Structure and Equivalence

Neil Dewar

October 2, 2020

Die Mathematiker sind eine Art Franzosen: redet man zu ihnen, so übersetzen sie es in ihre Sprache, und dann ist es alsobald ganz etwas anders.

Goethe

Contents

Introduction

This is a book about structure in the representations of physics, about equivalence between such representations, and about the relationship between these two concepts. In a slogan, that relationship is as follows: for two representations to be equivalent is for them to posit the same structure, and the structure of a representation is that which it has in common with equivalent representations. The question of which half of this slogan is primary—that is, whether we should take equivalence to be a derivative notion from structure, or the other way round—is the animating question behind much of what follows. That said, the main aim of this book is not to answer that question, but to introduce students to the tools and ideas that can, I think, be useful in seeking to answer it.

The book is split into four parts, with each part comprising three chapters (with each of the twelve chapters being roughly the same length). Part I looks at issues of structure and equivalence in the context of formal, logical languages. Chapters 1 and 2 introduce notions of definability and translation with regards to (respectively) models and theories of first-order languages, and how these notions can be used to make precise ideas about equivalence, while Chapter 3 looks at whether Ramsey sentences provide a plausible way of explicating the structure of such a theory. The goal is to lay some ideas on the table, in the (admittedly artificial) context of formal languages, that we can use in studying the more physics-oriented structures introduced subsequently. This part of the book presumes familiarity with standard predicate logic.

Parts II and III engage most directly with questions of structure and equivalence as they arise within physics, especially as concerns symmetries in physics. My original intention was to write something much more general about symmetry and equivalence in physics, with appositely chosen case studies to illustrate those general lessons. However, this ambition foundered on three problems: a strict word-count, the desire to make this book even remotely pedagogically accessible, and the inverse relationship between tractability and generality. So instead, I offer two case studies, and leave it to the reader to consider how the lessons drawn from them might (or might not) generalise to other theories.

Thus, Part II is about N-particle Newtonian mechanics, and how its spacetime symmetries can be used as a way of recognising the presence of 'surplus structure' in representations of this theory: Chapter 4 introduces the theory and its symmetries, Chapter 5 discusses the reasons for thinking symmetry-variant structure is surplus structure, and Chapter 6 outlines how to formulate the theory without using such structure. Part III takes up these questions about symmetry and surplus structure, but applied to the symmetries of electromagnetism. These symmetries include both its spacetime symmetries (in Chapter 7) and its internal gauge symmetries (in Chapters 8 and 9). These parts should be mostly accessible to readers provided they have some undergraduate-level knowledge of physics, although some of the tools used are more abstract than one would typically find in an undergraduate physics course. The appendices—on vector and affine spaces, group theory, and differential forms—provide a guide to these tools, although one more suitable for reference or refreshment than introduction.

Finally, Part IV discusses the use of category-theoretic tools to study structure and equivalence: Chapters 10 and 11 introduce (respectively) categories and functors, and Chapter 12 describes how to apply them to categories formed from theories—including both the formal logical theories discussed in Part I, and the physical theories discussed in Parts II and III. So, this part of the book also seeks to bring together the ideas articulated in the earlier parts of the book. It is also perhaps the part of the book that will be most novel to students, at least to those coming from a philosophy or physics background; I hope, however, that it illustrates how the basic concepts, at least, are more readily understandable than one might have expected.

[TO ADD: ACKNOWLEDGMENTS AND THANKS]

Munich, October 2020

Part I.

Logic

1. Models

We begin our investigations by looking at the structure of, and equivalence relations between, models of first-order logic. This is a highly stylised context, especially if our ultimate goal is an inquiry into structure and equivalence in physics: it is a commonplace that for most theories in physics, a presentation of those theories in first-order logic is likely neither possible nor desirable. Nevertheless, as we will see, the first-order case will prove to have complexities enough for us to start with; and it will teach us some lessons that we can use when we turn to physics in the later chapters of this book.

1.1. Review of first-order semantics

We begin with a brief review of the terminology and notation of standard first-order model theory.[1] The foundational concept here is that of a *first-order language*, which consists of a logical vocabulary (common to all first-order languages) and a non-logical vocabulary or signature (different for each first-order language). The logical vocabulary comprises a set Var of *variables*, $x_1, x_2, \ldots, y_1, y_2, \ldots, z_1, z_2, \ldots$; the *equality* symbol $\ulcorner = \urcorner$; the *negation*, *conjunction*, *disjunction* and *implication* symbols $\ulcorner \neg \urcorner$, $\ulcorner \wedge \urcorner$, $\ulcorner \vee \urcorner$ and $\ulcorner \rightarrow \urcorner$; and the *universal and existential quantifiers* $\ulcorner \forall \urcorner$ and $\ulcorner \exists \urcorner$.

In general, a signature consists of both predicate-symbols and function-symbols. However, we will confine ourselves to signatures that only contain predicate-symbols: many of the notions in which we are interested (concerning definability and translatability) are much easier to handle when we exclude function-symbols, and those same notions demonstrate that any theory employing function-symbols is, in a certain sense, equivalent to a theory that uses only relation-symbols. In an ideal world we would have the space to investigate and discuss this notion of equivalence; but in this (as in so many respects), the world is far from ideal. Thus:

[1] Much of the notation and conventions follow Hodges (1997).

Definition 1. A *signature* consists of a set Σ of *predicate-symbols* (denoted by letters such as P, Q, R, etc.), each of which is associated with a natural number known as its *arity*. ♠

So unary predicate-symbols (those which take a single argument) have an arity of 1; binary predicate-symbols (those which take two arguments, also known as binary relation-symbols) have an arity of 2; and so on. Where appropriate, the arity of a symbol will be indicated by adding a parenthetical superscript: to introduce R as a binary predicate-symbol, for instance, we will write its first appearance as $R^{(2)}$.

Given a signature Σ, one can define the set $\mathrm{Form}(\Sigma)$ of well-formed Σ-*formulae*, using the standard compositional rules of predicate logic. A variable in a formula is *free* if it is not bound by any quantifier; we will use $\phi(\xi_1, \ldots, \xi_n)$ to denote a formula with the variables $\xi_1, \ldots, \xi_n$ free, and $\phi(\eta_1/\xi_1, \ldots, \eta_n/\xi_n)$ to denote the result of uniformly substituting η_i for ξ_i throughout such a formula. The set of Σ-*sentences* is the set of closed Σ-formulae (formulae with no free variables).

