sum_i a_i(x_i - x_0)$ is a point).

Lemma 1. *The affine combination* (2) *does not depend on* x_0 *if the* a_i *sum to one.*

Proof.

$$\left[x + \sum\nolimits_i a_i(x_i - x)\right] - \left[y + \sum\nolimits_i a_i(x_i - y)\right] = (x - y)\left(1 - \sum\nolimits_i a_i\right)$$

□

3.3 Dimension

We say the dimension of a non-empty affine space is the dimension of its translation space. The empty affine space does not have a dimension.

When we consider affine spaces as subsets of vector spaces, we can always consider finite-dimensional affine spaces as subsets of finite-dimensional vector spaces. If $A = x + V$, where x is a point and V is a vector subspace, and $\mathcal{B}$ is a finite basis for V, then $\{x\} \cup \mathcal{B}$ spans a finite-dimensional vector space containing A and V.

3.4 Topology

When we consider affine spaces as subspaces of an enclosing vector space, they get the subspace topology: if A is an affine subspace of a finite-dimensional vector space E, then O is open in A if and only if there exists an open set W in E such that $O = A \cap W$.

Theorem 2. *If A is a finite-dimensional affine space, V is its translation space, and $x \in A$, then the map $V \to A$ given by $v \mapsto x + v$ is a homeomorphism (isomorphism of topological spaces).*

Note that the inverse of $v \mapsto x + v$ is $y \mapsto y - x$.

Proof. What must be shown is that O is open in A if and only if $O - x$ is open in V. Let E be a finite-dimensional vector space enclosing A and V. Suppose O is open in A, so there exists W open in E such that $O = W \cap A$. Then $W - x$ is open in E because $y \mapsto y - x$ is an invertible linear function, hence a topological vector space isomorphism of E. Hence $O - x = (W - x) \cap V$ is open in V. The other direction of the proof is similar. □

Thus when we consider affine spaces without mentioning an enclosing vector space we can use Theorem 2 to define the topology. A set O is open in A if and only if $O - x$ is open in V. And any $x \in A$ can be used here.

Of course the empty affine space has the only topology it can have. In every topological space the empty set is both closed and open, and the empty set is the only subset of the empty space.

Corollary 3. *Every affine subspace of a finite-dimensional affine space is closed.*

Proof. We already know that every vector subspace of a finite-dimensional vector space is closed (Section 2.4 above). The empty set is closed in any topology. Suppose A is an affine subspace of a finite-dimensional affine space B, and suppose $x \in A$. Then $A - x$ is a vector subspace of the translation space of B, hence a closed subset of this translation space. Theorem 2 says that the map $v \mapsto x + v$ is a topological isomorphism. Since it maps $A - x$ to A, that proves A is a closed subset of B. □

3.5 Structure-Preserving Functions

In many areas of mathematics there are structure-preserving functions, which are considered just as important as the mathematical objects they go between. In set theory, all functions are structure-preserving because abstract sets have no structure to preserve. In general topology, the structure-preserving functions are the continuous functions. In linear algebra, the structure-preserving functions are the linear functions. In the study of affine spaces, the structure-preserving functions are the affine functions.

We say a function f that maps points to points and vectors to vectors is structure-preserving if

$$\begin{aligned} f(x - y) &= f(x) - f(y) \\ f(x + v) &= f(x) + f(v) \end{aligned} \tag{3}$$

but it seems strange to have one function work on two types of elements. So we denote the part of f that maps vectors to vectors by a different symbol g. This gives

$$\begin{aligned} g(x - y) &= f(x) - f(y) \\ f(x + v) &= f(x) + g(v) \end{aligned}$$

which no longer looks so structure-preserving but conforms to ordinary usage. We want g to be structure-preserving between vector spaces, so it must be a linear function.

Formalizing this discussion gives the following definition. A function $f : A \to B$ between affine spaces having translation spaces U and V, respectively, is *affine* if there exists a linear function $g : U \to V$ such that

$$g(x - y) = f(x) - f(y), \qquad x, y \in A \tag{4a}$$

$$f(x + v) = f(x) + g(v), \qquad x \in A,\ v \in U \tag{4b}$$

Note that (4b) holds if and only if

$$f(x+v) - f(x) = g(v), \qquad x \in A,\ v \in U \tag{5}$$

and this holds if and only if (4a) holds. Thus each of (4a), (4b), and (5) implies the others, and we can take any one of them as our characterization of affine functions.

Theorem 4. *With the setup above,* (5) *holds with g a linear function if and only if*

$$f(x+v) - f(x) = g(v), \qquad v \in U, \tag{6}$$

for some $x \in A$.

That is, if (6) holds for one $x \in A$, then it holds for all $x \in A$.

Proof. One direction is trivial. So assume (6) holds for some $x \in A$. For $x^* \in A$ and $v \in V$ we have

$$\begin{aligned} f(x^* + v) - f(x^*) &= f(x + [x^* - x + v]) - f(x + [x^* - x]) \\ &= g(x^* - x + v) - g(x^* - x) \\ &= g(v) \end{aligned}$$

□

We call the function g defined by (4a) (4b), (5), or (6) the *associated linear function* of f. It is clear from (5) that the linear function g associated with an affine function f is unique.

Later on (Section 4.3 below) we shall learn another name for the associated linear function (it's a derivative).

Theorem 5. *An affine function between vector spaces is a linear function plus a constant function.*

Proof. Taking the case $x = 0$ in (4b) gives $f(v) = f(0) + g(v)$. □

There is precisely one function from the empty set to any other set, the empty function, which has an empty graph (the graph of a function is the set of its argument-value pairs). If A is the empty affine space and B is any other affine space, then we consider the empty function $A \to B$ to be an affine function. There are no functions from any nonempty set to the empty set, hence no functions from any nonempty affine space to the empty affine space,

Empty affine functions have no associated linear functions.

3.6 Category Theory

3.6.1 Introduction

Set theory was "new math" peddled as the foundations of mathematics around 1900. It percolated through mathematics in the first half of the twentieth century. An example is how families of sets (sigma-algebras) are basic objects in probability theory.

Category theory was "new math" peddled as the foundations of mathematics around 1950. It percolated through mathematics in the second half of the twentieth century. It has as yet had little effect on probability theory and mathematical statistics. We will make only the most naive use of it here.

3.6.2 Axioms

Category theory is like set theory except that functions are treated as first class things and also the function concept is generalized. A category consists of *objects* and *morphisms*. The morphisms are function-like in that they go between objects. We write $f : A \to B$ for a morphism f from A to B. Or in displays

$$A \xrightarrow{f} B$$

Like with functions, we call A the *domain* of f and B the *codomain* of f. Another terminology says *arrow* rather than morphism, *source* rather than domain, and *target* rather than codomain. Like functions, morphisms are composable. If we have morphisms

$$A \xrightarrow{f} B \xrightarrow{g} C$$

then there is also a morphism $g \circ f : A \to C$.

For every object A there is a morphism $\mathrm{id}_A : A \to A$ that "does nothing" in composition, that is,

$$\begin{aligned} f \circ \mathrm{id}_A &= f, \qquad \forall f : A \to B \\ \mathrm{id}_A \circ g &= g, \qquad \forall g : C \to A \end{aligned}$$

These are called the *identity axioms* or *identity laws*, and id_A is called the *identity morphism* for A. It is easy to show from the identity axiom that every object has a unique identity morphism.

We also have one other axiom or law, *associativity of composition:* whenever

$$A \xrightarrow{f} B \xrightarrow{g} C \xrightarrow{h} D \tag{7}$$

we have $(h \circ g) \circ f = h \circ (g \circ f)$. Thus it is clear what (7) means even though there are no parentheses indicating order of operations.

The final "tool" of category theory is the commutative diagram, which is a picture of objects and morphisms such that all (possibly composite) morphisms between any two objects are equal. For example, the assertion that

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ & \searrow^{h} & \downarrow{\scriptstyle g} \\ & & C \end{array}$$

is a commutative diagram says $h = g \circ f$, and the assertion that

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ {\scriptstyle g}\downarrow & \searrow^{h} & \downarrow{\scriptstyle i} \\ C & \xrightarrow{j} & D \end{array}$$

is a commutative diagram says $i \circ f = h = j \circ g$.

This doesn't seem very powerful at first glance but commutative diagrams can organize whole proofs. A few examples will appear below.

3.6.3 Isomorphism

A morphism $f : A \to B$ is an *isomorphism* if there exists a morphism $g : B \to A$ such that

$$g \circ f = \mathrm{id}_A$$
$$f \circ g = \mathrm{id}_B$$

Then g is called the *inverse* of f and written $g = f^{-1}$. It is easy to show from this definition and the identity and associativity axioms that every isomorphism has a unique inverse.

3.6.4 When Morphisms are Functions

When morphisms for some category are actually functions, we always define composition to be the usual composition of functions, that is $h = g \circ f$ means $h(x) = g(f(x))$ for all x in the domain of f (which is also the domain of h), and we always define identities to be the usual identity functions that map $x \mapsto x$.

It is easy to show that with these definitions the identity axiom and the associativity of composition axiom are satisfied.

And with this definition of identity morphisms, the inverse morphism (if it exists) is the inverse function in the usual sense.

We will be only interested in categories in which objects are sets with structure and morphisms are structure-preserving functions.

3.6.5 When Morphisms are Not Functions

(This section can be skipped.) There are many interesting categories in which morphisms are not functions, but statisticians are not familiar with the concepts. But here are two simple examples when morphisms are not functions.

Preorders Consider a category in which for any objects A and B there is at most one morphism $A \to B$. Then we usually change notation writing $A \le B$ instead of $A \to B$. For any object A, there is always the identity morphism $A \to A$, so this change of notation says $A \le A$ for all objects A. And the existence of compositions says that $A \le B \le C$ implies $A \le C$. And these two properties say that $\le$ is a reflexive and transitive relation. Such a relation is called a *preorder*.

All partial orders and total orders are special cases of preorders. In a preorder it is possible that $A \le B$ and $B \le A$ but $A \ne B$ (so A and B are isomorphic but not equal). If we add an axiom that $A \le B$ and $B \le A$ implies $A = B$, then this makes the preorder a partial order. If we add an axiom that for any objects A and B either $A \le B$ or $B \le A$, then this makes the partial order a total order.

Thus any order (preorder, partial order, or total order) can be thought of as a category in which there is a morphism $A \to B$ if and only if $A \le B$.

Groupoids Consider a category in which every morphism is an isomorphism. Such a mathematical structure is called a *groupoid*.

If we add an axiom that every pair of morphisms be composable then in order that id_A and id_B be composible we must have $A = B$. Thus we must have only one object.

This is now a *group* if we think of

- the elements of the group as the morphisms of the category and
- the group multiplication operation as the composition of morphisms in the category.

Then the category axioms imply the group axioms.

- Associativity of the group multiplication operation is implied by the associativity axiom for categories (and the assumption that every pair of morphisms are composable).
- The identity element of the group is the identity morphism of the only object of the category (and that there is only one object is implied by the assumption that every pair of morphisms are composable).
- Then inverses in the group are the same as inverses in the category.

Thus any group can be thought of as a category in which there is only one object and every morphism is an isomorphism. And groupoid is seen to be an obvious generalization of group.

3.7 The Category of Finite-Dimensional Vector Spaces

This is the category in which the objects are finite-dimensional vector spaces and the morphisms are linear functions.

In order for this to be a category it must be that identity functions are linear and compositions of linear functions are linear. These are shown as follows:

$$\mathrm{id}_A(x+y) = x+y = \mathrm{id}_A(x) + \mathrm{id}_A(y)$$

and

$$\mathrm{id}_A(cx) = cx = c \cdot \mathrm{id}_A(x)$$

And, if $h = g \circ f$, then

$$h(x+y) = g(f(x+y)) = g(f(x)+f(y)) = g(f(x)) + g(f(y)) = h(x) + h(y)$$

and

$$h(cx) = g(f(cx)) = g(cf(x)) = cg(f(x)) = ch(x)$$

Isomorphisms in this category are linear functions whose inverses are linear. But we already know this holds for every invertible linear function (Section 2.2 above). We also see that the category theoretic notion of isomorphism agrees with the usual notion of isomorphism taken from linear algebra.

We can also consider this category as the category of finite-dimensional topological vector spaces, if we give each finite-dimensional vector space the only vector topology it can have (Section 2.4 above). The morphisms are still linear functions, and the isomorphisms are still invertible linear functions.

3.8 The Category of Finite-Dimensional Affine Spaces

This is the category in which the objects are finite-dimensional affine spaces and the morphisms are affine functions. In order for this to be a category it must be that identity functions are affine functions, and compositions of affine functions are affine functions. These are easily shown as follows. Let A be an affine space and V its translation space. Then

$$\mathrm{id}_A(x + v) = x + v = \mathrm{id}_A(x) + \mathrm{id}_V(v)$$

and

$$\mathrm{id}_V(y - x) = y - x = \mathrm{id}_A(y) - \mathrm{id}_A(x)$$

so this agrees with (3). And, using the notation in (3) where the same letter denotes both an affine function and its associated linear function, if $h = g \circ f$, then

$$h(x + v) = g(f(x + v)) = g(f(x) + f(v)) = g(f(x)) + g(f(v)) = h(x) + h(v)$$

and

$$h(y - x) = g(f(y - x)) = g(f(y) - f(x)) = g(f(y)) - g(f(x)) = h(y) - h(x)$$

so this agrees with (3).

We also need to check empty affine functions. The identity morphism on the empty affine space is the empty function, which is an affine function by definition. The only way the empty affine space can appear in a composition is

$$\varnothing \longrightarrow A \longrightarrow B$$

(because there are no arrows to the empty set except for the empty morphism) and this composition is the empty function (because the empty function is the only function whose domain is the empty set), hence this composition is an affine function by definition.

Isomorphisms in this category are affine functions whose inverses are affine. So we need a theorem about that.

Theorem 6. *If an affine function is invertible, then its inverse is affine.*

Before we prove this, we rewrite Theorem 4 so it looks more category theoretic.

Lemma 7. *Suppose A and B are affine spaces having translation spaces U and V, respectively, and $a \in A$. For any affine space containing a point x, let s_x denote the map $y \mapsto y - x$ between that affine space and its translation space. Consider the commutative diagram*

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ {\scriptstyle s_a}\downarrow & & \downarrow{\scriptstyle s_{f(a)}} \\ U & \xrightarrow{g} & V \end{array} \tag{8}$$

Then f is an affine function if and only if g is a linear function.

Note that s_x is invertible and its inverse is $v \mapsto x + v$.

Proof. Commutative diagram means $g = s_{f(a)} \circ f \circ s_a^{-1}$, or

$$\begin{aligned} s_a^{-1}(u) &= a + u \\ f(s_a^{-1}(u)) &= f(a + u) \\ s_{f(a)}(f(s_a^{-1}(u))) &= f(a + u) - f(a) \end{aligned}$$

Thus this lemma merely translates the characterization of Theorem 4 into category theoretic language. □

Proof of Theorem 6. In (8) we know that s_a and $s_{f(a)}$ are invertible by the comment following following the statement of the lemma, and we assume in the statement of the theorem that f is invertible. Thus g is invertible and its inverse is $s_a \circ f^{-1} \circ s_{f(a)}^{-1}$. Then we know from linear algebra (Section 2.2 above) that g^{-1} must be a linear function. Now it follows from the lemma that f^{-1} defined by the commutative diagram

$$\begin{array}{ccc} A & \xleftarrow{f^{-1}} & B \\ {\scriptstyle s_a}\downarrow & & \downarrow{\scriptstyle s_{f(a)}} \\ U & \xleftarrow{g^{-1}} & V \end{array}$$

is an affine function. □

Thus every invertible affine function is an isomorphism. We know from linear algebra that finite-dimensional vector spaces are isomorphic if and only if they have the same dimension (Halmos, 1974, Section 9).

Theorem 8. *Nonempty finite-dimensional affine spaces are isomorphic if and only if they have the same dimension. The empty affine space is isomorphic only to itself.*

Proof. In any category any object is isomorphic to itself because the identity morphism is always an isomorphism (it is its own inverse). The empty affine space can only be isomorphic to itself, because there are no functions that map from a nonempty set to the empty set.

This leaves us with the case that both affine spaces are nonempty. We know from the proofs of Theorem 6 and Lemma 7 that such affine spaces have an isomorphism if and only if we have a commutative diagram (8) with both f and g invertible functions. But there is such an invertible g if and only if U and V have the same dimension (from linear algebra), and s_a and $s_{f(a)}$ map between spaces of the same dimension by definition of the dimension of an affine space (Section 3.3 above). It follows that A and B have the same dimension if and only if there exist invertible f and g that give such a diagram. □

Corollary 9. *Every d-dimensional affine space is isomorphic to $\mathbb{R}^d$.*

We can also consider this category as the category of finite-dimensional topological affine spaces, if we give each finite-dimensional affine space the only topology it can have (Section 3.4 above). The morphisms are still affine functions, and the isomorphisms are still invertible affine functions. But we need a theorem to establish that.

Theorem 10. *Every affine function between finite-dimensional affine spaces is continuous.*

Proof. Theorem 2 asserts that s_a and $s_{f(a)}$ in (8) are homeomorphisms. Every linear function is continuous (Section 2.4 above). And the composition of continuous functions

$$f = s_{f(a)}^{-1} \circ g \circ s_a$$

is continuous. □

We also tie up one loose end.

Theorem 11. *The function $s_x : y \mapsto y - x$ is a topological affine space isomorphism.*

Proof. We already know s_x is invertible. So we only have to show it is affine, because then Theorems 6 and 10 show that both s_x and s_x^{-1} are affine and continuous.

To check that it is affine we use Theorem 4.

$$g(v) = s_x(y + v) - s_x(y) = (y + v - x) - (y - x) = v$$

so g is the identity function, which is linear. □

4 Calculus

4.1 Integration

If we have a topology, then we know the open sets and the Borel sigma-algebra (the smallest sigma-algebra containing the open sets). So we can identify Borel measures and Borel-measurable functions and integrals of real-valued Borel-measurable functions with respect to Borel measures.

Transfer of integrals by affine isomorphism from $\mathbb{R}^d$ to any abstract finite-dimensional affine space is accomplished by the change-of-variable theorem for abstract integration.

For any measurable function f from a measurable space $(A, \mathcal{A})$ to a measurable space $(B, \mathcal{B})$, any measure μ on A induces a measure ν on B defined by

$$\nu(C) = \mu\big(f^{-1}(C)\big), \qquad C \in \mathcal{B},$$

where

$$f^{-1}(C) = \{\, x \in A : f(x) \in C \,\}$$

defines the set-to-set inverse of f. This ν is called the *image* of μ under f, and this operation is denoted $\nu = \mu \circ f^{-1}$. Moreover, when μ and ν have this relation, a real-valued function g on B is integrable with respect to ν if and only if $g \circ f$ is integrable with respect to μ, in which case

$$\int (g \circ f)\, d\mu = \int g\, d\nu = \int g\, d(\mu \circ f^{-1}) \tag{9}$$

(Billingsley, 1979, Theorem 16.12).

Discussion of the other change-of-variable theorem, for densities with respect to Lebesgue measure (the one involving Jacobian determinants), which involves differentiation, will have to wait until differentiation has been discussed (Section 4.4 below).

Using the same notation f^{-1} for the set-to-set inverse (which every function has) and the point-to-point inverse (which only invertible functions have) is commonplace in math. It usually does not cause confusion because it does not conflict with the forward image operation

$$f(B) = \{\, f(x) : x \in B \,\}$$

When f is an invertible function, $f^{-1}(B)$ is the same set, whether we apply the set-to-set inverse notion or think of this as meaning the forward image through the point-to-point inverse.

4.2 Differentiation on Vector Spaces

4.2.1 Definition

The abstract theory of differentiation is less familiar to statisticians than the abstract theory of integration, but it can also be found in functional analysis and differential geometry. Here we follow Lang (1993). The reader must excuse the appearance of Banach spaces (complete normed vector spaces, possibly infinite-dimensional). We will specialize to the finite-dimensional special case as soon as the definitions are finished. The reason we introduce the functional analysis definitions is to show that coordinates (isomorphism to $\mathbb{R}^d$) play no essential role in differentiation, contrary to the impression one gets from multivariable calculus.

Let U and V be Banach spaces. Then $L(U, V)$ denotes the set of all continuous linear maps $U \to V$, which is itself a vector space, the operations being given by (1a) and (1b). It is also a Banach space (Lang, 1993, pp. 65–66) when given the norm defined by

$$\|f\| = \sup_{\substack{x \in U \\ \|x\| \le 1}} \|f(x)\| \tag{10}$$

in which the $\|\cdot\|$ notation refers to three different norms: the expression $\|x\|$ refers to the norm of U, the expression $\|f(x)\|$ refers to the norm of V, and the expression $\|f\|$ refers to the norm for $L(U, V)$ which (10) defines.

Let O be open in U and let $f : O \to V$ be a map. Then f is *differentiable* at a point $x \in O$ if there exists $g \in L(U, V)$ such that

$$\lim_{h \to 0} \frac{f(x + h) - f(x) - g(h)}{\|h\|} = 0. \tag{11}$$

in which case g is the unique element of $L(U, V)$ having this property (Lang, 1993, p. 334) and we say that g is the *derivative of f at x* and write $f'(x) = g$.

If f is differentiable at every point of O, then it defines a map $x \mapsto f'(x)$ from O to $L(U, V)$. If this map is continuous, we say f is *continuously differentiable* (on O).

When this differentiation theory on Banach spaces is specialized to abstract finite-dimensional vector spaces, it becomes simpler. On a finite-dimensional vector space, *every* linear function is continuous (Section 2.4 above), so $L(U, V)$ consists of *all* linear functions $U \to V$ (one doesn't need to say "continuous linear function" in the finite-dimensional case). Moreover, on a finite-dimensional vector space, all norms are equivalent (Lang, 1993, Corollary 3.14 of Chapter II), meaning for any norms $\|\cdot\|_1$ and $\|\cdot\|_2$

there exist constants c_1 and c_2 such that $\|x\|_1 \leq c_2\|x\|_2$ and $\|x\|_2 \leq c_1\|x\|_1$ for all x, so the choice of norm does not affect the derivative. This also means that every norm on a finite-dimensional vector space induces the same topology, but we already knew that. Norms induce vector topologies and there is only one vector topology a finite-dimensional vector space can have (Section 2.4 above).

4.2.2 Philosophy

This is very different from the conceptualization of the derivative one gets from calculus. There the derivative of a scalar-to-scalar function is just a number; here it is a linear function. The correspondence is that the slope of the linear function is the derivative in the ordinary calculus sense.

In multivariate calculus the derivative of a vector-to-vector function $f : \mathbb{R}^d \to \mathbb{R}^e$ is the matrix of partial derivatives $\partial f_i(x)/\partial x_j$ (the so-called Jacobian matrix); here it is a linear function $f'(x) : \mathbb{R}^d \to \mathbb{R}^e$, the linear function represented by the Jacobian matrix (that is, if J is the Jacobian matrix, then the linear function is $x \mapsto Jx$).

On an abstract finite-dimensional vector space there are no coordinates hence no Jacobian matrix. We can introduce coordinates, but there are many ways to do so, and each gives a different Jacobian matrix. But the abstract derivative is a unique linear function, which does not depend on coordinates. (Each Jacobian matrix is the representation of that unique linear function in some coordinate system.)