The semantics for first-order model theory is given by *Tarski-models*. A Tarski-model $\mathfrak{A}$ for a language with signature Σ will be referred to as a Σ-*model*:[2]

Definition 2. A Σ-*model* consists of a set $|\mathfrak{A}|$ (the *domain* of $\mathfrak{A}$), equipped with a subset $\Pi_{\mathfrak{A}} \subseteq |\mathfrak{A}|^n$ (the *extension* of Π in $\mathfrak{A}$) for every $\Pi \in \Sigma$. ♠

A Tarski-model $\mathfrak{A}$ determines truth-values for formulae, relative to an assignment of elements of $|\mathfrak{A}|$ to variables in Var, in the standard recursive fashion. If the formula ϕ has the variables $x_1, \ldots, x_n$ free, and if $\mathfrak{A}$ satisfies ϕ relative to the assignment of $a_i \in |\mathfrak{A}|$ to x_i, then we write $\mathfrak{A} \models \phi[a_1, \ldots, a_n]$. If ϕ is a sentence, then the variable-assignment no longer matters, and we write simply $\mathfrak{A} \models \phi$.

1.2. Relationships between models

Most of the above should be familiar if you have taken a standard logic course before. Here, however, we are interested in exploring the use of these ideas to make precise concepts of structure and equivalence. For these purposes, it is very useful to start thinking about the kinds of relationships that Tarski models can bear to one another. First, for a given signature Σ, a *homomorphism* from a Σ-model $\mathfrak{A}$ to another Σ-model $\mathfrak{B}$ is a map which, in a certain sense, maps the structure of $\mathfrak{A}$ onto that of $\mathfrak{B}$. Formally:

[2] In the definition below, $|\mathfrak{A}|^n$ is the n-fold Cartesian product of $|\mathfrak{A}|$ with itself: that is, the set of ordered n-tuples of elements of $|\mathfrak{A}|$.

Definition 3. Let $\mathfrak{A}$ and $\mathfrak{B}$ be two Σ-models. A *homomorphism* $h : \mathfrak{A} \to \mathfrak{B}$ is a function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for every predicate-symbol $\Pi^{(n)} \in \Sigma$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\text{If } \langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}}, \text{ then } \langle h(a_1), \ldots, h(a_n) \rangle \in \Pi_{\mathfrak{B}} \tag{1.1}$$

♠

When two models are homomorphic to one another, there is a certain kind of 'structural resemblance' between them—but it's reasonably weak. However, we can strengthen it in successive degrees. First, the definition of homomorphism requires only a left-to-right implication. If we also require the right-to-left implication to hold, and require that it hold of all atomic formulae (including those using the equality-symbol), then we obtain the notion of *embedding*:[3]

Definition 4. An *embedding* $h : \mathfrak{A} \to \mathfrak{B}$ is an injective function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for every predicate-symbol $\Pi^{(n)} \in \Sigma$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}} \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in \Pi_{\mathfrak{B}} \tag{1.2}$$

♠

Finally, recall that $f : X \to Y$ is *surjective* if for any $y \in Y$, there is some $x \in X$ such that $f(x) = y$; a function which is both injective and surjective is *bijective*. This enables us to state the strongest kind of relationship between models that we will be interested in:

Definition 5. An *isomorphism* $h : \mathfrak{A} \to \mathfrak{B}$ is a surjective embedding. That is, it is a bijective function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for every predicate-symbol $\Pi^{(n)} \in \Sigma$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}} \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in \Pi_{\mathfrak{B}} \tag{1.3}$$

♠

An isomorphism between a Tarski-model and itself is known as an *automorphism*. For any Tarski-model, the identity map on its domain is an automorphism; but many Tarski-models also possess 'non-trivial' automorphisms, i.e., automorphisms which are not the identity map.

The notion of isomorphism seems to naturally capture a notion of 'structural identity', that is, of what it is for two Tarski-models to have 'the same structure'. (Note that

[3] Recall that a function $f : X \to Y$ is *injective* if for any $x_1, x_2 \in X$, if $x_1 \neq x_2$ then $f(x_1) \neq f(x_2)$.

I only say it captures *a* notion of structural identity; as we shall explore, there are other ways of capturing that idea that we should also investigate.) After all, at least intuitively: from the existence of a bijective function, we can infer that the domains of the two models have the same number of elements; and the condition (1.3) indicates that the extensions of the various symbols in Σ are 'distributed' over those elements in the same way.

As a result, if a given formula ϕ holds of a certain n-tuple in $\mathfrak{A}$, then that same formula ϕ will hold of that n-tuple's image under h in $\mathfrak{B}$. In other words, an isomorphism preserves the satisfaction of formulae. However, the converse to this is not true: it is possible for a homomorphism to preserve the satisfaction of formulae without being an isomorphism. Such a homomorphism is known as an *elementary embedding*; formally,

Definition 6. An *elementary embedding* $h : \mathfrak{A} \to \mathfrak{B}$ is a function $h : |\mathfrak{A}| \to |\mathfrak{B}|$ such that for any n-place Σ-formula ϕ and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\text{If } \mathfrak{A} \models \phi[a_1, \ldots, a_n], \text{ then } \mathfrak{B} \models \phi[h(a_1), \ldots, h(a_n)] \tag{1.4}$$

♠

The notion of elementary embedding is also of great model-theoretic significance: indeed, as we will discuss in Chapter 12, there are good reasons for thinking that elementary embeddings, rather than homomorphisms, should be thought of as the structure-preserving mappings in the context of model theory.

However, we can also think about relationships between models of different signatures; indeed, such relationships will be our main concern in the remainder of this chapter. For example, suppose that we have two signatures Σ and Σ^+, such that $\Sigma \subset \Sigma^+$. Then, given any Σ^+-model $\mathfrak{A}$, the *reduct* of $\mathfrak{A}$ to Σ is, intuitively, what we get by 'forgetting' the extensions of all those predicate-symbols that are in Σ^+ but not in Σ. More formally,

Definition 7. Let $\Sigma \subset \Sigma^+$, and let $\mathfrak{A}$ be a Σ^+-model. The *reduct* of $\mathfrak{A}$ to Σ is denoted by $\mathfrak{A}_\Sigma$, and is defined as follows: the domains are identical (i.e. $|\mathfrak{A}_\Sigma| = |\mathfrak{A}|$), and for any $\Pi \in \Sigma$,

$$\Pi_{\mathfrak{A}_\Sigma} = \Pi_{\mathfrak{A}} \tag{1.5}$$

♠

The converse notion to reduct is that of *expansion*.