4.2.3 Higher Order Derivatives

Second and higher derivatives are just derivatives of derivatives. If the map $f' : O \to L(U, V)$, where O is open in U, is differentiable at x, then we write its derivative as $f''(x)$. It is, by definition, an element of $L\big(U, L(U, V)\big)$. Its value at some point $h_1 \in U$, written $f''(x)(h_1)$ is an element of $L(U, V)$. And in turn, the value of this at some point $h_2 \in U$, written $f''(x)(h_1)(h_2)$ is an element of V.

The map $(h_1, h_2) \mapsto f''(x)(h_1)(h_2)$ is bilinear (linear in both arguments) and continuous $U \times U \to V$. Thus we can also consider $f''(x)$ a continuous bilinear form on U (Lang, 1993, p. 343, ff.). If f is twice continuously differentiable, meaning the map $x \mapsto f''(x)$ is continuous from some neighborhood of x in O to $L\big(U, L(U, V)\big)$, then this bilinear form is symmetric, meaning

$$f''(x)(h_1)(h_2) = f''(x)(h_2)(h_1), \qquad h_1, h_2 \in U.$$

Similarly, a continuous third derivative can be identified with a symmetric trilinear form, a continuous fourth derivative with a symmetric tetralinear form, and so forth.

4.2.4 Type Theory

It is convenient to steal some notation from type theory as used in functional programming (computer languages like Haskell or R) and also the twenty-first century's new foundations of mathematics, homotopy type theory. We say the type of f'' is $U \to U \to V$, the interpretation being that this notation is right associative $U \to (U \to V)$. This is easier to write and easier to read than $L(U, L(U, V))$, because $U \to U \to V$ reads left to right and the other reads inside out. In $U \to U \to V$ we have lost the L's in $L(U, L(U, V))$ that told us the functions are linear, so we will have to get that from the context.

Of course, the official type of the other interpretation of f'' as a symmetric bilinear form is $U \times U \to V$. But the equivalence of these two types is well known in functional programming, where it is called *currying*. The programming language Haskell is named after the logician Haskell Curry (1900–1982), and currying is also named after him.

Writing $f''(x)(h_1)(h_2)$ is thinking of f'' in curried type $U \to U \to V$. It is when we write it as a function of two arguments that it has the uncurried type $U \times U \to V$. We can write it as a function of two arguments $f''(x)(h_1, h_2)$, but this is sloppy. To be pedantically correct (a set theorist is a person who thinks all functions have only one argument), the uncurried form is a function with one argument (in $U \times U$), which is a pair, so we should write the really ugly $f''(x)\big((h_1, h_2)\big)$ for it to actually have the type $U \times U \to V$.

In Haskell and other functional programming languages the curried form is considered prettier, so we, like the functional programmers, prefer to write $f''(x)(h_1)(h_2)$ for the mathematical object and $U \to U \to V$ for the type.

4.2.5 More Philosophy

So higher derivatives are even more different from the conceptualization of higher derivatives one gets from calculus. Instead of second, third, fourth, etc. derivatives of scalar-to-scalar functions being just numbers, they are now symmetric bilinear, trilinear, tetralinear, etc. forms (or the alternate interpretation as linear functions between vector spaces). That seems crazy, but for vector-to-vector functions it is not so crazy. Once the num-

ber of indices gets to more than two, so partial derivatives can no longer be laid out in a matrix, as with second derivatives of a vector-to-vector function $\partial^2 f_i(x)/\partial x_j \partial x_k$, the conventions of multivariable calculus become inconvenient too.

Crazy or not, we will use the PhD level real analysis theory of derivatives as linear functions or multilinear forms in the rest of this document.

4.2.6 The Constant Rule

As in ordinary calculus, and as is obvious from the definition of differentiation (and uniqueness of the derivative), the derivative of a constant function is zero.

In detail, let U and V be vector spaces and let O be open in U. If $f : O \to V$ is a constant function, then $f'(x)$ is the zero function $U \to V$ for all $x \in O$.

4.2.7 The Other Constant Rule

As in ordinary calculus, and as is obvious from the definition of differentiation (and uniqueness of the derivative), constants come out of derivatives. If $h = af$, where a is a scalar constant, then $h'(x) = af'(x)$.

4.2.8 The Addition Rule

As in ordinary calculus, and as is obvious from the definition of differentiation (and uniqueness of the derivative), the derivative of a sum is the sum of the derivatives. If $h = f + g$, then $h'(x) = f'(x) + g'(x)$.

4.2.9 The Linear Function Rule

Not as in ordinary calculus, but as is obvious from the definition of differentiation (and uniqueness of the derivative), the derivative of a linear function is that linear function.

If f is a linear function between vector spaces, then $f'(x) = f$ for all x. Thus f' is a constant function, and $f''(x) = 0$ for all x.

In detail, let U and V be vector spaces, and let $f : U \to V$ be a linear function. Then $f'(x) = f$ for all $x \in U$, and $f''(x)$ is the zero function $U \to U \to V$ for all $x \in U$. So $f'(x)(h) = f(h)$ for all $x, h \in U$ and $f''(x)(h_1)(h_2) = 0$ for all $x, h_1, h_2 \in U$.

4.2.10 Linearity of Differentiation

The combination of the properties in Sections 4.2.7 and 4.2.8 above is sometimes called *linearity of differentiation*, by which it is meant that differentiation is a linear operator on certain vector spaces.

In detail, let U and V be vector spaces and let O be open in U. Then the set of all functions $f : O \to V$ that are differentiable at $x \in O$ is itself a vector space because it is closed under vector addition (Section 4.2.8) and scalar multiplication (Section 4.2.7). Call that vector space $\mathcal{F}$. Then the function $\mathcal{F} \to L(U, V)$ defined by $f \mapsto f'(x)$ is a linear function, again by Sections 4.2.7 and 4.2.8 above.

And, conversely, to say that $f \mapsto f'(x)$ is a linear function is just to say that the properties in Sections 4.2.7 and 4.2.8 above hold.

4.2.11 The Multiplication Rule

This section is about the rule, familiar from ordinary calculus

$$(fg)'(x) = f'(x)g(x) + f(x)g'(x) \tag{12}$$

but it is not at all obvious what the analog could mean using our definitions of differentiation. For one thing, there is no single notion of multiplication of vectors, but rather several notions applicable in different situations (dot product, inner product, outer product, cross product, matrix multiplication, composition of linear functions, and perhaps others). Lang (1993, p. 336) gives one formulation of the multiplication rule that applies to all of these situations.

Let E, U, V, and W be Banach spaces, and let $U \times V \to W$ be a continuous bilinear function that we denote by juxtaposition. Let O be open in E, and let $f : O \to U$ and $g : O \to V$ be differentiable at $x \in O$. Then the product map $fg : O \to W$ is also differentiable at x and (12) holds, although it is not obvious what this equation means without further interpretation. What it means is

$$(fg)'(x)(h) = f'(x)(h)g(x) + f(x)g'(x)(h), \qquad h \in E. \tag{13}$$

In even more detail, $f'(x)$ is a linear function $E \to U$, so $f'(x)(h)$ is an element of U. Similarly, $g'(x)$ is a linear function $E \to V$, so $g'(x)(h)$ is an element of V. Hence both terms $f'(x)(h)g(x)$ and $f(x)g'(x)(h)$ are instances of our multiplication operation $U \times V \to W$. So the right-hand side of (13) is an element of W and the $+$ operation is just addition of elements of W. On the left-hand side of (13) fg is a function $O \to W$ so $(fg)'(x)$ is a linear function $E \to W$, and $(fg)'(x)(h)$ is an element of W.

4.2.12 The Inversion Rule

Let $\operatorname{Lis}(U, V)$ denote the set of invertible linear functions $U \to V$, and let $i : f \mapsto f^{-1}$ denote the inversion function $\operatorname{Lis}(U, V) \to L(V, U)$, then

$$i'(f)(h) = -f^{-1} \circ h \circ f^{-1}, \qquad f \in \operatorname{Lis}(U, V),\ h \in L(U, V) \tag{14}$$

(Lang, 1993, Exercise 3 of Chapter XIII, p. 357).

On the left-hand side of (14) $i'(f)$ maps $\operatorname{Lis}(U, V) \to L(V, U)$ so $i'(f)(h)$ is an element of $L(V, U)$. On the right-hand side of (14) we have

$$V \xrightarrow{f^{-1}} U \xrightarrow{h} V \xrightarrow{f^{-1}} U$$

and this is also an element of $L(V, U)$.

4.2.13 The Chain Rule

Let U, V, and W be Banach spaces, let O be open in U and P be open in V, and let $f : O \to P$ and $g : P \to W$ be maps. Then the *chain rule* says that if f is differentiable at x and g is differentiable at $f(x)$, then the composition $h = g \circ f$ is differentiable at x and its derivative is given by

$$h'(x) = g'\big(f(x)\big) \circ f'(x) \tag{15}$$

(Lang, 1993, p. 337). This says that $h'(x)$ is the composition of linear functions

$$\begin{array}{ccc} U & \xrightarrow{f'(x)} & V \\ & \searrow^{h'(x)} & \downarrow{\scriptstyle g'\left(f(x)\right)} \\ & & W \end{array} \tag{16}$$

4.2.14 The Inverse Function Theorem

We say a map $f : O \to V$, where O is open in U, is C^p if f is continuously differentiable p times on O. Then the inverse function theorem (Lang, 1993, p. 361–363) says the following: if $x \in O$ and $f'(x)$ is a topological vector space isomorphism, then f is a local C^p isomorphism at x, meaning there exists an open neighborhood O' of x in O such that the restriction of f to O' is one-to-one on O' and hence invertible considered as a map $O' \to f(O')$ and the inverse is C^p.

Lang's statement of the theorem does not give formulas for the derivatives, but these are given in his proof. Let $g : f(O') \to O'$ denote the local

inverse whose existence is asserted by the theorem and let $y = f(x)$. Then $g'(y)$ is the $f'(x)^{-1}$ whose existence is assumed in the theorem, that is,

$$g'(y) = f'\big(g(y)\big)^{-1}, \qquad y \in f(O') \tag{17}$$

(this is the second displayed equation on p. 363 in Lang).

Higher order derivatives are arrived at by using (17), the chain rule, and the rule for differentiating inversion. In more detail, (17) can be rewritten as

$$g' = i \circ f' \circ g$$

which gives g' explicitly as the composition of differentiable maps, that is, the diagram

$$\begin{array}{ccc} U & \xrightarrow{f'} & \operatorname{Lis}(U, V) \\ {\scriptstyle g}\uparrow & & \downarrow{\scriptstyle i} \\ f(O') & \xrightarrow{g'} & L(V, U) \end{array}$$

is commutative. For example, the second derivative of the local inverse whose existence is guaranteed by the inverse function theorem if f is twice continuously differentiable is given by

$$\begin{aligned} g''(y) &= i'\big((f' \circ g)(y)\big) \circ (f' \circ g)'(y) \\ &= i'\Big(f'\big(g(y)\big)\Big) \circ (f' \circ g)'(y) \\ &= i'\Big(f'\big(g(y)\big)\Big) \circ f''\big(g(y)\big) \circ g'(y) \end{aligned}$$

Higher order derivatives are messier but still follow from combining (15), (17), and (14).

4.2.15 Transfer

If $f : \mathbb{R}^d \to \mathbb{R}^e$ is differentiable, then the derivative is the linear function represented by the matrix of partial derivatives (Browder, 1996, Theorem 8.21). The converse statement is false (Browder, 1996, Example 8.22), but if the partial derivatives are continuous functions, then f is continuously differentiable (Browder, 1996, Theorem 8.23). So, as long as we restrict our attention to continuously differentiable functions, there is no difference for $\mathbb{R}^d \to \mathbb{R}^e$ functions, between abstract differentiability and what we know from multivariable calculus.

Now suppose U and V are abstract finite-dimensional vector spaces of dimensions d and e, respectively, suppose O is open in U, and suppose

$f : O \to V$ is continuously differentiable. If we want to calculate using multivariable calculus, we need isomorphisms $g : U \to \mathbb{R}^d$ and $h : V \to \mathbb{R}^e$. Then the map $j : \mathbb{R}^d \to \mathbb{R}^e$ defined by $j = h \circ f \circ g^{-1}$ "represents" f in multivariable calculus

$$\begin{array}{ccc} \mathbb{R}^d & \xrightarrow{\;j\;} & \mathbb{R}^e \\ {\scriptstyle g}\uparrow & & \uparrow{\scriptstyle h} \\ U & \xrightarrow{\;f\;} & V \end{array}$$

Having gotten ahold of j, we can differentiate it by multivariable calculus (represented by the matrix of partial derivatives) and, since $f = h^{-1} \circ j \circ g$, the chain rule and the linear function rule gives the derivative

$$f'(x) = h^{-1} \circ j'\big(g(x)\big) \circ g$$

that is, letting $y = g(x)$, the diagram

$$\begin{array}{ccc} \mathbb{R}^d & \xrightarrow{\;j'(y)\;} & \mathbb{R}^e \\ {\scriptstyle g}\uparrow & & \uparrow{\scriptstyle h} \\ U & \xrightarrow{\;f'(x)\;} & V \end{array}$$

is commutative. This is "transfer" for differentiation.

4.3 Differentiation on Affine Spaces

4.3.1 Definition

Differential geometry extends differentiation from vector spaces to manifolds, which we need not define here. Affine spaces are a special case of manifolds. We can see what differential geometry says about affine spaces without knowing any differential geometry by just applying the definition of differentiation (Section 4.2.1 above) to functions between affine spaces.

We repeat equation (11) above

$$\lim_{h \to 0} \frac{f(x + h) - f(x) - g(h)}{\|h\|} = 0. \tag{11}$$

but now we allow f to be a function $O \to B$ where A and B are finite-dimensional affine spaces having translation spaces U and V, respectively, and O is an open subset of A. Now x is a point, so for $x + h$ to make sense h must be a vector, and $f(x + h) - f(x)$ is a vector, so g maps vectors to vectors and as before we can take it to be a linear function $U \to V$. If (11)

holds for some linear function g, then g is unique (by the same argument as for derivatives of functions between vector spaces given on p. 334 of Lang, 1993), and call it the derivative of f at the point x and write it $f'(x)$.

Hence if f is a differentiable function between finite-dimensional affine spaces, the derivative $f'(x)$ is a linear function between the corresponding translation spaces. If we look at what differential geometry says about differentiation, it agrees with what we have found here except for terminology. In differential geometry, one says *tangent space* rather than translation space.

Higher-order derivatives follow the same rule and are just derivatives of derivatives. If f above is everywhere differentiable on O, then $f' : O \to L(U, V)$ is a function from an open subset of a finite-dimensional affine space to a finite-dimensional vector space so we can apply the same definition to it, obtaining a derivative $f''(x)$ which is a linear function $U \to L(U, V)$ having type $U \to U \to V$.

Lemma 12. *For any point x in an affine space, let s_x denote the map $y \mapsto y - x$ from this affine space to its translation space. Let A and B be finite-dimensional affine spaces having translation spaces U and V respectively. Let O be open in A. Then a function $f : O \to B$ is differentiable at $x \in O$ if and only if the function $k : (O - a) \to V$ defined by the commutative diagram*

$$\begin{array}{ccc} O & \xrightarrow{f} & B \\ {\scriptstyle s_a}\downarrow & & \downarrow{\scriptstyle s_b} \\ O - a & \xrightarrow{k} & V \end{array}$$

is differentiable at $x - a$, in which case $f'(x) = k'(x - a)$.

Proof. With less category theory

$$k(u) = f(a + u) - b$$

so

$$k(u + h) - k(u) = f(a + u + h) - f(a + u).$$

To say that f is differentiable at x is to assert the existence of a linear function $g : U \to V$ such that (11) holds, in which case g is the derivative of f at x. In that case

$$k(x - a + h) - k(x - a) = f(x + h) - f(x)$$

so g is also the derivative of k at $x - a$.

Conversely, to say that g is the derivative of k at $x - a$ is to assert the existence of a linear function $g : U \to V$ such that

$$\lim_{h \to 0} \frac{k(x - a + h) - k(x - a) - g(h)}{\|h\|} = 0$$

in which case (11) holds and g is also the derivative of f at x. □

Either the lemma or the definition can be used to move all of our facts about differentiation of functions between vector spaces to the analogous facts about differentiation of functions between affine spaces. We will just state these without proof, since the proofs are easy.

4.3.2 The Constant Rule

The derivative of a constant function is zero. In more detail the derivative of a constant function from an open subset of a finite-dimensional affine space to an affine space is the only constant linear function (from the tangent space of one to the tangent space of the other), which is everywhere equal to zero.

4.3.3 The Other Constant Rule

The rule that $h = af$, where a is a scalar constant, implies $h'(x) = af'(x)$ only makes sense when the codomain of f is a vector space, because we only have multiplication by scalars in vector spaces not in affine spaces. So this rule makes sense when f goes from an open subset of a finite-dimensional affine space to a finite-dimensional vector space.

4.3.4 The Addition Rule

The rule that $h = f + g$ implies then $h'(x) = f'(x) + g'(x)$ only makes sense when the codomain of f and g is a vector space, because we only have addition of vectors not addition of points. So this rule makes sense when f and g go from an open subset of a finite-dimensional affine space to a finite-dimensional vector space.

4.3.5 The Affine Function Rule

If f is an affine function and g is its associated linear function (defined following Theorem 4), then $f'(x) = g$ for all x. Hence $f''(x) = 0$ for all x.

So now we can replace the term "associated linear function" by "derivative." Every nonempty affine function has a derivative at every point of

its domain, and that derivative does not depend on the point where it is evaluated. The derivative is always the associated linear function.

4.3.6 The Multiplication Rule

The multiplication rule in Section 4.2.11 above required that the multiplication operation be bilinear, which means it only makes sense when the things multiplied are vectors not points.

Let U, V, and W be finite-dimensional vector spaces, and let $U \times V \to W$ be a continuous bilinear function denoted by juxtaposition. Let E be a finite-dimensional affine space, let O be open in E, and let $f : O \to U$ and $g : O \to V$ be differentiable at $x \in O$. Then the product map $fg : O \to W$ is also differentiable at x and (12) holds, which is interpreted as meaning (13) modified by replacing E with its translation space.

4.3.7 The Chain Rule

Let A, B, and C be finite-dimensional affine spaces, having translation spaces U, V, and W, respectively. Let O be open in A and P be open in B, and let $f : O \to P$ and $g : P \to C$ be maps. Then the chain rule says that if f is differentiable at x and g is differentiable at $f(x)$, then the composition $h = g \circ f$ is differentiable at x and its derivative is given by (15) or (16).

4.3.8 The Inverse Function Theorem

Let A and B be finite-dimensional affine spaces, let O be an open subset of A and let $f : O \to B$ be p times continuously differentiable. If $f'(x)$ is an invertible linear function for some $x \in O$, then there exists an open set O' such that $x \in O' \subset O$ and the restriction of f to domain O' and codomain $f(O')$ is invertible and the inverse is p times continuously differentiable.

Moreover, the derivative of the inverse of this restriction at the point $f(x)$ is the inverse of $f'(x)$. The comments in Section 4.2.14 above about higher-order derivatives carry over to this setting. We do not need an inversion rule for functions between affine spaces here. The inversion rule of Section 4.2.12 above suffices, because we are inverting derivatives, which are linear functions.

We say such an f is a *local C^p isomorphism at x*.

4.4 Lebesgue Measure

Suppose A is a finite-dimensional affine space, V is its translation space, $f : \mathbb{R}^d \to A$ is an affine isomorphism, λ is Lebesgue measure on $\mathbb{R}^d$, and $\mu = \lambda \circ f^{-1}$ (notation defined in Section 4.1 above). Then μ is a translation-invariant measure meaning

$$\mu(B + v) = \mu(B), \qquad \text{for all } v \in V \text{ and all Borel subsets } B \text{ of } A.$$

Moreover, every nonempty open set has positive μ measure. So μ has most of the properties of Lebesgue measure.

The only important property that is lacking is uniqueness. Different isomorphisms f may induce different measures μ. Lebesgue measure on $\mathbb{R}^d$ assigns measure one to the unit cube. In an abstract affine space there is no notion of unit cube because there is no preferred basis for its translation space.

To see what is happening, consider another affine isomorphism $g : \mathbb{R}^d \to A$, and let $\nu = \lambda \circ g^{-1}$. What is the relationship between μ and ν?

Let $\rho = \nu \circ f = \lambda \circ g^{-1} \circ f$. Since ρ is a measure on $\mathbb{R}^d$ we know how to relate it to λ. By the change-of-variable theorem for integrals with respect to Lebesgue measure on $\mathbb{R}^d$ (the theorem about Jacobians) $d\rho$ is $d\lambda$ times the Jacobian determinant of the matrix representing the affine function $f^{-1} \circ g$. Since every affine function has a constant derivative, the Jacobian determinant is a scalar constant. Thus we see that ρ is just a constant times λ. And, since $\mu = \lambda \circ f^{-1}$ and $\nu = \rho \circ f^{-1}$, it follows that μ and ν are also constant multiples of each other.

Thus Lebesgue measure on a finite-dimensional affine space is uniquely defined up to multiplication by positive scalars. But that is as unique as it gets. We say each measure like μ and ν defined above is a *version* of Lebesgue measure, and we consider no version in any way special.

If we want to define densities with respect to Lebesgue measure, they have to be unnormalized densities (because no version of Lebesgue measure is special). We can say let h be an unnormalized probability density with respect to Lebesgue measure on A, and this means h is a nonnegative function such that $\int h \, d\mu$ is strictly positive and finite, where μ is any version of Lebesgue measure. When we want to evaluate an expectation, it has the form

$$E_h\{g(X)\} = \frac{\int g(x) h(x) \, \mu(dx)}{\int h(x) \, \mu(dx)}$$

where μ is any version of Lebesgue measure (the arbitrary normalization cancels so the expectation is unique).

Lemma 13. *Suppose X is a random element of a finite dimensional affine space A having unnormalized probability density function f_X with respect to Lebesgue measure on A. Suppose $g : A \to B$ is an affine isomorphism. Define $Y = g(X)$. Then*

$$f_X(x) = f_Y(y), \qquad \text{whenever } y = g(x)$$

defines an unnormalized probability density function f_Y of Y with respect to Lebesgue measure on B.

We can rewrite the equation in the statement of the lemma as

$$f_Y(y) = f_X\big(g^{-1}(y)\big), \qquad y \in B, \tag{18}$$

or even more simply as $f_Y = f_X \circ g^{-1}$ or as $f_X = f_Y \circ g$.

Proof. We have to apply the general change-of-variable theorem (Section 4.1 above). For any real-valued function h for which $E\{h(Y)\}$ exists

$$E\{h(Y)\} = \frac{\int h(g(x)) f_X(x)\, \lambda(dx)}{\int f_X(x)\, \lambda(dx)}$$

where λ is a version of Lebesgue measure on A. Note that the denominator of the fraction is the special case of the numerator when h is the constant function everywhere equal to one. Thus it is enough to deal with the numerator

$$\int h(g(x)) f_X(x)\, \lambda(dx) = \int h(y) f_X(g^{-1}(y))\, (\lambda \circ g^{-1})(dy)$$

(equation 9 above), but $\lambda \circ g^{-1}$ is a version of Lebesgue measure on B, so $f_X(g^{-1}(y))$ serves as an unnormalized probability density function of Y with respect to Lebesgue measure on B. □

A function is called a C^p *diffeomorphism* if it is invertible and both the function and its inverse are p-times continuously differentiable. A C^1 diffeomorphism is also called a diffeomorphism (without the C^1).