Definition 8. Let $\Sigma \subset \Sigma^+$, let $\mathfrak{A}$ be a Σ^+-model, and let $\mathfrak{B}$ be a Σ-model. $\mathfrak{A}$ is an *expansion* of $\mathfrak{B}$ to Σ^+ if $\mathfrak{A}_\Sigma = \mathfrak{B}$. ♠

1.3. Definability

I remarked earlier that isomorphism provides a certain natural sense of structural identity. On this basis, we might argue that two mathematical representations should be thought of as possessing the same structure just in case they are isomorphic to one another. But taken literally, this criterion is far too restrictive. For, strictly speaking, no two models of different signatures can be isomorphic to one another. So, for example, a strict linear order with five elements that represents the order relation using the symbol $<$ is not isomorphic to a strict linear order with five elements that represents the order relation using the symbol $\prec$. But it would be very strange to think of these two models as having different structures—the difference between them is merely notational.

However, this might seem an unduly uncharitable construal of the proposal. When philosophers talk about isomorphism, they often seem to have a less literal understanding of the notion of isomorphism: one which requires only that the number and distributions of the extensions over the models are the same, independently of what those extensions are labelled. Following Lutz (2015), we will refer to this more liberal notion as *H-isomorphism*:[4]

Definition 9. Let $\mathfrak{A}$ be a Σ_1-model, and let $\mathfrak{B}$ be a Σ_2-model. An *H-isomorphism* consists of a bijection $h : \mathfrak{A} \to \mathfrak{B}$ and a bijection $k : \Sigma_1 \to \Sigma_2$ such that for any $\Pi^{(n)} \in \Sigma_1$ and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\mathfrak{A}} \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in k(\Pi)_{\mathfrak{B}} \tag{1.6}$$

♠

Nevertheless, a little reflection suggests that this is still too restrictive a notion of isomorphism. Consider first the standard model of the natural numbers, equipped with extensions for zero, successor, addition, and multiplication. Call this model $\mathfrak{M}$. Now consider the standard model of the natural numbers, equipped with extensions for zero, successor, addition, multiplication, and *evenness*. Call this model $\mathfrak{N}$. $\mathfrak{M}$ and $\mathfrak{N}$ are not H-isomorphic: there is no bijection between their signatures, since those contain four and five symbols respectively. And yet, it is natural to feel that $\mathfrak{M}$ and $\mathfrak{N}$ have the same structure. After all, it is not as though the notion of evenness is somehow 'missing' in $\mathfrak{M}$, just because $\mathfrak{M}$ does not come equipped with a special label for it. The only difference, we want to say, between $\mathfrak{M}$ and $\mathfrak{N}$ is that some piece of structure which is implicitly present in $\mathfrak{M}$ has been bestowed with a specific name in $\mathfrak{N}$.

[4] So-called since it plays a role in Halvorson (2012)'s argument against the semantic view of theories.

We can make this notion of 'implicit structure' precise through the concept of *definability*. Intuitively, a certain collection of elements, or of tuples, is definable if it consists of precisely those elements that match a certain description. More formally:

Definition 10. Let $\mathfrak{A}$ be a Σ-structure. A set $X \subseteq |\mathfrak{A}|^n$ is *definable in* $\mathfrak{A}$ if there is some Σ-formula $\phi(x_1, \ldots, x_n)$ such that

$$\langle a_1, \ldots, a_n \rangle \in X \text{ iff } \mathfrak{A} \models \phi[a_1, \ldots, a_n] \tag{1.7}$$

♠

For example, the set of even numbers is definable in $\mathfrak{M}$, being definable by the (one-place) formula

$$\exists y(y + y = x) \tag{1.8}$$

If we accept the idea that the definable sets should be considered an (implicit) part of a Tarski-model's 'structure', then this also suggests regarding two models as having the same structure when the extensions of one model are definable in the other. We make this precise via the notion of *codetermination* of models:[5]

Definition 11. Let $\mathfrak{A}$ be a Σ-model and let $\mathfrak{B}$ be a T-model. $\mathfrak{A}$ and $\mathfrak{B}$ are *codeterminate* if:

- $|\mathfrak{A}| = |\mathfrak{B}|$;
- for every $\Pi \in \Sigma$, $\Pi_{\mathfrak{A}}$ is definable in $\mathfrak{B}$; and
- for every $\Omega \in \mathrm{T}$, $\Omega_{\mathfrak{B}}$ is definable in $\mathfrak{A}$.

♠

We can shed further light on definability by reflecting on its relationship to *invariance*. A set of elements (or tuples) in a Tarski-model is invariant if it is 'fixed' by all automorphisms of that model, that is:

Definition 12. Let $\mathfrak{A}$ be a Σ-structure. A set $X \subseteq |\mathfrak{A}|^n$ is *invariant in* $\mathfrak{A}$ if, for any automorphism $h : \mathfrak{A} \to \mathfrak{A}$, and any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in X \text{ iff } \langle h(a_1), \ldots, h(a_n) \rangle \in X \tag{1.9}$$

♠

[5] See Barrett (nd), Winnie (1986).

Thus, for example, the extension of any predicate in Σ is guaranteed to be invariant (by the definition of automorphism). More generally, it turns out that the extension of any formula whatsoever is invariant: that is,

Theorem 1. *Let $\mathfrak{A}$ be a Σ-structure. For any set $X \subseteq |\mathfrak{A}|^n$, if X is definable then X is invariant.*

Proof. Left as exercise. □

However, the converse is not true: not all invariant sets are definable. For example, consider again the natural numbers (whether in the form of the model $\mathfrak{M}$ or the model $\mathfrak{N}$). This model is *rigid*, in that it possesses no non-trivial automorphisms. As a result, every subset in the domain is invariant. Since there are $\aleph_0$-many natural numbers, there are $2^{\aleph_0}$-many such subsets. But the signature is finite, and so there can only be at most $\aleph_0$-many formulae (since each formula is itself a finite construction); and hence, only at most $\aleph_0$-many definable subsets. That said, although in general a model might contain sets that are both indefinable and invariant, there is a partial result:

Theorem 2. *Let $\mathfrak{A}$ be a* finite *Σ-structure. For any set $X \subseteq |\mathfrak{A}|^n$, if X is invariant then X is definable.*

Proof. Left as exercise. □

Note a corollary: if $\mathfrak{A}$ is finite and rigid, then every subset of $|\mathfrak{A}|$ (or of $|\mathfrak{A}|^n$, for any n) is definable.

2. Theories

In the previous section, we looked at structure and equivalence between Tarski-models. In this section, we turn our attention to these concepts within the realm of *theories*. 'Theory' here will be meant in the usual sense of model theory:

Definition 13 (First-order theory)**.** Let Σ be a signature. A Σ-*theory* T is a set of Σ-sentences. ♠

Definition 14 (Model of a theory)**.** Let T be a Σ-theory. A Σ-model $\mathfrak{A}$ is a *model of* T if, for every sentence $\phi \in T$, $\mathfrak{A} \models \phi$. ♠

Definition 15 (Consequence)**.** Let T be a Σ-theory. A Σ-sentence ϕ is a *consequence* of T if, for every model $\mathfrak{A}$ of T, $\mathfrak{A} \models \phi$. In such a case, we say that T *entails* ϕ, and write $T \vDash \phi$; if ψ is a consequence of $\{\phi\}$, then we write $\phi \vDash \psi$.[1] ♠

We will denote the class of models of T by $\mathrm{Mod}(T)$. Within philosophy of science, there has been a great deal of discussion of the respective merits of the *syntactic view of theories* (roughly, that theories are sets of sentences) and the *semantic view of theories* (roughly, that theories are classes of models). The definition of a theory as a set of sentences is not intended to take a side on this debate, which may not be trading on quite such a sharp distinction as its protagonists suppose.[2] Any set of sentences brings a class of models in its wake, so any theory on the syntactic view will correspond to some theory on the semantic view; and although not any class of Tarski-models will be the class of models of some theory, many of the most interesting such classes are.[3]

[1] Note that the symbols for satisfaction and consequence are unfortunately similar: the former is $\models$, whilst the latter is $\vDash$. The easiest way to distinguish them is to look at what is on the left-hand-side of the symbol: if it is a Σ-model, then the relation is satisfaction; if it is a theory (or sentence), the relation is consequence.

[2] See, for instance, Lutz (2015).

[3] More specifically, if a class K of Σ-models is closed under isomorphism (i.e. if a model $\mathfrak{A}$ is in K, then so is any model isomorphic to $\mathfrak{A}$) and also closed under both 'ultraproducts' and 'ultraroots' (whose definition is too complex for a footnote), then there is some Σ-theory T such that $K = \mathrm{Mod}(T)$. See (Hodges, 1993, §9.5).

2.1. Translations between theories

What is it for two theories to be equivalent, i.e., to posit the same structure? The strictest criterion of equivalence that one might consider is that of *identity*: two theories are equivalent if they consist of the same sentences. Although this is surely a sufficient condition for equivalence, it seems overly restrictive. For instance, this criterion would consider the theories $\{\exists x Px\}$ and $\{\exists x \neg\neg Px\}$ to be inequivalent.[4]

A more relaxed condition is that of *logical equivalence*. For our purposes, this means having the same models:

Definition 16 (Logical equivalence)**.** Let T_1 and T_2 be Σ-theories. T_1 and T_2 are *logically equivalent* if $\mathrm{Mod}(T_1) = \mathrm{Mod}(T_2)$: that is, if for every Σ-model $\mathfrak{A}$, $\mathfrak{A}$ is a model of T_1 iff $\mathfrak{A}$ is a model of T_2. ♠

There's a natural relationship between isomorphism (as a criterion of equivalence between models) and logical equivalence (as a criterion of equivalence between theories), expressed by the following proposition.

Proposition 1. Let T_1 and T_2 be Σ-theories. T_1 and T_2 are logically equivalent iff for every model $\mathfrak{A}_1$ of T_1, there is an isomorphic model $\mathfrak{A}_2$ of T_2.

Proof. Left as exercise. □

In the previous section, we discussed the weaker notion of H-isomorphism, as a less language-dependent version of isomorphism. The corresponding notion for theories would be two theories that are logically equivalent 'up to a choice of notation'; we shall say that two theories related in this fashion are *notational variants* of one another.

Definition 17. Let T_1 be a Σ_1-theory, and let T_2 be a Σ_2-theory. T_1 and T_2 are *notational variants* of one another if there is an arity-preserving bijection $k : \Sigma_1 \to \Sigma_2$ such that $k(T_1)$ is logically equivalent to T_2, where $k(T_1)$ is the result of replacing every occurrence of any $P \in \Sigma_1$ in T_1 with $k(P)$. ♠

However, as a criterion of equivalence, notational variance is still very strict. In the previous chapter, we considered a further weakening from H-isomorphism, namely codetermination. This motivates us to consider a weaker kind of relationship between theories: that we can offer a *translation* between them.[5]

[4] If we had required that a theory be a set of sentences *closed under entailment* (so that if $T \vDash \phi$ then $\phi \in T$), then the identity criterion would coincide with the criterion of logical equivalence.

[5] For more detail on translations between theories, see (Halvorson, 2019, chap. 4).

The basic idea of translating one theory into another is that we can systematically replace expressions of the first theory's language by expressions of the second theory's language, in such a way that all theorems of the first theory are converted into theorems of the second theory. More precisely,

Definition 18 (Translation between theories)**.** Let T_1 be a Σ_1-theory, and T_2 a Σ_2-theory. A *translation* from T_1 to T_2 is a map $\tau : \mathrm{Form}(\Sigma_1) \to \mathrm{Form}(\Sigma_2)$, which:

1. Preserves variables: if the Σ_1-formula ϕ has exactly the variables $\xi_1, \ldots, \xi_n$ free, then $\tau(\phi)$ has exactly $\xi_1, \ldots, \xi_n$ free.

2. Commutes with substitution: for any Σ_1-formula ϕ with the variables $\xi_1, \ldots, \xi_n$ free, and any variables $\eta_1, \ldots, \eta_n$,

$$\tau(\phi(\eta_1/\xi_1, \ldots, \eta_n/\xi_n)) = \tau(\phi)(\eta_1/\xi_1, \ldots, \eta_n/\xi_n) \tag{2.1}$$

3. Commutes with the logical connectives: for any Σ_1-formulae ϕ and ψ, and any variable ξ,

$$\tau(\neg\phi) = \neg\tau(\phi) \tag{2.2}$$
$$\tau(\phi \wedge \psi) = \tau(\phi) \wedge \tau(\psi) \tag{2.3}$$
$$\tau(\forall\xi\phi) = \forall\xi\tau(\phi) \tag{2.4}$$

   etc.

4. Preserves consequence: for any Σ_1-formula ϕ,

$$\text{If } T_1 \vDash \phi \text{ then } T_2 \vDash \tau(\phi) \tag{2.5}$$

♠

When τ is a translation from T_1 to T_2, we will write $\tau : T_1 \to T_2$. Since a translation is required to commute with substitution and the logical connectives, we can specify such a translation just by specifying, for every $\Pi^{(n)} \in \Sigma_1$, how to translate $\Pi x_1 \ldots x_n$. In what follows, this is how we will usually specify translations.

How does the existence of a translation from one theory to another relate to the structures posited by the two theories? We can get some insight here by reflecting on how it is reflected in the relationships between the theories' classes of models. The key observation here is that a translation τ from one theory to another induces a 'dual map' τ^*

from the models of the latter theory to those of the former (so the dual map goes 'in the other direction' from the translation).