Theorem 14. *Suppose X is a random element of a finite dimensional affine space A having unnormalized probability density function h_X with respect to Lebesgue measure on A. Let O be an open subset of A that supports X, let B be a finite-dimensional affine space having the same dimension as A,*

and let $g : O \to B$ be any function such that the restriction $O \to g(O)$ is a diffeomorphism. Define $Y = g(X)$. Then

$$h_Y(y) = h_X(x)J(y), \qquad \text{whenever } y = g(x) \tag{19}$$

defines an unnormalized probability density function h_Y of Y with respect to Lebesgue measure on B, where J is defined as follows. Let $f_X : A \to \mathbb{R}^d$ and $f_Y : B \to \mathbb{R}^d$ be affine isomorphisms, and define a function k by the commutative diagram

$$\begin{array}{ccc} O & \xrightarrow{\ g\ } & B \\ {\scriptstyle f_X}\downarrow & & \downarrow{\scriptstyle f_Y} \\ f_X(O) & \xrightarrow{\ k\ } & \mathbb{R}^d \end{array}$$

then k is also a diffeomorphism. Define $J(y)$ to be the absolute value of the determinant of the Jacobian matrix of the map k^{-1} evaluated at the point $z = f_Y(y)$.

Proof. We know from the lemma that

$$\begin{aligned} h_W(w) &= h_X(x), \qquad \text{whenever } w = f_X(x) \\ h_Z(z) &= h_Y(y), \qquad \text{whenever } z = f_Y(y) \end{aligned}$$

define unnormalized probability density functions for W and Z with respect to Lebesgue measure on $\mathbb{R}^d$.

We know from the change-of-variable theorem from master's level probability theory that

$$h_Z(z) = h_W\big(k^{-1}(z)\big) \cdot J_{WZ}(z)$$

where $J_{WZ}(z)$ denotes the absolute value of the determinant of the Jacobian matrix of the function k^{-1} at the point z. Putting this all together, we get

$$h_Y(y) = h_Z(z) = h_W(w) \cdot J_{WZ}(z) = h_X(x) \cdot J_{WZ}(z),$$

whenever $w = f_X(x)$, $y = g(x)$, and $z = k(w) = f_Y(y)$. Thus we see we do have (19) with $J(y) = J_{WZ}(z)$. □

4.5 Duality

4.5.1 Definition

The dual space (Rudin, 1991, Section 3.1) of a topological vector space V is the set of all continuous linear functionals on V, where *linear functional*

means functions $V \to \mathbb{R}$. This set, denoted V^*, equipped with the operations of vector addition (1a) and scalar multiplication (1b) is itself a vector space. The only difference for finite-dimensionality is that every linear functional is continuous so there is no need for the word "continuous" in the definition.

4.5.2 Canonical Bilinear Form

In functional analysis a curious bracket notation (Halmos, 1974, Section 14; Rudin, 1991, Section 4.2) is used. We write $f(x)$ as $\langle x, f \rangle$ when $x \in V$ and $f \in V^*$. This is to emphasize that not only is $\langle \cdot, f \rangle$ a linear functional on V and hence an element of V^*, but also $\langle x, \cdot \rangle$ is a linear functional on V^* and hence an element of V^{**}, the dual space of V^*. The map $\langle \cdot, \cdot \rangle$ is called the *canonical bilinear form* that places V and V^* in duality.

4.5.3 Isomorphism

When V is finite-dimensional, V and V^* and V^{**} and as many more stars as you want all have the same dimension (Halmos, 1974, Theorem 2 of Section 15). They are all isomorphic to $\mathbb{R}^d$ for some d. Hence there is a tendency, if one has any nagging residual tendencies to think of finite-dimensional vector spaces as really being $\mathbb{R}^d$ for some d, of thinking of them as all the same space, but this is a mistake.

Here is a way to disabuse oneself of this mistake. If you think of V and V^* as both being $\mathbb{R}^d$, then you can think of $\langle x, f \rangle$ as meaning $x^T A f$, where A is any (fixed throughout the discussion) invertible $d \times d$ matrix. It is easy to check that every linear functional on $\mathbb{R}^d$ has the the form $x \mapsto x^T A f$ for some $f \in \mathbb{R}^d$ and, conversely for every $f \in \mathbb{R}^d$, the map $x \mapsto x^T A f$ is a linear functional. The default is to choose A to be the identify matrix so we get $\langle x, f \rangle = x^T f$, and this looks like the usual inner product on $\mathbb{R}^d$, but this default should be resisted. There is no reason for it other than computational convenience. The canonical bilinear form $\langle \cdot, \cdot \rangle$ is not an inner product at all. An inner product has both arguments in the same vector space. But (we pedantically reiterate) the arguments of the canonical bilinear form are in different spaces V and V^*.

4.5.4 Reflexivity

There is a special isomorphism $V \to V^{**}$ called the *natural isomorphism*, which identifies x in V with the linear functional $\langle x, \cdot \rangle$ in V^{**} (Halmos, 1974, Section 16).

So one can think of V and V^{**} as being "the same space" (in scare quotes, because, pedantically, they are not the same) via the natural isomorphism. Put another way, one is using a particular representation of V^{**} that represents every linear functional on V^* as $\langle x, \cdot \rangle$ for some $x \in V$.

We often use this representation, in effect making $V^{**} = V$, $V^{***} = V^*$, $V^{****} = V$, and so forth. So when we say we are using the representation $V^{**} = V$ we always mean that we are using the natural isomorphism to equate these spaces, and this also implies $V^{***} = V^*$, $V^{****} = V$, and so forth.

V and V^* are no more alike than they are like any other abstract vector spaces of the same dimension. There is no special isomorphism between them.

The section heading "reflexivity" comes from functional analysis. When V is an infinite-dimensional topological vector space there is still a map $V \to V^{**}$ that takes a vector x to the linear functional $\langle x, \cdot \rangle$ but this need not be an isomorphism. So we call this the natural injection $V \to V^{**}$. If the natural injection happens to be an isomorphism, then we say V is *reflexive*. Hence the section heading. Every finite-dimensional vector space is reflexive.

If V is a Hilbert space (a possibly infinite-dimensional complete inner product space) there is a special isomorphism between V and V^* (given by the Riesz representation theorem), but our vector and affine spaces have no inner product.

4.5.5 Adjoints

Suppose U and V are finite-dimensional abstract vector spaces, and suppose $f : U \to V$ is a linear function. Then there is a unique $f^* : V^* \to U^*$ satisfying

$$\langle f(x), y \rangle = \langle x, f^*(y) \rangle, \qquad y \in V^*, \ x \in U, \tag{20}$$

and f^* is called the *adjoint* of f (Halmos, 1974, Section 44). Note that the two canonical bilinear forms in (20) are different; on the left-hand side we have the canonical bilinear form placing V and V^* in duality, but on the right-hand side we have the canonical bilinear form placing U and U^* in duality.

Not using the bracket notation and remembering that elements of dual spaces are actually linear functionals as well as vectors makes the existence of the adjoint trivial, because (20) becomes

$$y\big(f(x)\big) = f^*(y)(x), \qquad y \in V^*, \ x \in U,$$

which says

$$f^*(y) = y \circ f. \tag{21}$$

If we specialize to $U = U^* = \mathbb{R}^d$ and $V = V^* = \mathbb{R}^e$, let the canonical bilinear forms be $\langle x, y \rangle = x^T y$, and confuse linear functions with the matrices representing them, then the adjoint is just the matrix transpose. If M is an $e \times d$ matrix, then the adjoint of the linear function $x \mapsto Mx$ is $y \mapsto M^T y$. But this depends on this particular choice of canonical bilinear forms. It is neither abstract nor general.

4.6 Points, Lines, Planes, and Hyperplanes

A *point* is a zero-dimensional affine space, a *line* is a one-dimensional affine space, a *plane* is a two-dimensional affine space, a *hyperplane* in a d-dimensional affine space A is an affine subspace having dimension $d - 1$.

If x is a point in an affine space and v is a nonzero vector in its translation space, then

$$\{ x + sv : s \in \mathbb{R} \}$$

is a line, and, conversely, all lines have this form.

If x and y are distinct points in an affine space, then

$$\{ x + s(y - x) : s \in \mathbb{R} \}$$

is a line. This is two points determine a line, one of Euclid's axioms.

Lemma 15. *If a subset V of a vector space U contains the origin and contains each line determined by any two of its points, then V is a vector subspace of U.*

Proof. If $s \in \mathbb{R}$ and $v \in V$, then $0 + sv = sv$ is on the line determined by 0 and v. Hence V is closed under scalar multiplication. If $v_1, v_2 \in V$, then $\frac{1}{2}v_1 + \frac{1}{2}v_2$ is on the line determined by v_1 and v_2 and $v_1 + v_2$ is a scalar multiple of that. Hence V is closed under vector addition. □

Theorem 16. *If a subset B of an affine space A contains each line determined by any two of its points, then B is an affine subspace of A.*

Proof. If B is empty or a singleton, then it is an affine subspace of A. Otherwise choose $x \in B$, and define $V = B - x$. For distinct points y_1 and y_2 in B and distinct vectors $v_i = y_i - x$ in V, the line determined by v_1 and v_2 is the set of vectors of the form

$$v_1 + s(v_2 - v_1) = y_1 - x + s(y_2 - y_1)$$

and hence is the image under the mapping $y \mapsto y - x$ of the line in B determined by y_1 and y_2. Hence, by the lemma, V is a vector subspace of the translation space of A. And $B = x + V$ is an affine subspace of A. □

Lemma 17. *If x is a point, v is a vector, s is a scalar, f is an affine function, then*

$$\begin{aligned} f(x + sv) &= f(x) + sf'(x)(v) \\ &= f(x) + s[f(x + v) - f(x)] \end{aligned} \tag{22}$$

Proof. Using $f'(x) = g$, the associated linear function, applying (4b) above gives $f(x + sv) = f(x) + g(sv)$, linearity of linear functions gives $g(sv) = sg(v)$ and (6) above gives $g(v) = f(x + v) - f(x)$. □

Theorem 18. *Every image and preimage of an affine subspace through an affine function is an affine subspace.*

With more symbols, suppose $f : A \to B$ is an affine function. If C is an affine subspace of A, then $f(C)$ is an affine subspace of B, If C is an affine subspace of B, then $f^{-1}(C)$ is an affine subspace of A,

Proof. Suppose C is an affine subspace of B. If $f^{-1}(C)$ is the empty set or a singleton set, then it is an affine subspace. Otherwise suppose $f^{-1}(C)$ contains distinct points x and y. Then by Lemma 17

$$f\big(x + s(y - x)\big) = f(x) + s[f(y) - f(x)] \tag{23}$$

Since C contains every line determined by any two of its points, the right-hand side of (23) is contained in C. Hence so is the left-hand side. Hence $x + s(y - x) \in f^{-1}(C)$.

Now suppose C is an affine subspace of A. If $f(C)$ is the empty set or a singleton set, then it is an affine subspace. Otherwise suppose $f(C)$ contains distinct points $f(x)$ and $f(y)$. We still have (23), which now shows that if C contains every line determined by two of its points, so does $f(C)$. □

Corollary 19. *Every image and preimage of a vector subspace through a linear function is a vector subspace.*

Proof. Apply the theorem and the fact that if f is a linear function, then $f(0) = 0$. □

For $f : A \to B$ and $C \subset A$, let $\tilde{f}$ denote the domain-codomain restriction of f that has domain C, codomain $f(C)$ and rule $x \mapsto f(x)$. (The rule remains the same, the domain and codomain are changed.)

Lemma 20. *The domain-codomain restriction $\tilde{f}$ described above, when $f : A \to B$ is a vector isomorphism and C is a vector subspace of A, is also a vector isomorphism that makes C isomorphic to $f(C)$.*

Proof. Because f is one-to-one, so is $\tilde{f}$. Because $\tilde{f}$ is onto by definition, it is thus invertible. So we only need to check that $\tilde{f}$ is linear. Because f and $\tilde{f}$ have the same rule, we have $\tilde{f}(x + y) = \tilde{f}(x) + \tilde{f}(y)$ and $\tilde{f}(ax) = a\tilde{f}(x)$. Since $f(C)$ is a vector space (Corollary 19 above), $\tilde{f}(x) + \tilde{f}(y)$ and $a\tilde{f}(x)$ are contained in $f(C)$. □

Lemma 21. *The domain-codomain restriction $\tilde{f}$ described above, when $f : A \to B$ is an affine isomorphism and C is an affine subspace of A, is also an affine isomorphism that makes C isomorphic to $f(C)$.*

Proof. Because f is one-to-one, so is $\tilde{f}$. Because $\tilde{f}$ is onto by definition, it is thus invertible. So we only need to check that $\tilde{f}$ is affine.

If C is empty then so is $f(C)$ and $\tilde{f}$ is the empty function, which is affine by definition.

Otherwise, let $x \in C$, let U be the translation space of C, and define $\tilde{g}$ by

$$\tilde{g}(u) = \tilde{f}(x + u) - \tilde{f}(x), \qquad u \in U.$$

If g is the associated linear function of f, then

$$\begin{aligned} \tilde{g}(u) &= \tilde{f}(x + u) - \tilde{f}(x) \\ &= f(x + u) - f(x) \\ &= g(u) \end{aligned}$$

So $\tilde{g}$ is the domain-codomain restriction of g and $g(U)$ is the translation space of $f(C)$. So $\tilde{g}$ is a linear function by Lemma 20 above. □

In particular, affine isomorphisms map empty affine spaces to empty affine spaces, points to points, lines to lines, planes to planes, and hyperplanes to hyperplanes.

Theorem 22. *If A is a finite-dimensional affine space and $f : A \to \mathbb{R}$ is a nonconstant affine function, then*

$$\{\, x \in A : f(x) = c \,\} \tag{24}$$

is a hyperplane, and, conversely, all hyperplanes have this form.

Proof. Let g denote the associated linear function of f, then g is not the zero function. By the rank plus nullity theorem (Halmos, 1974, Theorem 1 of Section 50), the rank of g is one, so the rank of the null space of g is $d-1$ where d is the rank of A and its translation space. We already know from Theorem 18 that (24) is an affine subspace. The null space of g is its own translation space, so (24) has dimension $d-1$.

Conversely, suppose H is a hyperplane in A. Let V be the translation space of A, let U be the translation space of H, and let x be a vector in $V \setminus U$. By the separating hyperplane theorem (Rudin, 1991, Theorem 3.4) there exists a linear functional g on V and a real number r such that $g(x) > r$ and $g(u) < r$ for $u \in U$. Since U is a vector subspace, it contains the origin. Hence $0 = g(0) < r$. Unless U is zero-dimensional, it contains a vector $u \neq 0$. Since U is closed under scalar multiplication we have $sg(u) = g(su) < r$ for all real numbers s. This implies $g(u) = 0$ for all $u \in U$. Furthermore, by the rank plus nullity theorem (cited above) the null space of g has dimension $d-1$ so must be a hyperplane containing H, so it is H. Thus $g(u) = 0$ if and only if $u \in U$.

Now let y be a point in H. Define $f : A \to \mathbb{R}$ by

$$f(x) = g(x - y) + c, \qquad x \in A.$$

Then H is given by (24). □

5 Convexity

5.1 Definitions

Do our notions of convexity and concavity have to change? The basic definitions for convexity and concavity are relative to lines (Rockafellar, 1970; Rockafellar and Wets, 1998). If the intersection of a set with each line is convex, then the set is convex. If the restriction of a function to each line is convex (resp. concave), then the function is convex (resp. concave). Rockafellar and Wets only give these definitions for $\mathbb{R}^n$. Others give them for abstract vector spaces. As far as I know, no one else gives them for abstract affine spaces. But, since we know what lines are in abstract affine spaces, the generalization seems obvious.

The only change in our habits needed in affine spaces is the triviality discussed in Section 3.2.3 above. If a and b are distinct vectors, the open line segment with endpoints a and b can be written

$$\{\, sa + (1 - s)b : 0 < s < 1 \,\}$$

but the formula $sa + (1 - s)b$ makes no sense when a and b are distinct points, because points cannot be multiplied by scalars, only vectors can be. As we saw in Section 4.6 above, the correct way to parameterize points on a line is $x + sv$, where x is a point, v is a vector, and s is a scalar, so the correct way to write the line segment with endpoints a and b in an affine space is

$$\{\, a + s(b - a) : 0 < s < 1 \,\}$$

So a set S in an abstract affine space is *convex* if

$$a + s(b - a) \in S, \qquad \text{whenever } a, b \in S \text{ and } 0 < s < 1. \tag{25}$$

Following Rockafellar (1970) and Rockafellar and Wets (1998) we consider convex functions that are allowed to have infinite values. So the codomain is the *extended real number system* $\overline{\mathbb{R}}$, which topologically is the two-point compactification of the real line. The order is the obvious one with $-\infty < x < +\infty$ for any real number x. The topology is the order topology. And the arithmetic is mostly obvious except it is not obvious how to define $\infty - \infty$ or $0 \cdot \infty$. Rockafellar and Wets (1998, Section 1.E) define both, although they adopt different definitions of $\infty - \infty$ in different contexts. We will adopt $0 \cdot \infty = 0$, which agrees with Rockafellar and Wets and which is widely used in probability theory, but will leave $\infty - \infty$ undefined. For details see Section 1.E of Rockafellar and Wets (1998).

One virtue of having $\overline{\mathbb{R}}$ as the codomain is that one can always have a whole affine space as the domain of convex functions (rather than a convex subset thereof; Rockafellar and Wets, 1998, Chapter 1), and this greatly simplifies theory, especially the theory of exponential families (Section 7.4 below).

An extended-real-valued function f on an abstract affine space A is *convex* if

$$\begin{aligned} &f(x + t[y - x]) \le tf(x) + (1 - t)f(y), \\ &\qquad \text{whenever } x, y \in A \text{ and } 0 < t < 1 \text{ and } f(x) < \infty \text{ and } f(y) < \infty. \end{aligned}$$

If all the conditions are satisfied, the right-hand side of the inequality is always well defined. If $f(x)$ and $f(y)$ are both finite, then the right-hand side is finite. If either $f(x)$ or $f(y)$ is infinite, then the only allowed value is $-\infty$, and the right-hand side is $-\infty$. We call the inequality in this definition the *convexity inequality*.

The set

$$\operatorname{dom} f = \{\, x \in V : f(x) < \infty \,\}$$

is called the *effective domain* of the convex function f (the domain is by definition all of the affine space on which f is defined).

A convex function is said to be *proper* if it nowhere takes the value $-\infty$ and does not everywhere have the value $+\infty$ (so the effective domain is nonempty).

Theorem 23. *The effective domain of a convex function is a convex set.*

Proof. Immediate from the convexity inequality. □

A proper convex function is said to be *strictly convex* if we have strict inequality in the convexity inequality whenever $x \neq y$ and $f(x)$ and $f(y)$ finite.

Theorem 24. *If f is a convex function and g is an affine function, and the domain of f is the codomain of g, then $f \circ g$ is a convex function.*

If f is an extended-real-valued function and g is an affine isomorphism, then f is a convex function if and only if $f \circ g$ is a convex function.

The latter also holds when "convex" is replaced by "strictly convex".

Proof. Write $h = f \circ g$. Assume f is convex. Apply Lemma 17 obtaining

$$\begin{aligned} h(x + s[y - x]) &= f(g(x + s[y - x])) \\ &= f(g(x) + s[g(y) - g(x)]) \\ &\leq sf(g(x)) + (1 - s)f(g(y)) \\ &= sh(x) + (1 - s)h(y) \end{aligned} \tag{26}$$

whenever $h(x) < \infty$ and $h(y) < \infty$ and $0 < s < 1$. So h is convex.

Now assume h is convex and g is an isomorphism. Now we know

$$h(x + s[y - x]) \leq sh(x) + (1 - s)h(y).$$

For any points u and v in the domain of f, which is the codomain of g, we can define $x = g^{-1}(u)$ and $y = g^{-1}(v)$, and we have

$$\begin{aligned} f(u + s[v - u]) &= f(g(x) + s[g(y) - g(x)]) \\ &= h(x + s[y - x]) \\ &\leq sh(x) + (1 - s)h(y) \\ &= sf(g(x)) + (1 - s)f(g(y)) \\ &= sf(u) + (1 - s)f(v) \end{aligned} \tag{27}$$

So f is convex.

For the assertion about strict convexity we repeat the proof above with g an affine isomorphism. If we assume f is strictly convex and $x \neq y$, then we also have $g(x) \neq g(y)$ and hence get strict inequality in (26), so h is also strictly convex. If we assume h is strictly convex and $u \neq v$, then we also have $x \neq y$ and hence get strict inequality in (27), so f is also strictly convex. □

We take the second and third sentences of the theorem statement to be how transfer works for convexity and strict convexity. Rockafellar (1970) and Rockafellar and Wets (1998) only work with convex functions on $\mathbb{R}^d$. But transfer says that f is convex if and only if $f \circ g$ is convex when g is an affine isomorphism from $\mathbb{R}^d$ to the abstract affine space that is the domain of f, and similarly for strictly convex.

5.2 Convexity and Optimization

Convex functions are useful in optimization because of the property that every local minimizer is a global minimizer, which follows directly from the convexity inequality. Ignore the trivial case where the objective function is everywhere equal to $+\infty$. Assume to get a contradiction that x is a local minimizer and there exists a point y such that $f(y) < f(x)$, then the convexity inequality says

$$f\big(x + t(y - x)\big) \leq tf(y) + (1 - t)f(x) < f(x), \qquad 0 < t < 1,$$

which contradicts x being a local minimizer.

Proper strictly convex functions are useful in optimization because of the property that they have at most one local minimizer (which if it exists is the unique global minimizer). This too follows immediately from the definitions. Assume to get a contradiction that x and y are distinct local minimizers. We already know (from convexity) that x and y must be global minimizers, so $f(x)$ and $f(y)$ are finite and equal. Then the strict convexity inequality says

$$f\big(x + t(y - x)\big) < tf(y) + (1 - t)f(x) = f(x), \qquad 0 < t < 1, \tag{28}$$

but this contradicts x and y being global minimizers.

5.3 Concavity

A function f is *concave* if $-f$ is convex. A function f is *strictly concave* if $-f$ is strictly convex. A concave function f is *proper* if $-f$ is proper. The *effective domain* of a concave function f is $\operatorname{dom}(-f)$.

Every local maximizer of a concave function is a global maximizer. A proper strictly concave function has at most one local maximizer (which if it exists is the unique global maximizer).

All the theory of concave functions follows from the theory of convex functions (just stand on your head so concave functions look convex).

5.4 Convexity and Derivatives

We say a real-valued function on an open convex set O of a finite-dimensional affine space A is *convex* if the convexity inequality holds with A replaced by O. It is *strictly convex* if the convexity inequality holds strictly with A replaced by O and $x \neq y$.

Theorem 25. *For a real-valued function f defined on an open convex set O in a finite-dimensional affine space, each of the following conditions is both necessary and sufficient for f to be convex on O, assuming the derivatives that appear in them exist.*

(a) $\langle y - x, f'(y) - f'(x) \rangle \geq 0$, *for all* $x, y \in O$.