Definition 19 (Dual map to a translation)**.** Let τ be a translation from the Σ_1-theory T_1 to the Σ_2-theory T_2. Given any Σ_2-model $\mathfrak{A}$, we define the Σ_1-model $\tau^*(\mathfrak{A})$ as follows. First, the domain of $\tau^*(\mathfrak{A})$ is the same as that of $\mathfrak{A}$, that is, $|\tau(\mathfrak{A})| = |\mathfrak{A}|$. Second, for any $\Pi^{(n)} \in \Sigma_1$, we define the extension of Π in $\tau^*(\mathfrak{A})$ as follows: for any $a_1, \ldots, a_n \in |\mathfrak{A}|$,

$$\langle a_1, \ldots, a_n \rangle \in \Pi_{\tau^*(\mathfrak{A})} \text{ iff } \mathfrak{A} \models \tau(\Pi)[a_1, \ldots, a_n] \tag{2.6}$$

For any model $\mathfrak{A}$ of T_2, $\tau^*(\mathfrak{A})$ is a model of T_1 (see Proposition 3 below). So τ^* is a function from $\mathrm{Mod}(T_1)$ to $\mathrm{Mod}(T_2)$, which we refer to as the *dual map* to the translation τ. ♠

To prove the claim used in this definition—that if $\mathfrak{A}$ is a model of T_2, then $\tau^*(\mathfrak{A})$ is a model of T_1—we need the following useful proposition.

Proposition 2. Let τ be a translation from the Σ_1-theory T_1 to the Σ_2-theory T_2. For any Σ_2-model $\mathfrak{A}$, and any Σ_1-sentence ϕ,

$$\tau^*(\mathfrak{A}) \models \phi \text{ iff } \mathfrak{A} \models \tau(\phi) \tag{2.7}$$

Proof. By induction on the length of formulae; left as exercise. □

Given this proposition, the proof that the dual map to a translation preserves model-hood is straightforward.

Proposition 3. Let τ be a translation from the Σ_1-theory T_1 to the Σ_2-theory T_2. If $\mathfrak{A}$ is a T_2-model, then $\tau^*(\mathfrak{A})$ is a T_1-model.

Proof. Suppose, for *reductio*, that $\tau^*(\mathfrak{A})$ is not a T_1-model. Then there must be some sentence $\phi \in T_1$ such that $\tau^*(\mathfrak{A}) \not\models \phi$. Then by Proposition 2, $\mathfrak{A} \not\models \tau(\phi)$. Since $\mathfrak{A}$ is a T_2-model, it follows that $T_2 \nvdash \tau(\phi)$. But by the definition of a translation, $T_2 \vdash \phi$; so by contradiction, $\tau^*(\mathfrak{A})$ must be a T_1-model. □

Now, let us consider the question of how the notion of translation could be used to articulate a criterion of equivalence. The mere existence of a translation (as defined here) would be a very weak condition, and would have some very counter-intuitive consequences: it would mean, for example, that any theory would be equivalent to any strictly stronger theory (since inclusions are always translations)—indeed, that any

theory is equivalent to some inconsistent theory! A more plausible criterion is to require the existence of a *pair* of translations. This criterion is known as *mutual interpretability*.

Definition 20. Let T_1 and T_2 be theories of signatures Σ_1 and Σ_2 respectively. T_1 and T_2 are *mutually interpretable* if there exist translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$. ♠

However, mutual interpretability is still a relatively weak notion, as the following examples indicate.

Example 1. Let $\Sigma_1 = \{P^{(1)}\}$, and let $\Sigma_2 = \{Q^{(1)}, R^{(1)}\}$. Let

$$T_1 = \varnothing \tag{2.8}$$

$$T_2 = \{\forall x(Qx \to Rx)\} \tag{2.9}$$

One would expect that T_1 and T_2 should not be regarded as equivalent: intuitively, T_2 says something non-trivial, whereas T_1 does not. Yet T_1 and T_2 are mutually interpretable, since

$$\tau(Px) = Qx \tag{2.10}$$

is a translation from T_1 to T_2, and

$$\sigma(Qx) = Px \tag{2.11}$$

$$\sigma(Rx) = Px \tag{2.12}$$

is a translation from T_1 to T_2.

In light of this, we introduce a yet stronger condition: not just that there exist a pair of translations, but that those translations be, in a certain sense, inverse to one another. The intuition here is that if we take some expression of our first theory's language, translate it into the second language, and then translate it back into the first language, we should—if the pair of translations really express an equivalence between the theories—get an expression with the same meaning as the expression with which we began. Formally, we cash out this condition of 'having the same meaning' as 'equivalent modulo the ambient theory'; the resulting criterion is known as *intertranslatability*.[6]

Definition 21. Let T_1 and T_2 be theories of signatures Σ_1 and Σ_2 respectively. T_1 and T_2 are *intertranslatable* if there exist translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$, such that for

[6] Barrett and Halvorson (2016a)

any Σ_1-formula $\phi(x_1, \ldots, x_n)$ and any Σ_2-formula $\psi(y_1, \ldots, y_m)$,

$$T_1 \vDash \forall x_1 \ldots \forall x_n(\phi \leftrightarrow \sigma(\tau(\phi))) \tag{2.13}$$

$$T_2 \vDash \forall y_1 \ldots \forall y_m(\psi \leftrightarrow \psi(\sigma(\psi))) \tag{2.14}$$

In such a case, we will say that τ and σ are *inverse translations* to one another. ♠

Where we are dealing with a pair of inverse translations of this kind, we will often express them by writing

$$\Pi x_1 \ldots x_n \equiv \tau(\Pi x_1 \ldots x_n) \tag{2.15}$$

for every $\Pi \in \Sigma_1$, and

$$\Omega y_1 \ldots y_m \equiv \sigma(\Omega y_1 \ldots y_m) \tag{2.16}$$

for every $\Omega \in \Sigma_2$. Thus, the symbol $\equiv$ will typically have expressions from two different languages on either side of it.

In general, if we have a theory T_1, then its image $\tau[T_1]$ under a map $\tau : \mathrm{Form}(\Sigma_1) \to \mathrm{Form}(\Sigma_2)$ is not intertranslatable with T_1, even if we suppose that τ preserves free variables, and that it commutes with substitution and the logical connectives.[7] However, if τ is 'suitably invertible', then this does hold. More precisely:

Proposition 4. Suppose that the translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$ are inverse to one another. Then T_2 is logically equivalent to $\tau[T_1]$, and T_1 is logically equivalent to $\sigma[T_2]$.