(b) $f(y) \geq f(x) + \langle y - x, f'(x) \rangle$, *for all* $x, y \in O$.

(c) $f''(x)$ *is positive semidefinite for all* $x \in O$.

And each of the following conditions is both necessary and sufficient for f to be strictly convex on O, assuming the derivatives that appear in them exist.

(d) $\langle y - x, f'(y) - f'(x) \rangle > 0$, *for all* $x, y \in O$ *such that* $x \neq y$.

(e) $f(y) > f(x) + \langle y - x, f'(x) \rangle$, *for all* $x, y \in O$ *such that* $x \neq y$.

The following condition is sufficient (but not necessary) for f to be strictly convex on O, assuming the derivatives that appear in it exist.

(f) $f''(x)$ *is positive definite for all* $x \in O$.

Proof. This is Theorem 2.14 in Rockafellar and Wets (1998) except they state it for $\mathbb{R}^d$ rather than any finite-dimensional affine space and are considering $f'(x)$ as a vector and $f''(x)$ as a matrix.

So we need to use transfer. Let A be the finite-dimensional affine space containing O and let g be an invertible affine function $\mathbb{R}^d \to A$. Then $g^{-1}(O)$

is an open convex set in $\mathbb{R}^d$. Fix $x \in O$, and let $y = g^{-1}(x)$ and $h = f \circ g$. From the chain rule

$$\begin{aligned} h'(y) &= f'(x) \circ g'(y) \\ &= f'\big(g(y)\big) \circ g'(y) \\ &= (f' \circ g)(y) \circ g'(y) \end{aligned}$$

and from the affine function rule we know $g'(y)$ does not actually depend on y (it is the linear function associated with g). Let U be the translation space of A. Then $f'(x)$ is a linear function $U \to \mathbb{R}$. Hence so is $(f' \circ g)(y)$. The types match up

$$\begin{array}{ccc} \mathbb{R}^d & \xrightarrow{g'(y)} & U \\ & \searrow^{h'(y)} & \downarrow{\scriptstyle (f' \circ g)(y)} \\ & & \mathbb{R} \end{array}$$

It follows that $f' \circ g$ itself is a nonlinear function $\mathbb{R}^d \to L(U, \mathbb{R})$.

Composition of linear functions is bilinear (linear in both arguments), so by the multiplication rule

$$h''(y)(v) = (f' \circ g)'(y)(v) \circ g'(y), \qquad v \in \mathbb{R}^d$$

because $g''(y) = 0$.

This makes sense because h has type $\mathbb{R}^d \to \mathbb{R}$, so $h''(y)$ has type $\mathbb{R}^d \to \mathbb{R}^d \to \mathbb{R}$, and $h''(y)(v)$ has type $\mathbb{R}^d \to \mathbb{R}$. And we just learned that $f' \circ g$ maps $\mathbb{R}^d \to L(U, \mathbb{R})$. So its derivative $(f' \circ g)'(y)$ has the same type (except the former is nonlinear and the latter is linear), and $(f' \circ g)'(y)(v)$ is an element of $L(U, \mathbb{R})$. The types match up

$$\begin{array}{ccc} \mathbb{R}^d & \xrightarrow{g'(y)} & U \\ & \searrow^{h''(y)(v)} & \downarrow{\scriptstyle (f' \circ g)'(y)(v)} \\ & & \mathbb{R} \end{array}$$

By the chain rule

$$(f' \circ g)'(y) = f''\big(g(y)\big) \circ g'(y)$$

Here $f' \circ g$ maps $\mathbb{R}^d \to L(U, \mathbb{R})$. So its derivative $(f' \circ g)'(y)$ does the same (except is linear). Also f maps $O \to \mathbb{R}$, so its derivative $f'(x)$ is an element

of $L(U,\mathbb{R})$, so $f''(x)$ maps $U \to L(U,\mathbb{R})$. The types match up

$$\begin{array}{ccc} \mathbb{R}^d & \xrightarrow{g'(y)} & U \\ & \searrow^{(f'\circ g)'(y)} & \downarrow{\scriptstyle f''(x)} \\ & & L(U,\mathbb{R}) \end{array}$$

So if we evaluate $(f'\circ g)'(y)$ at a vector $v \in \mathbb{R}^d$, we get an element of $L(U,\mathbb{R})$

$$(f'\circ g)'(y)(v) = f''(x)\big(g'(y)(v)\big)$$

and

$$h''(y)(v) = f''(x)\big(g'(y)(v)\big)\circ g'(y)$$

and

$$h''(y)(v)(w) = f''(x)\big(g'(y)(v)\big)\big(g'(y)(w)\big)$$

Now let x_1 and x_2 be distinct elements of O and $y_i = g^{-1}(x_i)$. Then

$$\begin{aligned} h'(y_1) - h'(y_2) &= f'(x_1)\circ g'(y_1) - f'(x_2)\circ g'(y_2) \\ &= \big(f'(x_1) - f'(x_2)\big)\circ G \end{aligned}$$

where we have recalled that $g'(y)$ does not depend on y and is a linear function $\mathbb{R}^d \to U$ and written $g'(y) = G$. So

$$\begin{aligned} \langle y_1 - y_2, h'(y_1) - h'(y_2)\rangle &= \big\langle y_1 - y_2, \big(f'(x_1) - f'(x_2)\big)\circ G\big\rangle \\ &= \big(f'(x_1) - f'(x_2)\big)\big(G(y_1 - y_2)\big) \end{aligned}$$

Because G is the linear function associated with the affine function g, by Lemma 7 with commutative diagram

$$\begin{array}{ccc} \mathbb{R}^d & \xrightarrow{g} & A \\ {\scriptstyle s_{y_2}}\downarrow & & \downarrow{\scriptstyle s_{x_2}} \\ \mathbb{R}^d & \xrightarrow{G} & U \end{array}$$

we have $G(y_1 - y_2) = x_1 - x_2$. Hence

$$\langle y_1 - y_2, h'(y_1) - h'(y_2)\rangle = \langle x_1 - x_2, f'(x_1) - f'(x_2)\rangle$$

and this establishes the equivalence of statements (a) and (d) for our theorem and the theorem in Rockafellar and Wets.

Now

$$\begin{aligned} h(y_2) - h(y_1) - \langle y_2 - y_1, h'(y_1) \rangle &= f(x_2) - f(x_1) - h'(y_1)(y_2 - y_1) \\ &= f(x_2) - f(x_1) - \big(f'(x_1) \circ G\big)(y_2 - y_1) \\ &= f(x_2) - f(x_1) - f'(x_1)\big(G(y_2 - y_1)\big) \\ &= f(x_2) - f(x_1) - f'(x_1)(x_2 - x_1) \\ &= f(x_2) - f(x_1) - \langle x_2 - x_1, f'(x_1) \rangle \end{aligned}$$

and this establishes the equivalence of statements (b) and (e) for our theorem and the theorem in Rockafellar and Wets.

Now

$$h''(y)(v_1)(v_2) = f''(x)(G(v_1))(G(v_2))$$

Since G is invertible, it follows that $h''(y)$ is positive definite (resp. positive semi-definite) if and only if $f''(x)$ is, and this establishes the equivalence of statements (c) and (f) for our theorem and the theorem in Rockafellar and Wets. □

5.5 Fermat's Principle

We know from multivariable calculus what is called Fermat's principle: if f is a function $\mathbb{R}^d \to \mathbb{R}$ that is differentiable at x, then a necessary condition for x to be a local minimum or a local maximum is $f'(x) = 0$.

This is just as true for real-valued functions on abstract affine spaces. If f is a function $A \to \mathbb{R}$ that is differentiable at x, where A is an abstract finite-dimensional affine space, then a necessary condition for x to be a local minimum or a local maximum is $f'(x) = 0$.

This is obvious by transfer. Suppose $g : \mathbb{R}^d \to A$ is an affine isomorphism, and write $y = g^{-1}(x)$. Then f has a local minimum (resp. local maximum) at x if and only if $h = f \circ g$ has a local minimum (resp. local maximum) at y. By the chain rule

$$h'(y) = f'(x) \circ g$$

or

$$h'(y)(z) = f'(x)\big(g(z)\big).$$

By Fermat's principle for $\mathbb{R}^d$ we have $h'(y)(z) = 0$, for all $z \in \mathbb{R}^d$, and since g is an isomorphism, that implies $f'(x)(v) = 0$ for all $v \in V$.

But we get a much sharper condition from Theorem 25 (b).

Theorem 26. *If f is a differentiable real-valued convex (resp. concave) function defined on an open convex subset of a finite-dimensional affine*

space, then a necessary and sufficient condition for x to be a global minimizer (resp. maximizer) of f is $f'(x) = 0$.

5.6 Hulls

Lemma 27. *The intersection of a family of convex sets is convex. The intersection of a family of affine subspaces is an affine subspace. The intersection of a family of vector subspaces is a vector subspace.*

Proof. Suppose $\mathcal{C}$ is a family of convex subsets of the same affine space. Any points x and y in $\bigcap \mathcal{C}$ are contained in every element of $\mathcal{C}$. Hence the line segment (x, y) having end points x and y is contained in every element of $\mathcal{C}$. Hence (x, y) is contained in $\bigcap \mathcal{C}$.

A similar argument shows that if $\mathcal{A}$ is a family of affine subspaces of the same affine space, and x and y are distinct points in $\bigcap \mathcal{A}$, then the line determined by x and y is in $\bigcap \mathcal{A}$.

So if $\mathcal{V}$ is a family of vector subspaces of the same vector space we already know that $\bigcap \mathcal{V}$ is a affine subspace. But since $\bigcap \mathcal{V}$ must also contain the origin, it is a vector subspace. □

From the lemma we learn that for any set S there is a smallest convex set containing it (the intersection of all convex sets containing it) and similarly a smallest affine subspace containing it and a smallest vector subspace containing it. These are denoted $\operatorname{con}(S)$, $\operatorname{aff}(S)$, and $\operatorname{span}(S)$, respectively. Note that $\operatorname{con}(\varnothing)$ and $\operatorname{aff}(\varnothing)$ are empty, but $\operatorname{span}(\varnothing)$ is the zero subspace $\{0\}$.

5.7 Topology

The *relative interior* of a convex set C is the interior of C relative to $\operatorname{aff}(C)$, that is, the interior of C considered as a subset of the topological space $\operatorname{aff}(C)$. The relative interior of C is denoted $\operatorname{ri}(C)$.

Lemma 28. *Every image and preimage of a convex set through an affine function is a convex set.*

Proof. The proof is the same as the proof of Theorem 18 except for only considering line segments rather than lines. □

Theorem 29. *In a finite-dimensional affine space, the closure of a convex set is convex, the relative interior of a convex set is convex, and every nonempty convex set has a nonempty relative interior.*

Proof. For convex sets in $\mathbb{R}^d$, these assertions are found in Propositions 2.32 and 2.40 in Rockafellar and Wets (1998). They are transferred to any finite-dimensional affine space by affine isomorphism because closure and relative interior are topological affine space operations. □

6 Probability Theory

6.1 Random Vectors

In any topological vector space V, a random vector is a random object described by a Borel probability measure on V. Thus this concept also extends far beyond finite-dimensional vector spaces.

6.2 Ordinary Moments

In $\mathbb{R}^d$ the mean of a random vector X is the vector whose components are the means of the components of X. In an abstract vector space, vectors have no components (they can be given components by a choice of basis, but every different choice of basis leads to a different notion of components).

The good analog of components in $\mathbb{R}^d$ in an abstract vector space V is all of the $\langle X, \eta \rangle$ for all $\eta \in V^*$. This gives us an infinite number of "components" $\langle X, \eta \rangle$, but, of course, since V^* is finite-dimensional, this infinite number of "components" are determined by $\langle X, \eta_i \rangle$ where η_1, ..., η_d are a basis for V^*. This analogy also works the other way. In $\mathbb{R}^d$ the map $y \mapsto y_i$, where y_i denotes a component of y, is a linear functional (a vector-to-scalar linear function), hence an element of the dual space of $\mathbb{R}^d$, hence $\langle \, \cdot \, , \eta \rangle$, where η is another notation for the function $y \mapsto y_i$.

Having got the right analogy, it is fairly obvious how moments should be defined (and we get confirmation when we consider moment generating functions in Section 6.11 below).

If X is a random vector in an abstract finite-dimensional vector space V, then the first ordinary moment is the unilinear form $V^* \to \mathbb{R}$ defined by

$$\alpha_1(\eta) = E(\langle X, \eta \rangle), \qquad \eta \in V^*$$

provided all of the expectations exist. When $E(\langle X, \eta \rangle)$ does not exist for some η, we say the mean of X does not exist. "Unilinear form" is not a term we use often; it means the same thing as "linear functional" but goes with the terms we use for higher moments.

The second ordinary moment is the symmetric bilinear form $V^* \to V^* \to \mathbb{R}$ defined by

$$\alpha_2(\eta_1)(\eta_2) = E(\langle X, \eta_1 \rangle \langle X, \eta_2 \rangle), \qquad \eta_1, \eta_2 \in V^*$$

provided all of the expectations exist (otherwise, we say the second ordinary moment does not exist).

The third ordinary moment is the symmetric trilinear form $V^* \to V^* \to V^* \to \mathbb{R}$ defined by

$$\alpha_3(\eta_1)(\eta_2)(\eta_3) = E(\langle X, \eta_1 \rangle \langle X, \eta_2 \rangle \langle X, \eta_3 \rangle), \qquad \eta_1, \eta_2, \eta_3 \in V^*$$

provided all of the expectations exist (otherwise, we say the third ordinary moment does not exist).

And so on (you get the general idea, we hope, although we actually won't be interested in higher than second moments in this document).

The "ordinary" in "ordinary moment" is not a widespread usage. Your humble author uses it to contrast ordinary moments and central moments (Section 6.7 below). Most people say "moment" instead of "ordinary moment."

6.3 Mean

The first ordinary moment is also called the *mean*. So we, like everybody else, also use the notation μ for mean instead of α_1 for first ordinary moment.

As we said in the preceding section, if X is a random vector in V, then its mean μ is a unilinear form on V^*, which is the same thing as saying μ is a linear functional on V^*, which is the same thing as saying $\mu \in V^{**}$.

When V is finite-dimensional, we have the natural isomorphism that makes $V^{**} = V$, that allows us to consider μ an element of V, the same space where X lives.

6.4 Random Elements of Topological Affine Spaces

In any topological affine space A, a random element of A, also called a *random point*, is described by a Borel probability measure on A.

An empty affine space cannot support a probability measure because the measure of the empty set must be zero for any measure but the empty set is the whole space so must have probability one for any probability measure. Thus to say that X is a random point in an affine space A automatically implies that A is nonempty.

6.5 Supports

A topological space is *second countable* if there is a countable basis, which is a countable family $\mathcal{B}$ of open sets such that every open set is a union of elements of $\mathcal{B}$.

Lemma 30. *For any Borel probability measure P on a second countable topological space, there is a smallest closed set C such that $P(C) = 1$.*

The set C that the lemma asserts the existence of is called the *support* of P.

Proof. Let $\mathcal{B}$ be the countable basis, let

$$\mathcal{B}_0 = \{\, B \in \mathcal{B} : P(B) = 0 \,\},$$

and let $W = \bigcup \mathcal{B}_0$. Then by countable subadditivity, $P(W) = 0$. So if we let C be the complement of W, we have $P(C) = 1$.

Any open set U larger than W cannot have $P(U) = 0$, because otherwise U would be a countable union of elements of $\mathcal{B}_0$, and $U \subset W$ contrary to the assumption that U is larger. Hence W is the largest open set having probability zero, and C is the smallest closed set having probability one. □

Any finite-dimensional affine space A given the topology described in Section 3.4 above is a second-countable topological space. Let $f : \mathbb{R}^d \to A$ be an affine isomorphism. As is well known, $\mathbb{R}^d$ is second countable (the set of open balls having centers having rational coordinates and having rational radii is a countable basis). If $\mathcal{B}$ is a countable basis for $\mathbb{R}^d$, then

$$\{\, f(B) : B \in \mathcal{B} \,\}$$

is a countable basis for A. Similarly, for any finite-dimensional vector space.

Let cl denote the closure operator for topological spaces.

Theorem 31. *Suppose A is a finite-dimensional affine space and P is a Borel probability measure on A. Let C be the support of P. Then* $\operatorname{cl}(\operatorname{con}(C))$ *is the smallest closed convex set having P-probability one,* $\operatorname{aff}(C)$ *is the smallest affine subspace having P-probability one, and* $\operatorname{span}(C)$ *is the smallest vector subspace having P-probability one.*

Consequently, we say $\operatorname{cl}(\operatorname{con}(C))$ is the *convex support* of P, we say $\operatorname{aff}(C)$ is the *affine support* of P, and we say $\operatorname{span}(C)$ is the *vector support* of P.

Proof. The closure of a convex set is a convex set (Theorem 29 above). Hence $\operatorname{cl}(\operatorname{con}(C))$ is closed and convex and has P-probability one. If D is any other closed convex set such that $P(D) = 1$, then we must have $C \subset D$, hence $\operatorname{con}(C) \subset D$, hence $\operatorname{cl}(\operatorname{con}(C)) \subset D$.

By Corollary 3, $\operatorname{aff}(C)$ is closed and affine, so a similar argument works for it.

Also $\operatorname{span}(C)$ is a closed vector subspace, so a similar argument works for it. □

6.6 The Mean of a Random Element of an Affine Space

In general, ordinary moments make no sense for random elements of affine spaces because there is no dual space, double dual, etc. However, there ought to be a mean, and we define it as follows.

Theorem 32. *Suppose X is a random element of a finite-dimensional affine space A and $E\{f(X)\}$ exists for every affine function $f : A \to \mathbb{R}$. Then there exists a unique $\mu \in A$, such that*

$$E\{f(X)\} = f(\mu), \qquad \text{for every affine function } f : A \to \mathbb{R}.$$

Under the condition of the theorem we say that the μ asserted to exist by the theorem is the mean of X. Otherwise, we say the mean of X does not exist.

Proof. Let U be the translation space of A, and for any affine function $f : A \to \mathbb{R}$ let $\tilde{f} : U \to \mathbb{R}$ be the associated linear function. Fix $a \in A$. Then

$$\begin{aligned} f(X) &= f(a) + \tilde{f}(X - a) \\ &= f(a) + \langle X - a, \tilde{f} \rangle \end{aligned}$$

Hence

$$\begin{aligned} E\{f(X)\} &= f(a) + E\{\langle X - a, \tilde{f} \rangle\} \\ &= f(a) + \langle E\{X - a\}, \tilde{f} \rangle \\ &= f(a) + \langle \mu_{X-a}, \tilde{f} \rangle \end{aligned}$$

where μ_{X-a} is the mean of the random vector $X - a$, which exists because $E\{\langle X - a, \tilde{f} \rangle\} = E\{f(X)\} - f(a)$ and the right-hand side is always finite by assumption of the theorem.

Taking μ_{X-a} to be an element of U rather than U^{**} using the natural isomorphism, define $\mu = a + \mu_{X-a}$. Then

$$\begin{aligned} f(a) + \langle \mu_{X-a}, \tilde{f} \rangle &= f(a) + \tilde{f}(\mu_{X-a}) \\ &= f(a) + \tilde{f}(\mu - a) \\ &= f(\mu) \end{aligned}$$

□

As the proof shows, we already had a definition of the mean of a random vector as an element of the vector space where it lives, and this agrees with our new definition in this section.

We say a family of functions $\mathcal{F}$ from A to B *separates points* of A if whenever x and y are distinct points of A there exists an $f \in \mathcal{F}$ such that $f(x) \neq f(y)$.

Corollary 33. *The family of all real-valued affine functions on a finite-dimensional affine space A separates points of A.*

Proof. Consider the random variable X concentrated at the single point x in A. Then $E(X) = x$, and the theorem asserts that there is no other point y in A such that $E\{f(X)\} = f(y)$ for all real-valued affine functions f on A. Since we know that $E\{f(X)\} = f(x)$, this says it is not true that $f(x) = f(y)$ for all real-valued affine functions f on A. And this says the family real-valued affine functions on A separates points of A. □

Corollary 34. *Suppose A and B are finite-dimensional affine spaces, X is a random element of A having mean μ, and $f : A \to B$ is an affine function. Then $Y = f(X)$ has mean $f(\mu)$.*

Proof. For any affine function $g : B \to \mathbb{R}$, the function $h = g \circ f$ is affine so by definition $E\{h(X)\} = E\{g(Y)\}$ exists. So by the theorem there is a unique point $\nu \in B$ such that

$$E\{g(Y)\} = g(\nu)$$

for all affine functions $g : B \to \mathbb{R}$, and by definition $\nu = E(Y)$. We also have

$$g(\nu) = g(f(\mu))$$

for all affine functions $g : B \to \mathbb{R}$. So, by Corollary 33, $\nu = f(\mu)$. □

6.7 Central Moments of Random Vectors

If X is a random element of an abstract finite-dimensional vector space V having mean μ, then the first central moment of X is the unilinear form $V^* \to \mathbb{R}$ defined by

$$\mu_1(\eta) = E(\langle X - \mu, \eta \rangle), \qquad \eta \in V^*,$$

the second central moment is the symmetric bilinear form $V^* \to V^* \to \mathbb{R}$ defined by

$$\mu_2(\eta_1)(\eta_2) = E(\langle X - \mu, \eta_1 \rangle \langle X - \mu, \eta_2 \rangle), \qquad \eta_1, \eta_2 \in V^*, \tag{29}$$

the third central moment is the symmetric trilinear form $V^* \to V^* \to V^* \to \mathbb{R}$ defined by

$$\mu_3(\eta_1)(\eta_2)(\eta_3) = E(\langle X - \mu, \eta_1 \rangle \langle X - \mu, \eta_2 \rangle \langle X - \mu, \eta_3 \rangle), \qquad \eta_1, \eta_2, \eta_3 \in V^*,$$

and so forth. Again, these definitions assume that all expectations involved exist. When some expectation involved in the definition does not exist, then that central moment does not exist.

It is customary to use μ without subscripts for the mean and μ with subscripts for central moments, which can be confusing, but in this document we will have little possibility of confusion because we will only be interested in the second central moment and will give it a special name and notation.

The first central moment μ_1 is weird because, by linearity of expectation,

$$\mu_1(\eta) = E(\langle X - \mu, \eta \rangle) = E(\langle X, \eta \rangle) - \langle \mu, \eta \rangle = \langle \mu, \eta \rangle - \langle \mu, \eta \rangle = 0$$

so μ_1 is just another name for the zero unilinear form. We could have started the definitions with μ_2, but chose not to because it is sometimes nice to use the notation μ_1 to preserve symmetry of formulas (but we won't see that in this document).

It might appear at first sight that because of the appearance of $X - \mu$ in the formulas for central moments that this concept relies on the representation $V^{**} = V$ for the double dual (unless $\mu \in V$ we cannot do the subtraction $X - \mu$). But appearances are deceiving because

$$\langle X - \mu, \eta \rangle = \langle X, \eta \rangle - \langle \mu, \eta \rangle$$

and we can always consider the right-hand side to be well defined even if we consider μ to be an element of V^{**}. In that case, the two canonical bilinear forms on the right-hand side are different. In $\langle X, \eta \rangle$, it is the one placing V

and V^* in duality. In $\langle \mu, \eta \rangle$, which we should perhaps now write $\langle \eta, \mu \rangle$, it is the the one placing V^* and V^{**} in duality. Thus we could write

$$\mu_2(\eta_1)(\eta_2) = E([\langle X, \eta_1 \rangle - \langle \eta_1, \mu \rangle][\langle X, \eta_2 \rangle - \langle \eta_2, \mu \rangle])$$

and so forth to avoid reliance on $V^{**} = V$. This seems a little too pedantic for us, so we won't bother to fuss about this issue.

Also this technicality does not carry over to affine spaces. There we have only one definition of the mean: the mean of a random element of an affine space A is a point of A (Section 6.6). It cannot be an element of the double dual because an affine space has no double dual. Its translation space has a double dual, but there is no natural isomorphism between an affine space and the double dual of its translation space. If A is a finite-dimensional affine space having translation space V, there is of course the natural isomorphism $V \to V^{**}$ but there is no natural isomorphism $A \to V$. There are many isomorphisms, none of them more special than any other.

6.8 Central Moments of Random Elements of Affine Spaces

Having defined the mean μ of a random element X of a finite-dimensional affine space as in Section 6.6, central moments of X are by definition ordinary moments of the random vector $X - \mu$. So the theory for this is all already done.

If X is a random element of a finite-dimensional affine space A having translation space V, then the first central moment, if it exists, is the zero unilinear form on V^*, the second central moment, if it exists, is a symmetric bilinear form on V^*, the third central moment, if it exists, is a symmetric trilinear form on V^*, and so forth.

6.9 Variance

6.9.1 Generalities

The second central moment is also called the *variance*, and we also use the notation Σ for variance instead of μ_2 for second central moment.

As we said in the preceding section, if A is a finite-dimensional affine space having translation space V, and X is a random element of A, then its variance Σ is a symmetric bilinear form on V^*, which is the same thing as saying Σ has type $V^* \to V^* \to \mathbb{R}$, but every linear function $V^* \to \mathbb{R}$ is a linear functional on V^*, hence an element of V^{**}, so the type can also be written $V^* \to V^{**}$. When we are using the representation $V^{**} = V$, the type is also $V^* \to V$.

Thus we can consider variance to be either a symmetric bilinear form on V^* or a linear function $V^* \to V$. Considered as a bilinear form, it is

$$\Sigma(\eta_1)(\eta_2) = E\{\langle X - \mu, \eta_1 \rangle, \langle X - \mu, \eta_2 \rangle\}, \qquad \eta_1, \eta_2 \in V^*. \tag{30}$$

Considered as a linear function, it maps $\eta_1 \in V^*$ to the unique $x \in V$ such that $\langle x, \cdot \rangle$ is the linear function on V^* that is also denoted $\Sigma(\eta_1)$. If Σ denotes this linear function, then we can also write the bilinear form

$$(\eta_1, \eta_2) \mapsto \langle \Sigma(\eta_1), \eta_2 \rangle. \tag{31}$$

If we take V and V^* to both be $\mathbb{R}^d$ and the canonical bilinear form to be $\langle x, y \rangle = x^T y$, the usual conventions for probability theory on $\mathbb{R}^d$, then variance is usually defined to be a matrix M having components

$$m_{ij} = \operatorname{cov}(X_i, X_j) \tag{32}$$

In this case, the corresponding bilinear form is $(x, y) \mapsto x^T M y$, and the corresponding linear function is $x \mapsto Mx$. A square matrix can represent either a bilinear form or a linear function.

Many people do not like the term "variance matrix" for the matrix having components (32) because it involves covariances. Some call it the "covariance matrix" but that is really bad terminology, because what then do you call the covariance of two random vectors? Others call it the "variance-covariance matrix." Others call it the "dispersion matrix." But your humble author always uses "variance matrix" on the grounds that it is the vector analogue of the variance of a random scalar. This is seen in the formulas for the change of mean and variance under a linear transformation (Section 6.10 below), in the central limit theorem and the delta method (Sections 6.16 and 6.17 below), and in many other places. Hence we also call Σ defined by (30) the "variance" (considered as either a symmetric bilinear form or as a linear function).

6.9.2 Symmetric and Positive Semidefinite

A variance matrix is symmetric and positive semidefinite and every symmetric and positive semidefinite matrix is a variance matrix (there is, for example, a normal random vector with that variance matrix). Moreover, the variance matrix fails to be positive definite if and only if its random vector is concentrated on a hyperplane. Moreover, the variance matrix fails to be positive definite if and only if it is not invertible. What are the analogs of these properties in the abstract picture?

6.9.3 Symmetry and Adjoints

First we consider variance as a bilinear form, in which case symmetric refers to the bilinear form. We have

$$\Sigma(\eta)(\eta) = E\{\langle X - \mu, \eta\rangle^2\} \geq 0,$$

and this is the positive semi-definiteness property. A symmetric bilinear form $\Sigma : V^* \times V^* \to \mathbb{R}$ is *positive semidefinite* if

$$\Sigma(\eta)(\eta) \geq 0, \qquad \eta \in V^*. \tag{33}$$

It is *positive definite* if

$$\Sigma(\eta)(\eta) > 0, \qquad \eta \in V^*,\ \eta \neq 0. \tag{34}$$

Positive definiteness fails if there is an $\eta \neq 0$ such that

$$\Sigma(\eta)(\eta) = E\{\langle X - \mu, \eta\rangle^2\} = 0,$$

which happens if and only if $\langle X - \mu, \eta\rangle = 0$ almost surely, which is the same as saying that X is concentrated on the hyperplane

$$H = \{\, x \in A : \langle x - \mu, \eta\rangle = 0 \,\}. \tag{35}$$

When we think of Σ as a linear function $V^* \to V$, the corresponding bilinear form is (31), and symmetry of this bilinear form says

$$\langle \eta_1, \Sigma^*(\eta_2)\rangle = \langle \Sigma(\eta_1), \eta_2\rangle = \langle \Sigma(\eta_2), \eta_1\rangle \tag{36}$$

And this leads to the following.

Theorem 35. *If we are using the representation $V^{**} = V$, then a symmetric bilinear form, when considered as a linear operator, is its own adjoint.*

Proof. We need to be careful about (36) says. If we are fussy about which slot in $\langle \cdot, \cdot \rangle$ is in the dual space, says the dual space to the right, then (36) becomes

$$\underset{V \times V^*}{\langle \Sigma^*(\eta_2), \eta_1\rangle} = \underset{V^* \times V^{**}}{\langle \eta_1, \Sigma^*(\eta_2)\rangle} = \underset{V \times V^*}{\langle \Sigma(\eta_1), \eta_2\rangle} = \underset{V \times V^*}{\langle \Sigma(\eta_2), \eta_1\rangle}$$

where we have added another equality on the left that is not in (36), and reading from end to end gives $\Sigma^* = \Sigma$. But the first and second equalities just above are explicitly using $V^{**} = V$. □

If we don't use $V^{**} = V$, then $\Sigma : V^* \to V^{**}$ and $\Sigma^* : V^{***} \to V^{**}$ don't even have the same type.

6.9.4 Positive Definiteness and Invertibility

A one-to-one linear function going between finite-dimensional vector spaces is invertible only if it goes between vector spaces of the same dimension (Section 2.2 above). By definition, a function (linear or nonlinear) is invertible if and only if it is injective and surjective (one-to-one and onto). A linear function going between finite-dimensional vector spaces of the same dimension is injective if and only if it is surjective (Halmos, 1974, Theorem 1 of Section 50). Thus a necessary condition for a morphism in the category of finite-dimensional vector spaces to be iso is that its domain and codomain have the same dimension. And a sufficient condition is that it be injective. And another sufficient condition is that it be surjective.

Theorem 36. *A variance considered as a bilinear form is positive definite if and only if it is invertible considered as a linear function.*

Proof. If we consider a variance Σ as a bilinear form $V^* \to V^* \to \mathbb{R}$, then it fails to be positive definite if and only if there exists a nonzero η such that $\langle X - \mu, \eta \rangle = 0$ almost surely. But then

$$\Sigma(\eta)(\eta_2) = E\{\langle X - \mu, \eta \rangle, \langle X - \mu, \eta_2 \rangle\} = 0$$

for all η_2 which implies $\Sigma(\eta) = 0$ because linear functionals separate points (Corollary 33 above).

Conversely $\Sigma(\eta) = 0$ obviously implies $\Sigma(\eta)(\eta_2) = 0$. □

6.9.5 Comparison with the $\mathbb{R}^d$ Picture

If we treat V and V^* as both being $\mathbb{R}^d$ for some d and take the canonical bilinear form to be $\langle x, \eta \rangle = x^T A \eta$ for some (fixed throughout the discussion) symmetric positive definite matrix A, then the matrix corresponding to the linear function Σ will be a symmetric positive semidefinite matrix. But, as we said in Section 4.5.3 above, any invertible matrix A will do here, not necessarily either symmetric or positive definite. And then the matrix corresponding to Σ need not be symmetric or positive semidefinite. Of course, for practical calculations, one wouldn't do that because it would be confusing. But one could do that, which shows that (beating a dead horse here) abstract vector spaces aren't really $\mathbb{R}^d$ for some d, and linear functions between them aren't really matrices, and there is no unique way to associate matrices with them.

6.10 Change of Mean and Variance under Affine Functions

Theorem 37. *Suppose A and B are finite-dimensional abstract affine spaces and f is an affine function $A \to B$. Suppose X is a random element of A and $Y = f(X)$. Let μ_X and μ_Y denote the means of X and Y, and let Σ_X and Σ_Y denote the variances of X and Y. Then*

$$\mu_Y = f(\mu_X) \tag{37}$$

and when variances are considered as bilinear forms

$$\Sigma_Y(\eta_1)(\eta_2) = \Sigma_X\big(g^*(\eta_1)\big)\big(g^*(\eta_2)\big) \tag{38}$$

where g is the linear function associated with f and when variances are considered as linear functions

$$\Sigma_Y = g \circ \Sigma_X \circ g^* \tag{39}$$

The commutative diagram for (39) is

$$\begin{array}{ccccccc} V^* & \xrightarrow{g^*} & U^* & \xrightarrow{\Sigma_X} & U & \xrightarrow{g} & V \\ & \searrow & & & & \nearrow & \\ & & & \Sigma_Y & & & \end{array} \tag{40}$$

where U is the translation space of A where V is the translation space of B.

Using the equivalence of "associated linear function" and "derivative," we can also rewrite (39) as

$$\Sigma_Y = f'(x) \circ \Sigma_X \circ f'(x)^* \tag{41}$$

with the implicit understanding that $f'(x)$ does not depend on x.

Proof. Equation (37) follows from Theorem 32.

Equation (38) is derived as follows.

$$\begin{aligned} \Sigma_Y(\eta_1)(\eta_2) &= E\{\langle Y - \mu_Y, \eta_1\rangle\langle Y - \mu_Y, \eta_2\rangle\} \\ &= E\{\langle f(X) - f(\mu_X), \eta_1\rangle, \langle f(X) - f(\mu_X), \eta_2\rangle\} \\ &= E\{\langle g(X - \mu_X), \eta_1\rangle, \langle g(X - \mu_X), \eta_2\rangle\} \\ &= E\{\langle X - \mu_X, g^*(\eta_1)\rangle\langle X - \mu_X, g^*(\eta_2)\rangle\} \\ &= \Sigma_X\big(g^*(\eta_1)\big)\big(g^*(\eta_2)\big) \end{aligned}$$

Then (39) is derived as follows.

$$\begin{aligned}\bigl\langle \Sigma_Y(\eta_1), \eta_2 \bigr\rangle &= \bigl\langle \Sigma_X\bigl(g^*(\eta_1)\bigr), g^*(\eta_2) \bigr\rangle \\ &= \Bigl\langle g\Bigl(\Sigma_X\bigl(g^*(\eta_1)\bigr)\Bigr), \eta_2 \Bigr\rangle\end{aligned}$$

That this holds for all η_2 means

$$\Sigma_Y(\eta_1) = g\Bigl(\Sigma_X\bigl(g^*(\eta_1)\bigr)\Bigr) = (g \circ \Sigma_X \circ g^*)(\eta_1)$$

by Corollary 33, and this holding for all η_1 gives (39). □

The formulas in this section generalize the analogous formulas familiar from undergraduate probability theory. If $Y = a + BX$ where X and Y are random vectors, a is a constant vector, and B is a constant matrix, then

$$\begin{aligned}E(Y) &= a + BE(X) \\ \operatorname{var}(Y) &= B \operatorname{var}(X) B^T\end{aligned}$$

6.11 Moment Generating Functions

For an random vector X in an abstract finite-dimensional vector space V define an extended-real-valued function M on V^* defined by

$$M(\eta) = E\bigl\{e^{\langle X, \eta \rangle}\bigr\}, \qquad \eta \in V^*$$

where we take this to mean $M(\eta) = \infty$ at η such that the expectation does not exist. If M is finite on a neighborhood of zero, then we say M is the *moment generating function* (MGF) of X. Otherwise we say that X does not have an MGF.

The reason for the name is that derivatives of M evaluated at zero are the ordinary moments of X, that is,

$$\begin{aligned}M'(0) &= \alpha_1 \\ M''(0) &= \alpha_2 \\ M'''(0) &= \alpha_3\end{aligned}$$

and so forth (in the first line both sides are unilinear forms on V^*, in the second line both sides are symmetric bilinear forms on V^*, and so forth). If M is an MGF, then it is infinitely differentiable, and X has moments of all orders.

All of this is provable by transfer. Let X be a random vector in an abstract finite-dimensional vector space V, let $f : \mathbb{R}^d \to V$ be a vector space isomorphism, and let $Y = f^{-1}(X)$. Letting M_X and M_Y denote the MGF's of X and Y, if they exist, we have

$$\begin{aligned} M_X(\eta) &= E\big\{e^{\langle f(Y),\eta\rangle}\big\} \\ &= E\big\{e^{\langle Y, f^*(\eta)\rangle}\big\} \\ &= M_Y\big(f^*(\eta)\big) \end{aligned}$$

so $M_X = M_Y \circ f^*$. Hence X has an MGF if and only if Y does, and if they do, the derivatives of M_X are given by

$$\begin{aligned} M_X'(\eta)(\zeta) &= M_Y'\big(f^*(\eta)\big)\big(f^*(\zeta)\big) \\ M_X''(\eta)(\zeta_1)(\zeta_2) &= M_Y''\big(f^*(\eta)\big)\big(f^*(\zeta_1)\big)\big(f^*(\zeta_2)\big) \end{aligned}$$

and so forth. So

$$\begin{aligned} M_X'(0)(\zeta) &= M_Y'(0)\big(f^*(\zeta)\big) \\ &= E\big\{\langle Y, f^*(\zeta)\rangle\big\} \\ &= E\big\{\langle f(Y), \zeta\rangle\big\} \\ &= E\big\{\langle X, \zeta\rangle\big\} \\ M_X''(0)(\zeta_1)(\zeta_2) &= M_Y''(0)\big(f^*(\zeta_1)\big)\big(f^*(\zeta_2)\big) \\ &= E\big\{\langle Y, f^*(\zeta_1)\rangle\langle Y, f^*(\zeta_2)\rangle\big\} \\ &= E\big\{\langle X, \zeta_1\rangle\langle X, \zeta_2\rangle\big\} \end{aligned}$$

and so forth.

Since MGF involve ordinary moments, which do not make sense on affine spaces, MGF are inherently a vector space tool. Of course, if X is a random element of an affine space and a is a nonrandom point in that space, then $X - a$ is a random vector (in the translation space of that affine space) so $X - a$ may have an MGF, which, if it exists, can be used to calculate moments of $X - a$. And if we choose a to be the mean of X, these will be the central moments of X.

6.12 Cumulant Generating Functions

The log of the MGF of X, if it exists, is called the *cumulant generating function* (CGF) of X. Derivatives of the CGF evaluated at zero are called the *cumulants* of X.

Cumulants of order m are polynomial functions of the ordinary moments up to order m and vice versa (Cramér, 1951, Section 15.10). Here we will only be interested in the first two cumulants, which are the mean and the variance. If M is the MGF and K is the CGF, then

$$\begin{aligned} K'(\eta)(\zeta) &= \frac{M'(\eta)(\zeta)}{M(\eta)} \\ K''(\eta)(\zeta_1)(\zeta_2) &= \frac{M''(\eta)(\zeta_1)(\zeta_2)}{M(\eta)} - \frac{M'(\eta)(\zeta_1)}{M(\eta)} \cdot \frac{M'(\eta)(\zeta_2)}{M(\eta)} \end{aligned}$$

so, since $M(0) = 1$,

$$\begin{aligned} K'(0)(\zeta) &= M'(0)(\zeta) \\ K''(0)(\zeta_1)(\zeta_2) &= M''(0)(\zeta_1)(\zeta_2) - M'(0)(\zeta_1) \cdot M'(0)(\zeta_2) \end{aligned}$$

and this together with

$$\mu_2(\zeta_1)(\zeta_2) = \alpha_2(\zeta_1)(\zeta_2) - \alpha_1(\zeta_1) \cdot \alpha_1(\zeta_2),$$

which is the vector analog of $\operatorname{var}(X) = E(X^2) - E(X)^2$ and is proved the same way, by linearity of expectation, finishes the proof that the first two cumulants are mean and variance.

6.13 Law of Large Numbers

Theorem 38. *If X_1, X_2, ... is a sequence of IID random elements of an abstract finite-dimensional affine space having mean μ, and*

$$\overline{X}_n = \mu + \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) \tag{42}$$

then

$$\overline{X}_n \to \mu, \qquad \text{almost surely, as } n \to \infty.$$

The only surprise is that we define $\overline{X}_n$ differently than we would in a vector space as explained in Section 3.2.3 above. Of course, we could also define

$$\overline{X}_n = a + \frac{1}{n} \sum_{i=1}^{n} (X_i - a) \tag{43}$$

for any point a (Lemma 1 above).

Proof. We prove by transfer from the standard LLN for $\mathbb{R}^d$. Let A denote the affine space where X_1, X_2, ..., and μ live, let $f : A \to \mathbb{R}^d$ be an affine isomorphism, and define $Y_i = f(X_i)$ for all i, then we know Y_1, Y_2, ... is a sequence of IID random vectors having mean $f(\mu)$, and

$$\overline{Y}_n \to f(\mu), \qquad \text{almost surely, as } n \to \infty,$$

where

$$\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{44}$$

by the standard LLN. Let g be the linear function associated with f. Then

$$\overline{Y}_n - f(\mu) = \frac{1}{n} \sum_{i=1}^{n} \bigl[f(X_i) - f(\mu) \bigr] = \frac{1}{n} \sum_{i=1}^{n} g(X_i - \mu) = g\left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) \right)$$

by linearity. Hence

$$\overline{X}_n - \mu = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) = g^{-1}\left(\overline{Y}_n - f(\mu) \right)$$

converges almost surely to zero because g^{-1} is continuous. □

6.14 Normal Distributions

The *standard normal* distribution on $\mathbb{R}$ is the continuous distribution having unnormalized probability density with respect to Lebesgue measure $z \mapsto \exp(-z^2/2)$.

The *standard normal* distribution on $\mathbb{R}^d$ is the continuous distribution of a random vector having IID standard normal components. It has unnormalized density with respect to Lebesgue measure $z \mapsto \exp(-z^T z/2)$.

A general *normal distribution* on an abstract finite-dimensional affine space A is the image of a standard normal distribution on $\mathbb{R}^d$ for some d under an affine function $\mathbb{R}^d \to A$

Like general normal distributions on $\mathbb{R}^d$, general normal distributions on an abstract finite-dimensional affine space come in two kinds. *Degenerate* normal distributions are concentrated on hyperplanes. They cannot have unnormalized densities with respect to Lebesgue measure. They are the ones whose variances, considered as bilinear forms, are not positive definite (Section 6.9 above). *Nondegenerate* normal distributions are not concentrated on any proper affine subspace. They give positive probability to any open set and do have unnormalized densities with respect to Lebesgue measure.

We know the unnormalized density of a general normal distribution on $\mathbb{R}^d$ having mean μ and variance Σ, written in matrix notation, is

$$x \mapsto \exp\bigl(-(x-\mu)^T \Sigma^{-1} (x-\mu)/2\bigr).$$

Our notation is

$$(x-\mu)^T \Sigma^{-1} (x-\mu)$$

when we think of x, μ, and Σ as matrices. Our notation is

$$\langle x-\mu, \Sigma^{-1}(x-\mu)\rangle$$

when we think of x and μ as points, so $x-\mu$ is a vector, and Σ as a linear function whose inverse linear function is Σ^{-1}. Our notation is

$$\Sigma^{-1}(x-\mu)(x-\mu)$$

when we think of x and μ as points and Σ^{-1} as the bilinear form that corresponds to Σ^{-1} thought of as a linear function.

Now suppose X is a nondegenerate normal random element of an abstract finite-dimensional affine space A; suppose $f : \mathbb{R}^d \to A$ is an affine isomorphism, and let g denote the linear function associated with f defined by (4a). Then $Y = f^{-1}(X)$ is a nondegenerate normal random vector, and the variances of X and Y are related by (39). By $X = f(Y)$ and the change of variable theorem in Section 4.4 above we get that the unnormalized density of X is given by

$$h_X(x) = h_Y(f(x)) = \exp\bigl(-\langle f(x) - \mu_Y, \Sigma_Y^{-1}(f(x)-\mu_Y)\rangle/2\bigr)$$

and

$$\begin{aligned}
\langle f(x) - \mu_Y, \Sigma_Y^{-1}(f(x)-\mu_Y)\rangle &= \langle f(x) - f(\mu_X), \Sigma_Y^{-1}[f(x) - f(\mu_Y)]\rangle \\
&= \langle g(x-\mu_X), \Sigma_Y^{-1} g(x-\mu_Y)\rangle \\
&= \langle x-\mu_X, (g^* \circ \Sigma_Y^{-1} \circ g)(x-\mu_Y)\rangle
\end{aligned}$$

the second equality being linearity of g, and

$$\begin{aligned}
(g^* \circ \Sigma_Y^{-1} \circ g)^{-1} &= g^{-1} \circ \Sigma_Y \circ (g^*)^{-1} \\
&= g^{-1} \circ \Sigma_Y \circ (g^{-1})^* \\
&= \Sigma_X
\end{aligned}$$

the second equality being the inverse of an adjoint is the adjoint of the inverse (Halmos, 1974, Section 44) and (39). Hence

$$h_X(x) = \exp\bigl(-\Sigma_X^{-1}(x - \mu_X)(x - \mu_X)/2\bigr) \tag{45}$$

gives an unnormalized probability density with respect to Lebesgue measure on an abstract finite-dimensional affine space. (This is transfer for densities with respect to Lebesgue measure.) Densities have the same form on abstract finite-dimensional affine spaces as on $\mathbb{R}^d$. We just have to write them using the correct notation.

Now suppose X is a degenerate normal random element of an abstract finite-dimensional affine space A. Let B be the affine support of X. Let Y be the restriction of X to B. Then Y is a nondegenerate normal random element of B, and the preceding discussion says (45) is an unnormalized probability density of Y with respect to Lebesgue measure on B.