Proof. Suppose that there were some model $\mathfrak{B}$ of T_2 which was was not a model of $\tau[T_1]$. Then for some $\phi \in T_1$, $fB \not\models \tau(\phi)$; but then it follows that $T_2 \not\vDash \tau(\phi)$, which contradicts the assumption that τ is a translation. The other case is proven similarly. □

In particular, take as given some Σ_1-theory T_1, and suppose that $\tau : \mathrm{Form}(\Sigma_1) \to \mathrm{Form}(\Sigma_2)$ and $\sigma : \mathrm{Form}(\Sigma_2) \to \Sigma_1)$ preserve free variables, commute with substitution and the logical connectives. If $\sigma(\tau(\phi))$ is logically equivalent to ϕ and $\tau(\sigma(\psi))$ is logically equivalent to ψ (for every $\phi \in \mathrm{Form}(\Sigma_1)$ and $\psi \in \mathrm{Form}(\Sigma_2)$), then T_1 is intertranslatable with $\tau[T_1]$; this also holds if we have equivalence with respect to some background theory, rather than full logical equivalence. We will employ this observation in Part II.

Finally, we observe that intertranslatability is associated with codetermination between classes of models in a natural way.

[7] See Barrett and Halvorson (2016a).

Proposition 5. If $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$ are inverse translations, then:

- for any $\mathfrak{A} \in \mathrm{Mod}(T_1)$, $\mathfrak{A}$ is codeterminate with $\sigma^*(\mathfrak{A})$, and $\tau^*(\sigma^*(\mathfrak{A})) = \mathfrak{A}$; and
- for any $\mathfrak{B} \in \mathrm{Mod}(T_2)$, $\mathfrak{B}$ is codeterminate with $\tau^*(\mathfrak{B})$, and $\sigma^*(\tau^*(\mathfrak{B})) = \mathfrak{B}$.

Proof. Left as exercise. □

This suggests that intertranslatability is a fairly natural criterion for equivalence between theories. That said, we should bear in mind that all we have discussed here are criteria of *formal* equivalence: roughly, of two theories having the same form, independently of their content. However, merely being of the same form is manifestly insufficient for two theories to be equivalent in the full sense of 'saying the same thing'. For example, as Sklar (1982) famously observes, the two theories 'all lions have stripes' and 'all tigers have stripes' are intertranslatable but do not say the same thing. So we should bear in mind that formal criteria like those discussed in this chapter (and, to some extent, the whole of this book) can only be a partial guide to theoretical equivalence.

3. Ramsey sentences

We've now seen some of the ways in which we can use the resources of model theory to articulate different senses of equivalence between models and theories. In the course of doing this, I have made occasional remarks about how these different notions of equivalence might be thought to capture different notions of 'structure', in the sense that they point to different ways of understanding the claim that two models, or two theories, have the same structure. However, there is an alternative way of approaching the relationship between the notions of structure and equivalence: rather than using equivalences to reveal structure, one can seek to articulate a notion of 'structure' directly, and then use that to formulate a criterion of equivalence. In this section, we consider one well-known proposal for the 'structural content' of a theory: that this content can be identified with the theory's *Ramsey sentence*.

3.1. Second-order logic

As we shall see, the Ramsey sentence of a first-order theory is a second-order sentence; so we begin by reviewing the formalism of second-order logic. In second-order logic, we can—as people say—*quantify into predicate position*. Intuitively, this means that we can make quantified claims about properties (and relations): so rather than being limited to saying things like 'all whales are mammals', we can now say things like 'anything which is true of all mammals is true of all whales', or 'there are some properties which whales and dolphins both have'.

More formally, then, second-order logic is distinguished from first-order logic by having not only a stock Var of first-order variables $x, y, z, \ldots$, but also a stock VAR of second-order variables $X, Y, Z, \ldots$. Like predicates, every second-order variable has an associated arity $n \in \mathbb{N}$. And as with predicates, we will indicate the arity of a second-order variable (where helpful) by a parenthesised superscript, thus: $X^{(n)}$. The subset of VAR containing all the n-ary variables will be denoted VAR^n.

Other than this, the symbolic vocabulary of second-order logic is the same as that of first-order logic: we have the equality-symbol, the logical connectives, the quantifiers,

and a signature Σ consisting of predicates (of various arities). The rules for forming well-formed formulae are the same as for first-order logic, but with two additional clauses:

- If $X^{(n)} \in \text{VAR}$, and if $x_1, \ldots, x_n \in \text{Var}$, then $Xx_1 \ldots x_n$ is a formula
- If ψ is a formula, and $X \in \text{VAR}$, then $\forall X\psi$ is a formula

Respectively, these clauses tell us that the new variables can go into predicate position, and that they can be quantified over. As with the first-order case, we will use $\ulcorner\vee\urcorner$, $\ulcorner\to\urcorner$ and $\ulcorner\exists\urcorner$ as abbreviations (so $\exists X\psi$ abbreviates $\neg\forall X\psi$).

We now turn to the standard semantics of second-order logic.[1] As with the first-order case, we use Tarski-models: for signature Σ, a Σ-model $\mathfrak{A}$ consists of a set $\mathfrak{A}$ equipped with extensions for all predicates in Σ. Given a Σ-model $\mathfrak{A}$, a *second-order variable-assignment* G for $\mathfrak{A}$ consists of a map $g : \text{Var} \to |\mathfrak{A}|$, and for every $n \in \mathbb{N}$, a map $G^n : \text{VAR}^n \to \mathcal{P}(|\mathfrak{A}|^n)$. Here, $\mathcal{P}(|\mathfrak{A}|^n)$ is the *power set* of $|\mathfrak{A}|^n$, i.e. the set containing all subsets of $|\mathfrak{A}|^n$; thus, for any $\Xi^{(n)} \in \text{VAR}$, $G^n(\Xi)$ is some set of n-tuples from $|\mathfrak{A}|$.

Now let ϕ be some second-order Σ-formula, let $\mathfrak{A}$ be a Σ-model, and let G be a second-order variable-assignment. Truth is then defined in the same way as in the first-order case, but with two extra clauses (corresponding to the two new clauses for formulae):

- For any $\Xi^{(n)} \in \text{VAR}$ and any $\xi_1, \ldots, \xi_n \in \text{Var}$,

$$\mathfrak{A} \models_G X^{(n)}x_1 \ldots x_n \text{ iff } \langle g(x_1), \ldots, g(x_n)\rangle \in G^n(\Xi) \tag{3.1}$$

- $\mathfrak{A} \models_G \forall X^{(n)}\phi$ iff for all $A \subseteq |\mathfrak{A}|^n$, $\mathfrak{A} \models_{G^X_A} \phi$

where the variable-assignment G^X_A is defined by the condition that

$$G^X_A(Y) = \begin{cases} G(Y) & \text{if } Y \neq X \\ A & \text{if } Y = X \end{cases} \tag{3.2}$$

We then say that a sentence ϕ is true relative to a Tarski-model $\mathfrak{A}$ if, for every variable-assignment G over $\mathfrak{A}$, $\mathfrak{A} \models_G \phi$. In this case, we write $\mathfrak{A} \models \phi$. Consequence is defined and denoted as in the first-order case.