Let $i : B \to A$ denote the natural injection $x \mapsto x$. Then $X = i(Y)$. This i is an affine function (the proof is similar to the proof that the identity function is an affine function; Section 3.8 above). Let j be the linear function associated with i. Let U and V be the translation spaces of A and B, respectively. Then j is the natural injection $V \to U$ given by $x \mapsto x$.

From Section 6.10 above,

$$\mu_X = i(\mu_Y) = \mu_Y$$

and, when variances are interpreted as linear functions,

$$\Sigma_X = j \circ \Sigma_Y \circ j^*$$

and, when variances are interpreted as bilinear forms,

$$\Sigma_X(\eta_1)(\eta_2) = \Sigma_Y(j^*(\eta_1))(j^*(\eta_2)) = \Sigma_Y(\eta_1 \circ j)(\eta_2 \circ j)$$

(if η is a linear functional on U, then $j^*(\eta)$ is the linear functional on V that is the restriction of η to V). This says μ_X and μ_Y are the same, and Σ_X and Σ_Y are almost the same. But Σ_X is not invertible. Only Σ_Y is invertible and gives an unnormalized probability density.

6.15 Convergence in Distribution

If X_1, X_2, ... is a sequence of random elements of a finite-dimensional affine space A and X is another random element of A, then we say the sequence *converges in distribution* to X if

$$E\{f(X_n)\} \to E\{f(X)\}$$

for every bounded continuous function $f : A \to \mathbb{R}$, and to indicate this we write

$$X_n \xrightarrow{\mathcal{D}} X.$$

Convergence in distribution is also called *convergence in law* and *weak convergence*.

Some readers may not recognize this definition, being instead familiar with a definition involving convergence of distribution functions (Ferguson, 1996, Definition 1 of Chapter 1). However, our definition is equivalent to that one by a result known as the Helly-Bray theorem (Ferguson, 1996, Theorem 3 of Chapter 3). The definition adopted here is more general being the one always used in general complete separable metric spaces (Billingsley, 1999, Chapter 1).

One might wonder why we are using this definition for abstract finite-dimensional affine spaces, which don't have a metric. But they can be given a metric. There is just no unique way to do so. But the definition of convergence in distribution only depends on the topology (which is unique, Section 3.4 above), because the topology determines which functions are continuous (those for which inverse images of open sets are open). So the definition of convergence in distribution only depends on the topology not on the metric.

Our basic tool for working with convergence in distribution is the continuous mapping theorem (Billingsley, 1999, Theorem 2.7), which says that if $X_n \xrightarrow{\mathcal{D}} X$ and g is a continuous function, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$. (Actually the theorem says more than that. It is enough for g to be continuous on a set having probability one under the distribution of X. But we will not need this refinement.)

6.16 The Central Limit Theorem

Theorem 39. *Suppose X_1, X_2, ... is a sequence of IID random elements of a finite-dimensional affine space having mean μ and variance Σ. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \text{Normal}(0, \Sigma), \tag{46}$$

where $\overline{X}_n$ is defined by (42) *or* (43).

Proof. This is proved by transfer, which is just like transfer for the LLN *mutatis mutandis*.

Let A denote the affine space where X_1, X_2, ..., and μ live, let $f : A \to \mathbb{R}^d$ be an affine isomorphism, let g denote the linear function associated with f defined by (4a), and define $Y_i = f(X_i)$ for all i.

Then Y_1, Y_2, ... is a sequence of IID random vectors having mean $\mu_Y = f(\mu)$ and variance Σ_Y given by (39). Define $\overline{Y}_n$ in the usual way (44).

The CLT for $\mathbb{R}^d$ tells us

$$\sqrt{n}(\overline{Y}_n - \mu_Y) \xrightarrow{\mathcal{D}} Z,$$

where Z has a normal distribution with mean zero and variance Σ_Y. By linearity of g

$$\begin{aligned}
\sqrt{n}(\overline{Y}_n - \mu_Y) &= \sqrt{n} \left(\left[\frac{1}{n} \sum_{i=1}^n f(X_i) \right] - f(\mu) \right) \\
&= \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n [f(X_i) - f(\mu)] \right) \\
&= \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n g(X_i - \mu) \right) \\
&= g \left(\sqrt{n} \left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right] \right) \\
&= g \left(\sqrt{n} \left[\overline{X}_n - \mu_X \right] \right)
\end{aligned}$$

so

$$\sqrt{n}(\overline{X}_n - \mu_X) = g^{-1}\big(\sqrt{n}(\overline{Y}_n - \mu_Y)\big) \xrightarrow{\mathcal{D}} g^{-1}(Z)$$

by the continuous mapping theorem. And $g^{-1}(Z)$ is a linear function of multivariate normal, hence multivariate normal. Its mean is $g^{-1}(0) = 0$. and its variance is given by (39)

$$\begin{aligned}
g^{-1} \circ \Sigma_Y \circ (g^{-1})^* &= g^{-1} \circ g \circ \Sigma_X \circ g^* \circ (g^{-1})^* \\
&= g^{-1} \circ g \circ \Sigma_X \circ g^* \circ (g^*)^{-1} \\
&= \Sigma_X
\end{aligned}$$

□

6.17 The Delta Method

Theorem 40. *Let X_1, X_2, ... be a sequence of random elements of a finite-dimensional affine space A, let ζ be a nonrandom element of A, and let a_n be a sequence of real numbers going to infinity. Assume*

$$a_n(X_n - \xi) \xrightarrow{\mathcal{D}} Y \tag{47}$$

Suppose g is a function from an open subset O of A to another finite-dimensional affine space B, and suppose g is differentiable at ξ. Then

$$a_n\big(g(X_n) - g(\xi)\big) \xrightarrow{\mathcal{D}} g'(\xi)(Y). \tag{48}$$

It is part of the assertion of this theorem that when X_n is not concentrated on O, then $g(X_n)$ can be defined to be an arbitrary element of B when $X_n \notin O$. This does not affect the limit because $\Pr(X_n \in O) \to 1$ as $n \to \infty$.

Necessarily Y is a random element of the translation space of A, and the right-hand side of (48) is a random element of the translation space of B.

Proof. Let $e : \mathbb{R}^d \to A$ be an affine isomorphism, let f be its associated linear function, which is a linear isomorphism, let $W_n = e^{-1}(X_n)$, let $\omega = e^{-1}(\xi)$, and let $Z = f^{-1}(Y)$. Then, by assumption (47),

$$a_n\big(e(W_n) - e(\omega)\big) = a_n(X_n - \xi) \xrightarrow{\mathcal{D}} Y = f(Z).$$

Write $h = g \circ e$. Then by the $\mathbb{R}^d$ case of the delta method (Ferguson, 1996, Theorem 7 of Chapter 7) we have

$$a_n\big(g(X_n) - g(\xi)\big) = a_n\big(h(W_n) - h(\omega)\big) \xrightarrow{\mathcal{D}} h'(\omega)(Z)$$

And by the chain rule and the affine function rule

$$h'(\omega) = g'(\xi) \circ e'(\omega) = g'(\xi) \circ f,$$

so

$$h'(\omega)(Z) = g'(\xi)\big(f(Z)\big) = g'(\xi)\big(Y\big)$$

□

7 Exponential Families: The Vector Picture

7.1 Definitions

We are now ready for what Geyer (1990) calls the "vector picture" of exponential families. We will mix this with the definition of exponential families from Geyer (2009) which avoids the distinction between "standard" exponential families and "general" exponential families (see Geyer, 1990, Sections 1.2 and 1.3, for what the distinction is).

An *exponential family of distributions* is a statistical model (family of probability distributions) having log likelihood of the form

$$l(\theta) = \langle y, \theta \rangle - c(\theta) \tag{49}$$

where y is a vector-valued statistic taking values in some abstract finite-dimensional vector space V, where θ is a vector-valued parameter taking values in V^*, and where any additive terms not involving the parameter have been dropped from the log likelihood, such terms having no effect on either likelihood inference or Bayesian inference.

Either a change-of-variable or change-of-parameter (or both) may be necessary to get the log likelihood into the form (49). To recognize this the special statistic y and the special parameter θ that occur in (49) are called the *canonical statistic* and *canonical parameter* (also called the natural statistic and parameter). The function c appearing in (49) is called the *cumulant function* of the family.

The canonical statistic, canonical parameter, and cumulant function are not unique. Any one-to-one affine function of a canonical statistic is another canonical statistic, any one-to-one affine function of a canonical parameter is another canonical parameter, and any cumulant function plus any (real-valued) affine function is another cumulant function. These alterations are not algebraically independent. Changing any one requires changes in the others to maintain the form (49). Usually no fuss is made about this nonuniqueness. One fixes a choice of canonical statistic, canonical parameter, and cumulant function and leaves it at that. But we should keep this nonuniqueness in the back of our minds. Anything that depends on a particular choice of y, θ, and c is wrongheaded, not a general concept.

7.2 Philosophy

Distinguishing V and V^* is very important in exponential family theory. V is where the canonical statistic lives, and V^* is where the canonical parameter lives. So distinguishing V and V^* is just keeping straight the difference between statistics and parameters.

7.3 Full and Regular Families

Implicit in the definition is that the family has densities f_θ with respect to some sigma-finite measure λ on the underlying measurable space, and the log likelihood has the form

$$l(\theta) = \log f_\theta(\omega) - \text{term not containing } \theta$$

where ω is the variable ranging over the underlying measurable space, that is, we have

$$f_\theta(\omega) = e^{\langle y(\omega), \theta \rangle - c(\theta) + h(\omega)} \tag{50}$$

where $h(\omega)$ is the "term not containing θ" we are allowed to drop from the log likelihood.

For future reference we record the probability density of the distribution having canonical parameter value θ with respect to the distribution having canonical parameter value ψ, which is given by

$$f_{\theta,\psi}(\omega) = e^{\langle Y(\omega), \theta - \psi \rangle - c(\theta) + c(\psi)}. \tag{51}$$

Since these probability densities are everywhere positive, every distribution in the family has the same support.

Since densities must integrate to one, we have

$$\begin{aligned} 1 &= \int f_\theta(\omega) \lambda(d\omega) \\ &= \int e^{\langle y(\omega), \theta \rangle - c(\theta) + h(\omega)} \lambda(d\omega) \\ &= e^{-c(\theta)} \int e^{\langle y(\omega), \theta \rangle + h(\omega)} \lambda(d\omega) \end{aligned}$$

or

$$\begin{aligned} e^{c(\theta)} &= \int e^{\langle y(\omega), \theta \rangle + h(\omega)} \lambda(d\omega) \\ &= \int e^{\langle y(\omega), \theta - \psi \rangle + c(\psi)} f_\psi(\omega) \lambda(d\omega) \\ &= e^{c(\psi)} \int e^{\langle y(\omega), \theta - \psi \rangle} f_\psi(\omega) \lambda(d\omega) \\ &= e^{c(\psi)} E_\psi \left\{ e^{\langle y, \theta - \psi \rangle} \right\} \end{aligned}$$

or

$$c(\theta) = c(\psi) + \log E_\psi \left\{ e^{\langle Y, \theta - \psi \rangle} \right\} \tag{52}$$

It is useful to take (52) to define c on all of V^* up to an arbitrary constant $c(\psi)$, which does not matter (adding a constant to a cumulant function gives another cumulant function for the same family with the same canonical statistic and parameter), thinking of θ as the variable and ψ as fixed (at any arbitrarily chosen point in the canonical parameter space). At points θ such that the expectation in (52) does not exist we write $c(\theta) = \infty$. Since the integrand in (52) is strictly positive, the integral (expectation) must also

be strictly positive, and we do not have to worry about the existence of logarithms (we are using $\log(\infty) = \infty$).

The exponential family is said to be *full* if the canonical parameter space is

$$\Theta = \{ \theta \in V^* : c(\theta) < \infty \} \tag{53}$$

which is thus referred to as the canonical parameter space of the full family or as the *full canonical parameter space*. If the family is not full, then it can be enlarged to be a full family. For all $\theta \in \Theta$, where Θ is given by (53), there is a density with respect to λ given by (50), and the collection of all these densities makes a full exponential family. This construction can obviously start with a single distribution, say the one corresponding to parameter value ψ, in which case we say that this distribution and canonical statistic *generate* the exponential family. Since ψ was arbitrary, we see that any probability distribution combined with any vector-valued statistic y generates a full exponential family with y as the canonical statistic. This construction is not always interesting. For example, a Cauchy distribution with the usual variable considered the canonical statistic generates only the trivial family with only one distribution because (52) is infinite unless $\theta = \psi$.

The full family is said to be *regular* if the full canonical parameter space (53) is an open set (in the only vector topology the finite-dimensional vector space V^* can have (Section 2.4)).

7.4 Convexity of Cumulant Functions

Theorem 41. *A cumulant function defined on a whole vector space by* (52) *is a lower semicontinuous proper convex function.*

Proof. Convexity is Hölder's inequality. Lower semicontinuity is Fatou's lemma. □

The lower semicontinuity property will not be used in this document. It is only important on the boundary of the effective domain of the cumulant function (the boundary of the full canonical parameter space of the exponential family), because every convex function on a finite-dimensional affine space is continuous on the interior of its effective domain (Rockafellar, 1970, Theorem 10.1). Cumulant functions, of course, are infinitely differentiable on the interior of their effective domains, a stronger property than mere continuity.

Corollary 42. *The effective domain of a cumulant function (the full canonical parameter space) is a convex set.*

7.5 Derivatives of Cumulant Functions

The moment generating function (MGF) of the distribution of the canonical statistic corresponding to the parameter θ is

$$\begin{aligned} M_\theta(t) &= E_\theta\left\{e^{\langle Y,t\rangle}\right\} \\ &= \int e^{\langle y(\omega),t\rangle} f_\theta(\omega)\lambda(d\omega) \\ &= \int e^{\langle y(\omega),\theta+t\rangle - c(\theta)+h(\omega)}\lambda(d\omega) \\ &= \int e^{\langle y(\omega),\theta-\psi+t\rangle - c(\theta)+c(\psi)} f_\psi(\omega)\lambda(d\omega) \\ &= e^{-c(\theta)+c(\psi)} \int e^{\langle y(\omega),\theta-\psi+t\rangle} f_\psi(\omega)\lambda(d\omega) \\ &= e^{-c(\theta)+c(\psi)} E_\psi\left\{e^{\langle y,\theta-\psi+t\rangle}\right\} \\ &= e^{c(\theta+t)-c(\theta)} \end{aligned}$$

provided that this satisfies the condition to be an MGF, which is that it be finite on a neighborhood of zero. And $M_\theta(t)$ is finite for all t in a neighborhood of zero when $c(\eta)$ is finite for all η in a neighborhood of θ, that is, when θ is an interior point of the full canonical parameter space (53). So for a regular full family (when every point in the full canonical parameter space is an interior point), every distribution in the family has an MGF, and, consequently, moments of all orders determined by the MGF.

Distributions corresponding to canonical parameter values on the boundary of the full canonical parameter space (53) do not have moment generating functions and hence need not have moments of all orders or, indeed, any moments at all. The exponential family generated by the Cauchy distribution, mentioned above, is an example. The full canonical parameter space of this family is a single point, hence it has an empty interior and is its own boundary, and the single distribution in the family has no moments of any order.

For a slightly less trivial example of a nonregular exponential family, consider the one-dimensional exponential family generated by the left half Cauchy distribution, which has PDF

$$f_0(x) = \frac{2}{\pi(1+x^2)}, \qquad x < 0, \tag{54}$$

and the usual variable is the canonical statistic. The PDF of this family are

$$f_\theta(x) = \frac{e^{x\theta}}{(1+x^2)[\operatorname{Ci}(\theta)\sin(\theta) - \operatorname{Si}(\theta)\cos(\theta) + \pi\cos(\theta)/2]}, \qquad x < 0, \tag{55}$$

where Si and Ci denote the so-called sine and cosine integral functions (not well known to statisticians, they are in the GNU scientific library, CRAN package `gsl`, Hankin, 2006; Hankin, et al., 2023). The full canonical parameter space is $\{ \theta : \theta \geq 0 \}$. The expression (55) does not work when $\theta = 0$ because $\mathrm{Ci}(0) = -\infty$, but, of course, the PDF for $\theta = 0$ is (54). Every distribution for $\theta > 0$ has a moment generating function and moments of all orders. The distribution for $\theta = 0$ (54) does not have a moment generating function and does not have any moments of any order.

An example of a nonregular exponential family that could be used in real applications, is the Strauss process, which is a two-parameter exponential family, whose canonical parameter space is the closed left half space $\{ \theta : \theta_2 \leq 0 \}$. See Geyer and Møller (1994) and references cited therein for details.

The cumulant generating function (CGF) of the distribution corresponding to the parameter θ is

$$K_\theta(t) = c(\theta + t) - c(\theta)$$

provided that this satisfies the condition to be a CGF, which is that it be finite on a neighborhood of zero, that is, when θ is an interior point of the full canonical parameter space (53). Derivatives of the CGF evaluated at zero (the cumulants) are derivatives of the cumulant function evaluated at θ, hence the name "cumulant function."

The first two cumulants are

$$E_\theta(Y) = c'(\theta) \tag{56a}$$

$$\mathrm{var}_\theta(Y) = c''(\theta) \tag{56b}$$

when θ is in the interior of the full canonical parameter space (Section 6.12 above).

7.6 Identifiability

7.6.1 Canonical Parameter

Following Section 2.2 in Geyer (1990) and Theorem 1 in Geyer (2009) we present the following theorem about identifiability of the canonical parameter of an exponential family.

Theorem 43. *For a full exponential family having canonical statistic Y taking values in an abstract finite-dimensional vector space V and full canonical parameter space Θ, which is a subset of V^*, and for some vector $\delta \in V^*$, the following statements are equivalent.*

(a) *The parameter values θ and $\theta + s\delta$ correspond to the same probability distribution for some $\theta \in \Theta$ and some $s \neq 0$.*

(b) *The parameter values θ and $\theta + s\delta$ correspond to the same probability distribution for all $\theta \in \Theta$ and all real s.*

(c) *The statistic $\langle Y, \delta \rangle$ is almost surely constant for some distribution in the family.*

(d) *The statistic $\langle Y, \delta \rangle$ is almost surely constant for all distributions in the family.*

It is part of the assertion (b) that if $\theta \in \Theta$ then the whole line

$$\{\, \theta + s\delta : s \in \mathbb{R} \,\}$$

is contained in Θ.

Proof. For any $\theta \in \Theta$, let P_θ correspond to the distribution having parameter value θ. Suppose (a). Then by (51) the probability density of $P_{\theta+s\delta}$ with respect P_θ is given by

$$f_{\theta+s\delta,\theta}(\omega) = e^{s\langle Y(\omega),\delta\rangle - c(\theta+s\delta)+c(\theta)}. \tag{57}$$

In order for $P_\theta = P_{\theta+s\delta}$ to hold, we must have $f_{\theta+s\delta,\theta} = 1$ almost surely with respect to P_θ. This implies $\langle Y, \delta \rangle$ is constant almost surely with respect to P_θ. Thus (c) holds.

Now assume (c). Then (d) holds because all distributions in the family have the same support, as was noted following (51). And (a) holds because this implies (57) is constant almost everywhere with respect to P_θ. So now we know that (a), (c), and (d) are equivalent.

Now (d) implies (b) because (d) implies that (57) is constant almost everywhere with respect to every distribution in the family and for all $\theta \in \Theta$ and all real s. And (b) trivially implies (a). □

A vector δ satisfying any of the conditions of the theorem is called a *direction of constancy* of the family. It is clear from (c) or (d) of the theorem that the set of all directions of constancy is a vector subspace of V^*. It is called the *constancy space* of the family.

7.6.2 Mean Value Parameter

Theorem 44. *In an exponential family, different distributions have different means for the canonical statistic if these means exist.*

It follows that in a regular full exponential family, where the mean of the canonical statistic exists for every distribution, the means parameterize the family. Furthermore, this parameterization of the family is identifiable because no distribution can have two different means. Even if the family is not regular so not all distributions need have means, means of the canonical statistic still provide an identifiable parameterization for the subfamily of distributions such that these means exist (if there are any). This parameterization is called the *mean value parameterization*.

Proof. Suppose θ_1 and θ_2 are canonical parameter values of distributions in the family having means μ_1 and μ_2. We consider the subfamily having canonical parameter values in

$$\Theta_{\text{sub}} = \{\, s\theta_1 + (1 - s)\theta_2 : 0 \le s \le 1 \,\}.$$

This is contained in the full canonical parameter space by Corollary 42. From (51) we have

$$f_{s\theta_1+(1-s)\theta_2,\theta_2}(\omega) = e^{s\langle Y(\omega),\theta_1-\theta_2\rangle - c(s\theta_1+(1-s)\theta_2)+c(\theta_2)}$$

This shows us that the subfamily is a one-parameter exponential family having canonical statistic $\langle Y, \theta_1 - \theta_2 \rangle$, canonical parameter s, canonical parameter space $0 \le s \le 1$, and cumulant function given by

$$c_{\text{sub}}(s) = c(s\theta_1 + (1 - s)\theta_2) - c(\theta_2)$$

This subfamily need not be full, but s such that $0 < s < 1$ are in the interior of the (one-dimensional) full canonical parameter space of the full exponential family generated by this subfamily canonical statistic. Thus for $0 < s < 1$ we have by (56a) and (56b), the chain rule, and the affine function rule

$$\begin{aligned} E_{s\theta_1+(1-s)\theta_2}(\langle Y, \theta_1 - \theta_2\rangle) &= c'_{\text{sub}}(s) \\ &= c'(s\theta_1 + (1 - s)\theta_2)(\theta_1 - \theta_2) \\ \operatorname{var}_{s\theta_1+(1-s)\theta_2}(\langle Y, \theta_1 - \theta_2\rangle) &= c''_{\text{sub}}(s) \\ &= c''(s\theta_1 + (1 - s)\theta_2)(\theta_1 - \theta_2)(\theta_1 - \theta_2) \end{aligned}$$

(compare with the second derivative calculation in the proof of Theorem 25 above to see that the second derivative is correct).

Now we have two cases. If the variance above is zero, then $\langle Y, \theta_1 - \theta_2 \rangle$ is almost surely constant and 1 is a direction of constancy of this subfamily (considered a one-dimensional vector in the one-dimensional vector space $\mathbb{R}$ that contains Θ_{sub}), so every distribution in the subfamily is the same, and $\theta_1 - \theta_2$ is a direction of constancy of the original family, and $\mu_1 = \mu_2$. But if the variance above is not zero for a particular s, then it is not zero for any s. It follows that $E_{s\theta_1+(1-s)\theta_2}(\langle Y, \theta_1 - \theta_2 \rangle)$ is a strictly increasing function of s on the open interval $(0, 1)$, and $\theta_1 - \theta_2$ is not a direction of constancy of the original family.