[1] See (Shapiro, 1991, §4.2) or Manzano (1996) for more detailed treatments.

3.2. Ramseyfication

We can now turn our attention to the Ramsey sentence itself. The intuitive idea is that the 'structural core' of a theory T will make the same structural claims about the world as T, but without committing itself to which properties or relations it is that instantiate that structure. Thus, if a theory says something like 'positively charged particles repel one another', the structural claim thereby expressed is merely 'there is a property, such that any two particles possessing that property will repel one another'. One might object that even this does not go far enough, since it still speaks of 'repulsion'; or, one might distinguish between charge and repulsion on the basis that the notion of repulsion, unlike that of positive charge, is associated with a direct empirical content. In the first instance we will take the latter attitude, since the former (more extreme) view can be recovered as a special case.

Thus, suppose that our non-logical vocabulary Σ is divided into two classes, Ω and Θ: intuitively speaking, we suppose that Ω is the collection of 'observational' predicates (like 'repulsion'), while Θ is the collection of 'theoretical' predicates (like 'positive charge'). Suppose further that the theory T we are interested in (which is formulated in Σ) consists only of finitely many sentences; without loss of generality, we can suppose that T consists of a single sentence.[2]

We first form a 'skeleton' theory T^*, by replacing all the theoretical predicates that occur in T by second-order variables (of the appropriate arity): that is, if only $R_1, \ldots, R_p \in \Theta$ occur in T, then

$$T^* = T[X_1/R_1, \ldots, X_p/R_p] \tag{3.3}$$

where for each i, X_i is of the same arity as R_i. We then form the Ramsey sentence of T by existentially quantifying over all of these predicates:

$$T^R = \exists X_1 \exists X_2 \ldots \exists X_p T^* \tag{3.4}$$

We take the signature of the Ramsey sentence to be Σ (even though the sentence itself only contains predicates from Ω).

We can now ask the question: how much of a theory's structure does the Ramsey sentence capture? To answer this, first define the *observational reduct* of any Σ-model $\mathfrak{A}$

[2] In principle, we could apply the Ramseyfication procedure to a theory which consisted of infinitely many sentences. However, we would need the second-order language of T^R to be an *infinitary* second-order language: if T contained κ-many sentences, and if λ-many predicates from Θ occur in T, then T^R must be in a language that permits κ-size conjunction, and which admits the introduction of λ-many second-order quantifiers. In order to not have to deal with these complications, we will suppose that the original theory is finite.

to be its reduct $\mathfrak{A}_\Omega$ to Ω. Second, let $\mathfrak{W}$ be a first-order model of signature Σ which is a faithful representation of the world: i.e., which has the observational and theoretical predicates distributed over its elements in just the way that the corresponding observational and theoretical properties are distributed over the objects of the world. If this formulation makes you uncomfortable (which it probably should), then just think of $\mathfrak{W}$ as a 'preferred model', without worrying about in virtue of what it is preferred. We'll say that a Σ-theory is *true* if $\mathfrak{W}$ is one of its models; that it is *observationally adequate* if it has some model whose observational reduct is identical to $\mathfrak{W}_\Omega$; and that it is *numerically adequate* if it has some model whose domain coincides with $|\mathfrak{W}|$. Intuitively: a theory which is true admits a model which matches the actual number of objects, and the actual distribution of observational and theoretical properties over those objects; a theory which is observationally adequate admits a model which matches the actual number of objects, and the actual distribution of observational properties over those objects; and a theory which is numerically adequate admits a model which matches the actual number of objects[3].

This enables us to now make the following observation: for any Σ-theory T, its Ramsey sentence T^R is true just in case T is observationally adequate.[4] More formally:

Proposition 6. Let T be a theory of signature Σ, and let T^R be the Ramsey sentence of T. Then $\mathfrak{W} \models T^R$ if and only if T is observationally adequate (i.e., there is some model $\mathfrak{A}$ of T such that $\mathfrak{A}_\Omega = \mathfrak{W}_\Omega$).

Proof. First, suppose that $\mathfrak{W} \models T^R$: that is, that

$$\mathfrak{W} \models \exists X_1 \dots \exists X_p T^* \tag{3.5}$$

This is true just in case there is some second-order variable-assignment G for $\mathfrak{W}$ such that

$$\mathfrak{W} \models_G T^* \tag{3.6}$$

[3] Bear in mind here that $\mathfrak{W}$ is merely supposed to be a 'faithful representative' of the world, not 'the world itself' (whatever, exactly, these terms might mean). So it's harmless to define observational adequacy as the theory admitting a model identical to $\mathfrak{W}_\Omega$ (not just isomorphic to it), and to define numerical adequacy as the theory admitting a model with the same domain identical to $|\mathfrak{W}|$ (not just equinumerous with it)

[4] (Ketland, 2004, Theorem 2)

But now consider the model $\mathfrak{A}$, defined as follows:

$$|\mathfrak{A}| = |\mathfrak{W}|$$
$$P^{\mathfrak{A}} = P^{\mathfrak{W}}, \text{ for every } P \in \Omega$$
$$R_i^{\mathfrak{A}} = G(X_i), \text{ for every } R_i \in \Theta$$

A proof by induction shows that

$$\mathfrak{A} \models T \tag{3.7}$$

But by construction, $\mathfrak{A}_\Omega = \mathfrak{W}_\Omega$. So T is observationally adequate.

Second, suppose that T is observationally adequate: i.e., that there is some model $\mathfrak{A}$ of T such that $\mathfrak{A}_\Omega = \mathfrak{W}_\Omega$. Consider any second-order variable-assignment G for $\mathfrak{A}$ satisfying the condition that for every $R_i \in \Theta$,

$$G(X_i) = R_i^{\mathfrak{A}} \tag{3.8}$$

Since $|\mathfrak{A}| = |\mathfrak{W}|$, we can regard G as a variable-assignment for $\mathfrak{W}$. Then, again, a proof by induction shows that

$$\mathfrak{W} \models_G T^* \tag{3.9}$$

from which it follows immediately that $\mathfrak{W} \models T^R$. □

We also have the following corollary, which applies to the more radical view canvassed above (that the use of *all* predicates, not just the 'theoretical' ones, should be converted to existential quantifications).