Now we have to deal with the end points of the interval. In

$$\begin{aligned} E_{s\theta_1+(1-s)\theta_2}(\langle Y, \theta_1 - \theta_2 \rangle) &= \int \langle Y(\omega), \theta_1 - \theta_2 \rangle f_{s\theta_1+(1-s)\theta_2, \theta_2}(\omega) \, P_{\theta_2}(d\omega) \\ &= E_{\theta_2}\left[\langle Y, \theta_1 - \theta_2 \rangle e^{s\langle Y, \theta_1 - \theta_2 \rangle - c(s\theta_1+(1-s)\theta_2)+c(\theta_2)} \right] \\ &= e^{-c(s\theta_1+(1-s)\theta_2)+c(\theta_2)} E_{\theta_2}\left[\langle Y, \theta_1 - \theta_2 \rangle e^{s\langle Y, \theta_1 - \theta_2 \rangle} \right] \end{aligned}$$

the argument of the last expectation is a nondecreasing function of s for all values of Y (if $a > 0$, then ae^{sa} is an increasing function of s; if $a < 0$, then ae^{sa} is again an increasing function of s). Hence by monotone convergence

$$\begin{aligned} \lim_{s \downarrow 0} E_{s\theta_1+(1-s)\theta_2}(\langle Y, \theta_1 - \theta_2 \rangle) &= E_{\theta_2}(\langle Y, \theta_1 - \theta_2 \rangle) = \langle \mu_2, \theta_1 - \theta_2 \rangle \\ \lim_{s \uparrow 1} E_{s\theta_1+(1-s)\theta_2}(\langle Y, \theta_1 - \theta_2 \rangle) &= E_{\theta_1}(\langle Y, \theta_1 - \theta_2 \rangle) = \langle \mu_1, \theta_1 - \theta_2 \rangle \end{aligned}$$

Hence, if $\theta_1 - \theta_2$ is not a direction of constancy, then $E_{s\theta_1+(1-s)\theta_2}(\langle Y, \theta_1 - \theta_2 \rangle)$ is a strictly increasing function of s on the closed interval $[0, 1]$, and hence $\mu_1 \neq \mu_2$. □

7.7 Minimality

An exponential family has *minimal representation* or is just *minimal* for short if the dimensions of the canonical statistic vector and canonical parameter vector are as small as possible. By Theorem 43 this implies the canonical parameterization is identifiable (there are no directions of constancy). By Theorem 43 this implies the canonical parameterization is not concentrated on a hyperplane. By Theorem 29 this implies the interior of the full canonical parameter space is nonempty, because, as we shall presently see, otherwise we could find a canonical parameter space of lower dimension.

We now show that if the originally given canonical statistic vector and canonical parameter vector do not have minimal dimension, we can always find a minimal representation.

As usual, consider an exponential family with canonical statistic Y taking values in the abstract vector space V, canonical parameter θ taking values in V^*, and cumulant function c given by (52).

Let Θ defined by (53) be the full canonical parameter space. We want Θ to have full dimension. If it does not, let ψ be an arbitrary element of Θ, define a new canonical parameter $\delta = \theta - \psi$ and its new full canonical parameter space

$$\Delta = \Theta - \psi = \{\, \theta - \psi : \theta \in \Theta \,\},$$

and let U be the vector subspace of V^* spanned by Δ. The log likelihood for the parameter δ is now

$$\begin{aligned} l_{\text{new}}(\delta) &= \langle y, \psi + \delta \rangle - c(\psi + \delta) \\ &= \langle y, \psi \rangle + \langle y, \delta \rangle - c(\psi + \delta) \end{aligned}$$

and we may drop the term that does not contain the parameter δ obtaining

$$\begin{aligned} l_{\text{new}}(\delta) &= \langle y, \delta \rangle - c(\psi + \delta) \\ &= \langle y, \delta \rangle - c_{\text{new}}(\delta) \end{aligned}$$

And now we want to consider the canonical parameter δ to have full dimension, so we want to consider it living in the vector space U. But this means we need to define a new canonical statistic taking values in U^*. The formalism of abstract vector spaces does this automatically for us. For each data value ω, the linear functional $\delta \mapsto \langle Y(\omega), \delta \rangle$ is an element of U^* when we allow δ to range over U. So that is our new canonical statistic

$$X(\omega) = \langle Y(\omega), \cdot \rangle$$

and our log likelihood is now

$$l_{\text{new}}(\delta) = \langle x, \delta \rangle - c_{\text{new}}(\delta)$$

where now $\langle \cdot, \cdot \rangle$ is the canonical bilinear form that places U and U^* in duality. So now we have an exponential family with parameter δ that has full dimension (in U).

To avoid even more annoying changes of notation, we return to having an exponential family with canonical statistic y taking values in the abstract vector space V, canonical parameter θ taking values in V^*, and cumulant

function c given by (52). But now we assume that Θ has full dimension (possibly because we have gone through the procedure outlined above).

Now we consider the dimension of the canonical statistic. Let A be the affine support of the canonical statistic for any distribution in the family (Section 6.5 above). This is the affine support of every distribution in the family because every distribution in the family has the same support. Let U be the translation space of A. Fix an arbitrary point $a \in A$. Then we define a new canonical statistic by

$$X(\omega) = Y(\omega) - a$$

and consider X to take values in U rather than V. This may require that we change the underlying probability space to delete the set

$$\{\, \omega : Y(\omega) - a \notin U \,\}$$

which has probability zero under every distribution in the family. Now we need a new canonical parameter taking values in U^*. To do this we simply apply (52)

$$c_{\text{new}}(\theta) = \log E_\psi\bigl\{ e^{\langle X, \theta \rangle} \bigr\}, \qquad \theta \in U^*.$$

Then we have a new exponential family with minimal representation having canonical statistic vector X taking values in U, cumulant function c_{new}, and full canonical parameter space $\operatorname{dom} c_{\text{new}}$, which is a subset of U^*.

7.8 Asymptotics of Maximum Likelihood

For a minimal regular full exponential family the mapping from canonical to mean value parameter is a function h defined by

$$h(\theta) = c'(\theta) \tag{58}$$

which has derivative defined by

$$h'(\theta) = c''(\theta)$$

(that we take $c'(\theta)$ to have type $V^* \to V$ rather than type $V^* \to V^{**}$, using the natural isomorphism of V and V^{**}, is used everywhere in this section).

Lemma 45. *Let c be the cumulant function and Θ the canonical parameter space of a regular full exponential family. The following are equivalent.*

(a) *$c''(\theta)$ is invertible for some $\theta \in \Theta$.*

(b) $c''(\theta)$ *is invertible for all* $\theta \in \Theta$.

(c) $c''(\theta)$ *is positive definite for some* $\theta \in \Theta$.

(d) $c''(\theta)$ *is positive definite for all* $\theta \in \Theta$.

(e) *The family has no nonzero directions of constancy.*

Proof. It follows from Theorem 36 that (a) and (c) are equivalent and (b) and (d) are equivalent. It follows from (29) and (56b) that

$$\operatorname{var}_\theta(Y)(\delta)(\delta) = \operatorname{var}_\theta\{\langle Y, \delta\rangle\} = c''(\theta)(\delta)(\delta) \tag{59}$$

and this is zero for some nonzero δ if and only if (c) is false and if and only if $\langle Y, \delta\rangle$ is almost surely constant with respect to the distribution for parameter value θ. But (51) shows all distributions in the family are mutually absolutely continuous with respect to each other. Hence (59) is false for some θ if and only if it is false for all θ. Thus (e) is equivalent to all of the others. □

Theorem 46. *For a regular full exponential family having minimal representation the map from canonical to mean value parameter is invertible and its derivative at any parameter value is also invertible.*

Proof. A minimal representation guarantees no nonzero directions of constancy. Hence the rest follows from the lemma. □

The derivative of (49) is

$$l'(\theta) = \langle y, \cdot\rangle - c'(\theta)$$

Since the log likelihood is a function $\Theta \to \mathbb{R}$, its derivative is a function $V^* \to \mathbb{R}$, but each such function is an element of V^{**}. Again using $V^{**} = V$ via the natural isomorphism, we can also write

$$l'(\theta) = y - c'(\theta)$$

where now both y and $c'(\theta)$ are taken to be elements of V. This makes sense because

$$c'(\theta) = E_\theta(Y) = \mu$$

the mean value parameter corresponding to θ. Thus we can also write

$$l'(\theta) = y - \mu$$

with the understanding $\mu = c'(\theta)$.

One definition of Fisher information is the variance of the score (first derivative of the log likelihood function). For the parameter θ that is

$$\operatorname{var}_\theta(Y - \mu) = \operatorname{var}_\theta(Y) = c''(\theta)$$

And, as we already know, we can take this either to be a bilinear form $V^* \to V^* \to \mathbb{R}$ or a linear function $V^* \to V$ (which again uses $V^{**} = V$). As in (58), let h denote the map from canonical to mean value parameter, so $h(\theta) = c'(\theta)$ and $h'(\theta) = c''(\theta)$. When we have an identifiable canonical parameterization, h is invertible considered as a function from its domain (the full canonical parameter space) to its range. Let j denote this inverse. Then from the inverse function theorem it follows that j is infinitely differentiable at all points (because h is) and that its derivative is the inverse of the derivative of h, that is, if $\mu = c'(\theta)$, then $h'(\theta)$ and $j'(\mu)$ are inverse linear functions

$$V \underset{h'(\theta)}{\overset{j'(\mu)}{\rightleftarrows}} V^*$$

so either composition of these two functions is an identity function.

The log likelihood for μ is $l \circ j$. Its derivative is, by the chain rule

$$(l \circ j)'(\mu) = l'(\theta) \circ j'(\mu) \tag{60}$$

(still assuming $\mu = c'(\theta)$), the diagram for this being

$$V \xrightarrow{j'(\mu)} V^* \xrightarrow{l'(\theta)} \mathbb{R}$$

The value of this derivative at a vector $\xi \in V$ can be written

$$\langle y - \mu, j'(\mu)(\xi) \rangle$$

This makes sense because $y - \mu \in V$ and $j'(\mu)(\xi) \in V^*$. Now Fisher information for μ is

$$\operatorname{var}_\theta\{(l \circ j)'(\mu)\}$$

If we think of this variance as a bilinear form, we can write it as

$$\begin{aligned} \operatorname{var}_\theta\{(l \circ j)'(\mu)\}(\xi)(\zeta) &= E_\theta\{\langle y - \mu, j'(\mu)(\xi) \rangle \langle y - \mu, j'(\mu)(\zeta) \rangle\} \\ &= c''(\theta)\big(j'(\mu)(\xi)\big)\big(j'(\mu)(\zeta)\big) \end{aligned} \tag{61}$$

and this makes sense because this bilinear form has type $V \to V \to \mathbb{R}$ and $c''(\theta)$ considered as a bilinear form has type $V^* \to V^* \to \mathbb{R}$, and $j'(\mu)$ maps $V \to V^*$.

But now we need to take into account that $c''(\theta)$ and $j'(\mu)$ are inverse functions so $c''(\theta) \circ j'(\mu) = \mathrm{id}_V$

$$V \xrightarrow{j'(\mu)} V^* \xrightarrow{c''(\theta)} V$$

so

$$c''(\theta)\big(j'(\mu)(\xi)\big) = \xi \tag{62}$$

for all $\xi \in V$. But this is a bit confusing because in in (61) we are considering $c''(\theta)$ as a bilinear form and in (62) we are considering $c''(\theta)$ as a linear function. We can get the bilinear form back as

$$\mathrm{var}_\theta\{(l \circ j)'(\mu)\}(\xi)(\zeta) = \langle \xi, j'(\mu)(\zeta) \rangle$$

And this tells us Fisher information for μ is

$$\mathrm{var}_\theta\{(l \circ j)'(\mu)\} = j'(\mu)$$

considered as a linear function $V \to V^*$.

We summarize the calculations above as a theorem.

Theorem 47. *For a regular full exponential family, Fisher information for the canonical parameter is $c''(\theta)$, and Fisher information for the mean value parameter is $j'(\mu) = [c''(\theta)]^{-1}$.*

Theorem 47 is a bit disingenuous because it silently relies on the representation $V^{**} = V$. If we don't do that, then $c''(\theta)$ and $h'(\theta)$ have type $V^* \to V^{**}$. So $j'(\mu) = [h'(\theta)]^{-1} = [c''(\theta)]^{-1}$ has type $V^{**} \to V^*$. And (60) has type

$$V^{**} \xrightarrow{j'(\mu)} V^* \xrightarrow{l'(\theta)} \mathbb{R}$$

that is $V^{**} \to \mathbb{R}$ which is V^{***}. So its variance has type $V^{****} \to V^{****} \to \mathbb{R} = V^{****} \to V^{*****}$. And this matches the type of $j'(\mu)$ only if we use $V^{**} = V$.

Theorem 48. *For IID sampling from a minimal regular full exponential family, if $\hat{\theta}_n$ is the MLE for the canonical parameter vector θ sample size n and $\hat{\mu}_n$ is the MLE for the mean value parameter vector μ for sample size n, then*

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, c''(\theta)) \tag{63}$$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, j'(\mu)) \tag{64}$$

(this uses $V^{**} = V$).

Proof. The log likelihood for IID sampling is

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^n \bigl[\langle y_i, \theta \rangle - c(\theta) \bigr] \\ &= \left\langle \sum_{i=1}^n y_i, \theta \right\rangle - nc(\theta) \\ &= \langle n \bar{y}_n, \theta \rangle - nc(\theta) \\ &= n[\langle \bar{y}_n, \theta \rangle - c(\theta)] \end{aligned}$$

where y_1, y_2, ... are IID random vectors having the distribution for canonical parameter θ and mean value parameter μ from the exponential family under discussion and (as usual)

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$$

This log likelihood is concave for the same reasons that l_1 is concave. Hence any θ such that the first derivative is equal to zero is an MLE but the MLE is unique by minimality. Since

$$l_n'(\theta) = n[\bar{y}_n - c'(\theta)] = n[\bar{y}_n - h(\theta)] \tag{65}$$

the unique MLE for θ is

$$\hat{\theta}_n = j(\bar{y}_n)$$

provided $\bar{y}_n$ is in $\operatorname{dom}(j) = h(\Theta)$. It can happen that the MLE does not exist (this happens, for example, for the binomial distribution when $\bar{y}_n = 0$ or $\bar{y}_n = 1$). But we can add another point not in Θ to be the MLE when $\bar{y}_n \notin \operatorname{dom}(j)$. It does not matter what this point is because $\mu \in \operatorname{dom}(j)$ and $\bar{y}_n \to \mu$ (in probability and almost surely). So $\bar{y}_n \in \operatorname{dom}(j)$ for sufficiently large n with arbitrarily high probability. For more pedantic discussion of this issue, see Geyer (2013, Section 3.5). And by invariance of maximum likelihood the unique MLE for μ is

$$\hat{\mu}_n = h(\hat{\theta}_n) = (h \circ j)(\bar{y}_n) = \bar{y}_n$$

(again provided $\bar{y}_n \in \operatorname{dom}(j)$ because otherwise $\bar{y}_n$ is not a possible mean value).

Now the CLT applied to the latter gives (63) because $\hat{\mu}_n = \bar{y}_n$ with probability converging to one as $n \to \infty$ and $E_\theta(y_i) = \mu$ and $\operatorname{var}_\theta(y_i) = c''(\theta)$.

Then (64) follows by the delta method because of

$$\hat{\theta}_n - \theta = j(\bar{y}_n) - j(\mu)$$

so by the delta method

$$\begin{aligned}\sqrt{n}[\hat{\theta}_n - \theta] &= \sqrt{n}[j(\bar{y}_n) - j(\mu)] \\ &\xrightarrow{\mathcal{D}} \mathcal{N}(0, j'(\mu) \circ c''(\theta) \circ j'(\mu)^*)\end{aligned}$$

applying Theorem 40 and Section 6.10. But $j'(\mu) \circ c''(\theta)$ is the identity operator on V^*, and $j'(\mu)^* = j'(\mu)$ by Theorem 35, so that gives (64). □

Let us check that these make sense. If we do not use $V^{**} = V$, then (65) has type $V^* \to \mathbb{R}$ or V^{**}. Thus $\bar{y}_n$ and μ must be elements of V^{**}. And the variance must have type $V^{***} \to V^{***} \to \mathbb{R}$ or $V^{***} \to V^{****}$. But $c''(\theta)$ has type $V^* \to V^* \to \mathbb{R}$, So we already need to use $V^{**} = V$ to make (63) correct.

If we wanted (63) to be correct without assuming $V^{**} = V$, we would need to replace $c''(\theta)$ by $c''(\theta)^{**}$.

Alternatively, we can take $\bar{y}_n$ and μ to be elements of V except then that makes $l'(\theta)$ given by (65) a point of V, hence not a linear function, which disagrees with the PhD level view of differentiation explicated in Section 4.2 above. So this step, by itself, is already using $V^{**} = V$. (See also Section 6.3 about considering μ to be in V also assuming $V^{**} = V$.)

Moving on to (64) and still avoiding assuming $V^{**} = V$, we have $\hat{\mu}_n$ in V^{**} and $h = c'$ mapping $\Theta \to V^{**}$ so its inverse j maps $h(\Theta) \subset V^{**}$ to Θ. So $j'(\mu)$ maps $V^{**} \to V^*$. But the variance or asymptotic variance of $\hat{\theta}_n - \theta$ must have type $V^{**} \to V^{**} \to \mathbb{R}$ or $V^{**} \to V^{***}$. So again we need to use $V^{**} = V$ to make (64) correct.

If we did not want to use $V^{**} = V$, then the proof shows we would have to replace $j'(\mu)$ with $j'(\mu)^*$ in (64). This does give the correct types

$$V^{**} \xrightarrow{j'(\mu)} V^*$$

hence

$$V^{***} \xleftarrow{j'(\mu)^*} V^{**}$$

8 The Category of Exponential Families

8.1 Questions

What is the category (in the sense of category theory) of exponential families? The category of exponential families is a category in which the

objects are exponential families and the morphisms are what? Some sort of map or map-like thingummy between exponential families.

8.2 Objects of the Category

8.2.1 Densities

To answer these questions we need to know what an exponential family is. Combining the logic of Section 1.4 and Chapter 4 of Geyer (1990) with Section 3.1 of Geyer (2009), an exponential family of (probability) distributions is a statistical model having log probability density functions with respect to some sigma-finite positive measure (on some measurable space) of the form $h \circ y$ where h is an affine function on some finite-dimensional affine space (a different affine function for each distribution in the family) and y is a statistic, called the canonical statistic, (the same for all distributions in the family).

So an exponential family involves three spaces, and two functions

$$(\Omega, \mathcal{A}, \lambda) \xrightarrow{\ y\ } B \xrightarrow{\ h\ } \mathbb{R} \tag{66}$$

$(\Omega, \mathcal{A}, \lambda)$ is the sigma-finite measure space, B is the finite-dimensional affine space, y is the canonical statistic (which does not vary), and h the affine function (which varies to give the different distributions in the family).

This description is unsatisfactory because it does not explicitly indicate the family $\mathcal{H}$ of affine functions $B \to \mathbb{R}$ which h varies over. So we will rewrite our diagram as

$$(\Omega, \mathcal{A}, \lambda) \xrightarrow{\ y\ } B \xrightarrow{\ \mathcal{H}\ } \mathbb{R} \tag{67}$$

But both diagrams are intended to indicate the same thing. For each $h \in \mathcal{H}$ there is the composition $h \circ y$ and $\exp \circ h \circ y$ is a probability density function with respect to λ, that is,

$$\int e^{h \circ y} \, d\lambda = 1. \tag{68}$$

8.2.2 Measures

Actually, (67) is also unsatisfactory in that it does not incorporate the usual abstract nonsense of measure theory. We want an exponential family of *distributions* not of *densities*.

Thus we consider (67) a characterization or representation of the object rather than the object itself. The object is the family of probability measures

$$\{\, P_h : h \in \mathcal{H} \,\}, \tag{69}$$

where P_h is a probability measure on $(\Omega, \mathcal{A})$ given by

$$P_h(A) = \int_A e^{h \circ y} \, d\lambda, \qquad A \in \mathcal{A}. \tag{70}$$

We consider the objects equal if the families of probability measures are equal.

8.3 Morphisms of the Category

8.3.1 Densities

Morphisms between these objects consist of a pair of functions f and g going from parts of the source object (of the category) to parts of the target object such that the diagram

$$\begin{array}{ccccc} (\Omega, \mathcal{A}, \lambda) & \xrightarrow{\;y\;} & B & \xrightarrow{\;\mathcal{H}\;} & \mathbb{R} \\ \big\downarrow f & & \big\downarrow g & & \\ (\Omega', \mathcal{A}', \lambda') & \xrightarrow{\;y'\;} & B' & \xrightarrow{\;\mathcal{H}'\;} & \mathbb{R} \end{array} \tag{71}$$

has the following properties

(i) f is measurable, that is,

$$f^{-1}(A') \in \mathcal{A}, \qquad A' \in \mathcal{A}',$$

(ii) $\lambda' = \lambda \circ f^{-1}$, that is,

$$\lambda'(A') = \lambda\big(f^{-1}(A')\big), \qquad A' \in \mathcal{A}',$$

(iii) g is affine,

(iv) $g \circ y = y' \circ f$ almost everywhere with respect to λ, and

(v) the target object is a submodel of the source object, that is, for every $h' \in \mathcal{H}'$ there exists $h \in \mathcal{H}$ such that $h \circ y = h' \circ y' \circ f$ almost everywhere with respect to λ.

Items (i) and (ii) go together. The notation $\lambda' = \lambda \circ f^{-1}$ is not even defined unless f is measurable. For short, we say that item (iv) says the rectangle in the diagram (71) commutes.

8.3.2 Measures

The morphism (71) induces a mapping from the target object (exponential family of distributions) to the source object (exponential family of distributions). In the target object, the probability measure on $(\Omega', \mathcal{A}')$ having density $h' \circ y'$ with respect to λ' is mapped to the probability measure in the source object on $(\Omega, \mathcal{A})$ having density $h' \circ y' \circ f$ or density $h' \circ g \circ y$ with respect to λ. Item (iv) says these two densities are densities of the same probability measure. Item (v) says this probability measure also has density $h \circ y$ for some $h \in \mathcal{H}$. So this is a measure in the source object. Of course, $h' \circ y' \circ f$ and $h' \circ g \circ y$ and $h \circ y$ may be three different functions. But they must be equal almost everywhere with respect to λ and hence densities of the same probability measure (characterized by (70)).

We will consider this mapping between measures to be the morphism. So the pair (f, g) only represents or characterizes the morphism. So morphisms are equal if they are the same map of measures (this section). They do not have to have the same (f, g).

We do not have an explicit formula for this map of measures. As usual, we work mostly with densities, but only measures are unique. If we wanted we could write

$$\mathcal{O}_1 \xrightarrow{m} \mathcal{O}_2$$

where $\mathcal{O}_1$ and $\mathcal{O}_2$ are objects of the category, which are exponential families of distributions, and m is the morphism. Note that m corresponds to a function $\mathcal{O}_2 \to \mathcal{O}_1$, but which way the arrows go in a category is an arbitrary choice. We have chosen to say the arrows go the way the functions go in (71) rather than the way the function that maps measures to measures goes (and we still don't have a notation for that function because m is the morphism, which is denoted by an arrow going the opposite way of the function, so m isn't exactly that function).

Also note that the map of measures is not just any map. As described above, it is the map that takes the measure having log density $h' \circ y'$ with respect to λ' to the measure having log density $h' \circ g \circ y$ with respect to λ. And that is a very special map that embeds one exponential family $\mathcal{O}_2$ into another exponential family $\mathcal{O}_1$ as a canonical affine submodel (Sections 8.9 and 8.10 below).

8.4 Apology for Measure Theory

It is rare in practice that one would bother with the f in (71). Usually one uses the same dominating measure for source and target families giving

the diagram

$$\begin{array}{ccccc} (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y} & B & \xrightarrow{\mathcal{H}} & \mathbb{R} \\ \Big\downarrow \mathrm{id}_\Omega & & \Big\downarrow g & & \\ (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y'} & B' & \xrightarrow{\mathcal{H}'} & \mathbb{R} \end{array} \tag{72}$$

or, more simply,

$$\begin{array}{ccccc} (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y} & B & \xrightarrow{\mathcal{H}} & \mathbb{R} \\ & \searrow^{y'} & \Big\downarrow g & & \\ & & B' & \xrightarrow{\mathcal{H}'} & \mathbb{R} \end{array} \tag{73}$$

so now item (iv) of our characterization of morphisms just says $y' = g \circ y$.

In fact, I do not think that in all of my work on exponential families extending over more than forty years I have ever changed measures as (71) does. But the logic of measure theory (the underlying probability measure may be chosen arbitrarily) and the logic of exponential families (any dominating measure can be used for densities of the family) both suggest the change of measure should be allowed. Moreover the logic of category theory also suggests it. If (67) describes objects, then it seems obvious that all parts of the diagram could change to make another object.

So we keep the possibility of change-of-measure being part of a morphism although we rarely use it.

8.5 Is This a Category?

In order for the foregoing to specify a category, we need to say

- what the identity morphisms are,
- what composition of morphisms means, and
- show that the identity laws and associativity of composition law are satisfied.

The identity morphism for the object (66) is

$$\begin{array}{ccccc} (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y} & B & \xrightarrow{\mathcal{H}} & \mathbb{R} \\ \Big\downarrow f & & \Big\downarrow g & & \\ (\Omega, \mathcal{A}, \lambda') & \xrightarrow{y'} & B' & \xrightarrow{\mathcal{H}'} & \mathbb{R} \end{array} \tag{74}$$

satisfying our conditions to be a morphism and also

(vi) the source and target objects are equal (families of probability measures) and mapping of probability measures to probability measures described in Section 8.3.2 is the identity function.

Note that in (74) Ω and $\mathcal{A}$ are the same in both objects, so that the families of distributions are on the same measurable space $(\Omega, \mathcal{A})$. But the other parts of the specification may differ. Note also that

$$\begin{array}{ccccc} (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y} & B & \xrightarrow{\mathcal{H}} & \mathbb{R} \\ \Big\downarrow \mathrm{id}_\Omega & & \Big\downarrow \mathrm{id}_B & & \\ (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y} & B & \xrightarrow{\mathcal{H}} & \mathbb{R} \end{array}$$

is always one representation of the identity morphism. The point of (74) is that it is not the only representation.

And a composition of morphisms (represented by)

$$\begin{array}{ccccc} (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y} & B & \xrightarrow{\mathcal{H}} & \mathbb{R} \\ \Big\downarrow f & & \Big\downarrow g & & \\ (\Omega', \mathcal{A}', \lambda') & \xrightarrow{y'} & B' & \xrightarrow{\mathcal{H}'} & \mathbb{R} \\ \Big\downarrow f' & & \Big\downarrow g' & & \\ (\Omega'', \mathcal{A}'', \lambda'') & \xrightarrow{y''} & B'' & \xrightarrow{\mathcal{H}''} & \mathbb{R} \end{array} \tag{75}$$

is the morphism (represented by)

$$\begin{array}{ccccc} (\Omega, \mathcal{A}, \lambda) & \xrightarrow{y} & B & \xrightarrow{\mathcal{H}} & \mathbb{R} \\ \Big\downarrow f' \circ f & & \Big\downarrow g' \circ g & & \\ (\Omega'', \mathcal{A}'', \lambda'') & \xrightarrow{y''} & B'' & \xrightarrow{\mathcal{H}''} & \mathbb{R} \end{array}$$

It should be clear that the identity and associativity laws for the morphisms defined here follow from the identity and associativity laws for the category of sets and functions.

8.6 Full Families

An exponential family (67) is *full* if $\mathcal{H}$ is the set of all affine functions $h : B \to \mathbb{R}$ such that (68) holds (that is, $h \circ y$ is a log probability density with respect to λ).

Clearly, any exponential family can be enlarged to make it full without changing parts of the specification (67) except for the family of affine functions $\mathcal{H}$. Moreover, this enlargement, is unique. Given a point-valued statistic $y : (\Omega, \mathcal{A}, \lambda) \to B$ there is only one $\mathcal{H}$ that combines with it to give a full exponential family represented by (67).

8.7 Canonical Parameters

Section 1.4 of Geyer (1990) says that a canonical parameterization of an exponential family in the "affine picture" (which is what the category of exponential families formalizes) is given by the derivatives of the affine functions. So we define for the object (67) of the category, the canonical parameter space

$$\Theta = \nabla\mathcal{H} = \{ \nabla h(y) : h \in \mathcal{H} \} \tag{76}$$

and say the distribution in the exponential family having log density $h \circ y$ with respect to λ has canonical parameter $\theta = \nabla h(y)$. We are now using ∇ for derivatives rather than primes because we are using primes to denote different objects of the same type (objects or morphisms of the category). And (Section 4.3.5 above) the derivative $\nabla h(y)$ does not depend on y. It is the associated linear function of the affine function h.

Since h is an affine function $B \to \mathbb{R}$, its derivative θ is a linear function $V \to \mathbb{R}$, where V is the translation space (tangent space) of B. And a linear function $V \to \mathbb{R}$ is an element of the dual space V^*. Thus (76) is a subset of V^*.

When we compare what was just said with the vector picture, we see that it agrees. When we differentiate $\langle y, \theta \rangle - c(\theta)$ with respect to y, we get $\langle \cdot\, , \theta \rangle$ but this is just another notation for θ. Hence θ as described above is the canonical parameter that is dual to the canonical statistic y.

We rewrite Theorem 43 so it applies to the affine picture rather than the vector picture.

Theorem 49. *For an object* (67) *in the category of exponential families the following statements are equivalent.*

(a) The affine functions h_1 and h_2 in $\mathcal{H}$ correspond to the same probability distribution.

(b) The affine function $h_1 - h_2$ is constant on the affine support of $\lambda \circ y^{-1}$ for all real s.

(c) *The function* $[sh_1 - (1 - s)h_2] \circ y$ *is an unnormalized probability density with respect to* λ *for all real* s*, and all correspond to the same probability distribution.*

Let $\theta_1 = \nabla h_1(y)$ and $\theta_2 = \nabla h_2(y)$, then $\delta = \theta_1 - \theta_2$ is a direction of constancy of the exponential family. And this definition agrees with that of Section 7.6.1 in case the family is full. In case the family is non-full, we only obtain directions of constancy corresponding to $\theta_1 - \theta_2$ that actually occur in the family, whereas Theorem 43 insists the family in question is full. So we obtain the same directions of constancy as Theorem 43 if we start with a full exponential family.

8.8 Regular Full Families

As in the vector picture, a full family is regular if its canonical parameter space (76) is an open subset of the vector space containing it.

8.9 Canonical Affine Submodels

In the diagram (71) we have two exponential families of distributions. The first of these (the source) has

- canonical statistic y and
- canonical parameter θ ranging over the
- canonical parameter space (76).

The second of these (the target) has

- canonical statistic $g \circ y$ and
- canonical parameter θ' ranging over the
- canonical parameter space $\Theta' = \nabla \mathcal{H}'$.

From item (v) of our axioms for morphisms we have for every $h' \in \mathcal{H}'$ there exists an $h \in \mathcal{H}$ such that $h' \circ g \circ y = h \circ y$ almost everywhere with respect to λ, hence $h' \circ g = h$ everywhere on the support of $\lambda \circ y^{-1}$.

And this implies $\theta = \nabla h(y)$ and $\theta' = \nabla h'(y') \circ \nabla g(y)$ differ by a direction of constancy (because they correspond to the same probability measure).

If we insist the source family is full, then the mapping $h' \mapsto h' \circ g$ maps $\mathcal{H}' \to \mathcal{H}$.

Now from (21) we have $\theta = (\nabla g(y))^*(\theta')$ where θ is the canonical parameter of the source model and θ' is the canonical parameter of the target model. Thus we see that the target object is a canonical affine submodel of the the source object. And we see that we have captured the important concept of canonical affine submodel in our use of category theory here. (We can choose g to make $(\nabla g(y))^*$ be any linear function we want it to be. Hence we can express any canonical affine submodel this way.)

The other mapping for a canonical affine submodel, the one for mean value parameters is just g. From item (iv) of our axioms for morphisms we have $g \circ y = y' \circ f$ almost everywhere, which we interpret as

$$g(\text{canonical statistic of source family}) = \text{canonical statistic of target family}$$

and hence $g(\mu) = \mu'$ where μ and μ' are the mean value parameter vectors of the source and target families, respectively.

8.10 Morphisms Instead of Elements and Subsets

It may seem strange that a canonical affine submodel is not literally a subset of the supermodel it is a submodel of, but this is the way category theory works in general.

In the category of sets and functions, there is no element-of operation, but morphisms can take the place of elements. If 1 denotes some singleton set (it does not matter which one), then functions $x : 1 \to S$ pick out elements of S. So "elements are a special case of functions" (Leinster, 2014, Section 1). And function evaluation is a special case of composition:

$$1 \xrightarrow{x} S \xrightarrow{f} T$$

takes the place of the evaluation $f(x)$. There is no subset-of operation either. We just take injective functions $A \to B$ to treat A as a subset of B (Leinster, 2014, bottom of p. 409).

So this suggests the following.

Theorem 50. *The function mapping probability measures to probability measures described in Section 8.3.2 (what morphisms of the category really are) is always injective.*

Proof. In diagram (71) let h_1 and h_2 be elements of $\mathcal{H}'$ such that $h_1 \circ g \circ y = h_2 \circ g \circ y$ almost surely with respect to λ. Then by item (iv) of our

characterization of morphisms we also have $h_1 \circ y' \circ f = h_2 \circ y' \circ f$ almost surely with respect to λ. Hence $h_1 \circ y' = h_2 \circ y'$ almost surely with respect to λ'. □

So we are just following the way category theory works in general. There are no subsets, just injective morphisms. So we should not think of statistical submodels as subsets but rather as injective morphisms, as we do here. (Of course, the theorem says all morphisms are injective in this category.)

8.11 Isomorphisms of the Category

In light of the preceding section, two exponential families of distributions are isomorphic in the sense of category theory (with the category as defined above) if each is a canonical affine submodel of the other.

Another way to say this is that an isomorphism of the category is a morphism that does not do dimension reduction.

8.12 Identifiability

We rewrite Section 7.7 above to show that it becomes entirely trivial in the affine picture. In this we follow Geyer (1990, Section 1.5).

For an exponential family (67), let L denote the affine support of the measure $\lambda \circ y^{-1}$. Then log densities h_1 and h_2 in $\mathcal{H}$ that agree on L correspond to the same distribution. Hence if we replace B by L and restrict all of the densities to L, this will give us an identifiable canonical parameterization. (This may require us to delete a set of measure zero from the measure space $(\Omega, \mathcal{A}, \lambda)$.)

Simple. The vector picture makes the discussion of this subject messy.

8.13 Minimality

We still have the problem that even after we have done the above (so the affine support of $\lambda \circ y^{-1}$ is B), we still have the problem that the full canonical parameter space has empty interior because it is contained in a proper affine subspace of the vector space in which it lives (the dual space of the translation space of B).

An example of this is the exponential family generated by the Cauchy distribution with the usual variable as the canonical statistic. The support of the canonical statistic is the whole real line, but the canonical parameter space is zero-dimensional. This issue never arises in practice but is a theoretical consideration.

The affine picture is no help here. We have to follow the procedure described in Section 7.7. But as we shall see in Section 9, we do not need minimality in the affine picture.

8.14 Standard Exponential Families

Geyer (1990), following Barndorff-Nielsen (1978), stressed *standard* exponential families: a *standard* exponential family of probability densities with respect to a positive Borel measure on a finite-dimensional affine space is one such that the log densities are affine functions.

We could make a category of standard exponential families by putting the measures on the affine spaces

$$(B, \lambda) \xrightarrow{\mathcal{H}} \mathbb{R} \tag{77}$$

where B is still an affine space, λ is a positive Borel measure on B, and H is still a family of affine functions.

Then the theory of the category is developed as above, *mutatis mutandis*. But Geyer (2009) avoids standard exponential families, and we have also done so here. There are many reasons for this.

- We have lost track of the canonical statistic y. We can say it is the identity function on B. But this seems weird to many statisticians. Also it does not allow y to be a dimension reduction.

- The measure λ can seem weird to many statisticians. For example, if we want a binomial family of distributions with sample size n, we must take λ to put mass $\binom{n}{y}$ at the points $y = 0, 1, \ldots, n$. We cannot use counting measure. That would not give densities e^h for affine functions h.

- Probability theory in general is not fussy about the underlying probability space. Use whatever is convenient. But here we are insisting on an affine space.

These are different aspects of the same problem. Hence the category of exponential families is defined the way we do above, with arbitrary measure spaces and canonical statistics.

8.15 Cumulant Functions

Cumulant functions are inherently a vector space tool, so we have to go to the vector picture to get them. We replace (52) with

$$c(\theta) = c(\psi) + \log E_\psi\left\{e^{\langle Y - y_0, \theta - \psi\rangle}\right\} \tag{78}$$

where y_0 is an arbitrary point in the affine space where the canonical statistic takes values.

We have to do this because we need $Y - y_0$ to be a vector. And, of course, $c'(\theta)$ no longer gives the mean of Y but rather the mean of $Y - y_0$. So the mean value parameter is $\mu = y_0 + c'(\theta)$.

But then everything else is OK. We can write the log likelihood as

$$l(\theta) = \langle Y - y_0, \theta\rangle - c(\theta), \qquad \theta \in \Theta$$

(but we cannot take the minus sign out of the angle brackets).

9 Category of Closures of Exponential Families

9.1 Definition

Almost nothing needs to be done to the theory of Section 8. Simply replace families of affine functions $\mathcal{H}$, $\mathcal{H}'$, and so forth with families of *generalized* affine functions (Geyer, 1990, Chapters 3 and 4).

9.2 Generalized Affine Functions

Generalized affine functions on finite-dimensional affine spaces are pointwise limits of sequences of real-valued affine functions when $-\infty$ and $+\infty$ are allowed as limits. (This is not the definition of generalized affine function given in Section 3.1 of Geyer (1990), but it is a characterization of them, Section 3.3 of Geyer (1990). The definition is an extended-real-valued function that is both convex and concave.)

9.3 Subprobability Density Functions

One issue is that when one takes limits of sequences of probability densities, one needn't have a probability density limit. If $h_n \to h$ pointwise, $\exp \circ h_n \circ y$ being a probability density for each n, then Fatou's lemma only guarantees

$$\int e^{h \circ y}\, d\lambda \le 1. \tag{79}$$

So we have to replace equation (68) in Section 8 with (79) when we use generalized affine functions.

In (79) we are using the conventions $e^{-\infty} = 0$ and $e^{+\infty} = +\infty$. One might ask how an integral with possibly infinite-valued integrand can have a finite integral. Simple. The set where the integrand has infinite value has measure zero, and the conventions of probability theory say $0 \times \infty = 0$. We say the integrand in (79) is a *subprobability* density with respect to λ.

Since the set where $h \circ y$ is $-\infty$ has probability density zero, and the set where it is $+\infty$ has measure zero. The distribution having log density $h \circ y$ is concentrated on the set where it is finite, which we write as $(h \circ y)^{-1}(\mathbb{R})$.

If we start with a full exponential family, then the MLE in the completion is always a probability density (not sub) but the argument for that is long and complicated (Geyer, 1990, most of Chapter 2). But maximum likelihood in a non-full exponential family can give rise to subprobability MLE (Geyer, 1990, Examples 4.2 through 4.8).

9.4 Maximum Likelihood Estimation

Let $\mathcal{H}$ be a family of affine functions such that (68) holds, so (67) represents an exponential family. Then the log likelihood is given by

$$l(h) = h(y), \qquad h \in \mathcal{H},$$

where y is the observed value of the canonical statistic. Define

$$m(y) = \sup_{h \in \mathcal{H}} h(y), \qquad y \in B.$$

Call it the *log likelihood supremum function*. Then m is a convex function (a lower-semicontinuous, proper convex function Geyer, 1990, Theorem 4.2). Hence its effective domain

$$M = \{\, y \in B : m(y) < \infty \,\} \tag{80}$$

is a convex set. If the family we started with was full, then $\operatorname{cl} M$ is the convex support of $\lambda \circ y^{-1}$, but in any case $\operatorname{cl} M$ is a support of $\lambda \circ y^{-1}$ (Geyer, 1990, Theorem 4.2).

MLE always exist in the closure of the family. The space of generalized affine functions on a finite-dimensional affine space is sequentially compact in the topology of pointwise convergence (every sequence has a convergent subsequence) (Geyer, 1990, Section 3.3). Hence, for any $y \in B$, there is a sequence h_n in $\mathcal{H}$ such that $h_n(y) \to m(y)$ as $n \to \infty$. And this sequence

has a pointwise convergent subsequence, say $h_{n_k} \to \hat{h}$. Then $\hat{h}(y) = m(y)$, so $\hat{h}$ is an MLE, when y is the observed value of the canonical statistic.

Such an MLE may be a subprobability density and need not be unique. In a non-full family, non-uniqueness is typical. Consider the binomial model having only two distributions, with usual parameter π either $1/4$ or $3/4$. And suppose we observe $y = n/2$ (which requires even n). Then the MLE in the full family is $\hat{\pi} = 1/2$, but the MLE in the non-full family is either $1/4$ or $3/4$ or both (if you don't mind set-valued estimators). If you want non-full, you have to accept non-unique. As mentioned above, Examples 4.2 through 4.8 in Geyer (1990) show MLE that are subprobability distributions.

A subprobability distribution may even be the zero measure (that gives measure zero to every event). This will always be the case when the observed data y has $m(y) = +\infty$. So this is even more unsatisfactory than other subprobability MLE.

The closure of a full exponential family (considered as in this section) is in a sense (more on exactly what sense later) a union of exponential families. This is why Brown (1986) calls it an *aggregate exponential family*.

9.5 Faces of Convex Sets

A nonempty *face* of a convex set M is the set of points where some generalized affine function h achieves its supremum over M and that supremum is finite (Geyer, 1990, Theorem 3.9). The empty set is always a face too (by definition). And M is always a face, (the set where constant affine functions achieve their maximum over M).

The M we are interested in is the set (80) where the supremum of the log likelihood is finite. Let $\mathcal{F}$ be the family of nonempty faces of M.

9.5.1 Aggregate Exponential Families

Suppose now that $\mathcal{H}$ in (67) is a family of generalized affine functions closed in the topology of pointwise convergence (so the object of the category is now a generalized exponential family of subprobability measures).

For $F \in \mathcal{F}$ define

$$\mathcal{H}_F = \{\, h \in \mathcal{H} : h^{-1}(\mathbb{R}) = \text{aff } F \text{ and } \textstyle\int e^{h \circ y} \, d\lambda = 1 \,\}$$

Then the probability distributions having log density $h \in \mathcal{H}_F$ with respect to $\lambda \circ y^{-1}$ are concentrated on aff F. So if we think of $\mathcal{H}_F$ as being a family of affine functions on aff F, it is itself an exponential family of distributions

$$(\Omega, \mathcal{A}, \lambda) \xrightarrow{\ y\ } \text{aff } F \xrightarrow{\ \mathcal{H}_F\ } \mathbb{R} \tag{81}$$

so long as $\mathcal{H}_F$ is nonempty (emptiness is possible). This (81) is what Geyer (2009) calls a *limiting conditional model* because one can obtain it by conditioning the original family on aff F rather than by taking limits. And the union of all the $\mathcal{H}_F$ (thought of as exponential families) for $F \in \mathcal{F}$ is what Brown (1986) calls an *aggregate exponential family*, what Geyer (1990) calls the *relative closure* of an exponential family, and what Geyer (2009) calls the *Barndorff-Nielsen completion* of an exponential family (although that really only makes sense for full families).

9.5.2 Aggregate Exponential Subprobability Families

If one allows non-full families, then (as was discussed above) one must also allow subprobability densities. Then one defines

$$\mathcal{H}_F = \{\, h \in \mathcal{H} : h^{-1}(\mathbb{R}) = \text{aff } F \,\}$$

(not worrying about whether densities integrate to one). Then, of course, the $\mathcal{H}_F$ can no longer be thought of as exponential families of probability distributions but rather (if nonempty) as exponential families of subprobability distributions. Such are the complications of non-full families.

9.5.3 Conclusion

This section has been a bit sketchy. But Chapters 3 and 4 of Geyer (1990) have all the details and are fully rigorous. We included it both for generality, and to show where the affine picture really shines. It would be much messier to try to explain everything discussed here using the vector picture (and horribly messier using the $\mathbb{R}^d$ picture).

References

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families*. Wiley, Chichester.

Billingsley, P. (1979). *Probability and Measure*. Wiley, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*, second edition. Wiley, New York.

Bourbaki, N. (1970). *Algèbre.* Book II of *Éléments de Mathématique.* New edition, Chapters 1–3. Hermann, Paris. English translation, Addison-Wesley, 1974.

Browder, A. (1996). *Mathematical Analysis: An Introduction*. Springer-Verlag, New York.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA.

Cramér, H. (1951). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

Geyer, C. J. (1990). *Likelihood and Exponential Families*. PhD thesis, University of Washington. `http://purl.umn.edu/56330`.

Geyer, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electronic Journal of Statistics*, **3**, 259–289.

Geyer, C. J. (2013). Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, G. L. Jones and X. Shen eds. IMS Collections, Vol. 10, pp. 1–24. Institute of Mathematical Statistics: Hayward, CA.

Geyer, C. J. and Møller, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, **21**, 359–373.

Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, **94**, 415–426.

Godement, R. (1963). *Cours d'Algèbre*. Hermann, Paris. English translation as *Algebra,* Hermann, 1968.

Halmos, P. (1974). *Finite Dimensional Vector Spaces*. Springer-Verlag, New York. Reprint of second edition (1958), originally published by Van Nostrand

Hankin, R. K. S. (2006). .Special functions in R: Introducing the `gsl` package. *R News*, **6**, 24–26. `https://journal.r-project.org/articles/RN-2006-030/`.

Hankin, R. K. S., Clausen, A., and Murdoch, D. (2023). R package `gsl`: Wrapper for the Gnu Scientific Library, version 2.1-8. `https://cran.r-project.org/package=gsl`.

Lang, S. (1993). *Real and Functional Analysis*, third edition. Springer-Verlag, New York.

Leinster, T. (2014). Rethinking set theory. *American Mathematical Monthly*, **121**, 403–415.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

Rockafellar, R. T., and Wets, R. J.-B. (1998). *Variational Analysis*. Springer-Verlag, Berlin. (The corrected printings contain extensive changes. We used the third corrected printing, 2010.)

Roman, S. (2008). *Advanced Linear Algebra,* third edition. Springer, New York.

Rudin, W. (1991). *Functional Analysis*, second edition. McGraw-Hill, Boston.

Talagrand, M. (2005). *The Generic Chaining*. Springer-Verlag, Berlin.