Corollary 1. Suppose that $\Theta = \varnothing$ (equivalently, that $\Sigma = \Omega$); that is, consider the case where we Ramsefy away *all* the vocabulary. Then $\mathfrak{W} \models T^R$ if and only if T is numerically adequate (i.e., there is some model $\mathfrak{A}$ of T such that $|\mathfrak{A}| = |\mathfrak{W}|$).

Philosophically, this observation is usually taken as a problem for the proposal that a theory's structure is captured by its Ramsey sentence: simply put, the concern is that Proposition 6 shows that the Ramsey sentence fails to capture anything about a theory beyond its empirical or observational content.[5] So if we do indeed take the 'structure' of a theory to be that which is captured by its Ramsey sentence, then we appear to

[5] In the literature, this objection is referred to as 'Newman's objection', since a version of this objection was discussed in Newman (1928). (Note that this is *before* the introduction of the Ramsey sentence in Ramsey (1931)—even allowing for the fact that Ramsey's essay was written in 1929. The reason for this is that Newman's objection was, originally, offered as a criticism of Russell (1927), and only later applied to the Ramsey-sentence approach to theories.)

have the corollary that a theory simply has no non-observational structure. Moreover, there is something faintly paradoxical to this, insofar as the observational predicates were precisely the ones that we did not Ramseyfy. So the Ramsey-sentence approach to structure seems to hold that although the Ramsey sentence of a theory articulates that theory structure, the only structure a theory in fact possesses is expressed by that part of the theory's language which is not subject to Ramseyfication!

If the Ramsey sentence is taken as expressing a theory's structure, then it is natural to take two theories as equivalent if they have logically equivalent Ramsey sentences. A further way of thinking about Newman's objection is to observe that, if we use standard second-order semantics, then this criterion of equivalence degenerates into observational equivalence. More precisely, let us say that two first-order Σ-theories, T_1 and T_2, are *observationally equivalent* if for every model $\mathfrak{A}$ of T_1, there is some model $\mathfrak{B}$ of T_2 such that $\mathfrak{A}_\Omega$ is isomorphic to $\mathfrak{B}_\Omega$, and vice versa. Then:

Proposition 7. T_1 and T_2 are observationally equivalent if and only if, with respect to standard second-order semantics, their Ramsey sentences are logically equivalent.

Proof. Left as exercise. □

What can be said in response? One option is to bite the bullet, and argue that—in fact—it is *good* that a theory's structure should turn out to be exhausted by its observational structure. In other words, the Ramsey sentence can be regarded as a useful vehicle for expressing a (fairly strong) form of empiricism about scientific theories: it offers one way of making precise the idea that the real content of a scientific theory is its observational or empirical 'core'. This is, roughly speaking, the attitude that Carnap (1958) took in advocating the Ramsey sentence as expressing the 'synthetic part' of a theory, with the 'analytic part' expressed by the so-called *Carnap sentence*, $T^C = (T^R \to T)$.[6]

For non-empiricists, it is less clear what the best response is. One option is to argue that the way we have formalised the Ramsey sentence failed to capture the intuitive idea. In particular, note that our intuitive gloss above quantified over *properties*, whereas the standard semantics for second-order logic permits the second-order variables to range over *arbitrary subsets of the domain*. So one might argue that the Ramsey-sentence approach to structural content should use some other semantics for second-order logic, where the range of the second-order quantifiers is somehow restricted.

A natural way of doing this is to use so-called *Henkin semantics*. In a *Henkin model* $\mathfrak{H}$, for each $n \in \mathbb{N}$ a subset of $\mathcal{P}(|\mathfrak{H}|^n)$ is picked out as the permitted range for the

[6] For commentary and discussion, see Psillos (2000) or Andreas (2017).

second-order n-ary variables to range over.[7] This suffices to avoid the Newman objection; however, it turns out that this still captures a relatively weak notion of structural content. In particular, Dewar (2019a) shows that two theories T_1 and T_2 have logically equivalent Ramsey sentences under Henkin semantics if there exist translations from T_1 to T_2 and vice versa (without any requirement that these translations are inverse to one another); and as Example 1 showed, this is a fairly weak notion of equivalence.

In sum, then, we've seen that the notion of 'the structure of a theory' is slippier than one might expect, and admits of a variety of different formal explications. In particular, we now have a hierarchy of criteria of equivalence, in descending order of strictness:

- Logical equivalence
- Notational variance
- Intertranslatability
- Mutual translatability
- Logically equivalent Ramsey sentences (on Henkin semantics)
- Logically equivalent Ramsey sentences (on standard semantics)

We now move away from logic, and turn to theories of physics; we will bear in mind the lessons from these chapters, however, as guides to these more complex and interesting cases.

[7] Moreover, we require that these privileged subsets are, in an appropriate sense, closed under definability; see Manzano (1996) for details.

Part II.

Newtonian mechanics

4. Newtonian mechanics

In this part of the book, we consider *N-particle Newtonian mechanics*: this chapter introduces the theory and its symmetries, the next discusses the rationale for regarding symmetry-related models as physically equivalent, and the one after that considers how we might seek to revise the theory to incorporate that judgment of physical equivalence.

4.1. Introduction to N-particle Newtonian mechanics

In the first instance, we will describe this theory in terms of coordinates. We therefore expect that at least some of the structure of the theory will be unphysical, a mere 'artefact of the coordinate system'. However, we will refrain from making intuitive judgments about which structural features correspond to physical features: in due course, we will use symmetry considerations to make such judgments.

To set this up, we introduce a gadget that we will use repeatedly in what follows. Given some mathematical structure Ω, an *Ω-valued variable* is one whose intended range is Ω—thus, given a Tarski-model $\mathfrak{A}$, a first-order variable is a $|\mathfrak{A}|$-valued variable, and a second-order variable of arity n is a $\mathcal{P}(|\mathfrak{A}|^n)$-valued variable. Then, given a set $\{\xi_1, \ldots, \xi_m\}$ of Ω-valued variables, the *value space* associated with that set consists of all maps from $\{\xi_1, \ldots, \xi_m\}$ to Ω. Note that given an ordering on the variables, such a map can also be thought of as an m-tuple of elements of Ω: such an m-tuple, after all, is simply a map from $\{1, \ldots, m\}$ to Ω. Thus, the value space is isomorphic to Ω^m.[1]

Now, without worrying too much about what doing this might mean, we take as given a coordinate system with x-, y- and z-axes, which persists over time; and we take as given some clock that measures the passage of time. We introduce one $\mathbb{R}$-valued variable $\ulcorner t \urcorner$, representing time, and three $\mathbb{R}$-valued variables $\ulcorner x^1 \urcorner$, $\ulcorner x^2 \urcorner$ and $\ulcorner x^3 \urcorner$, representing (respectively) the x-, y- and z-components of position. From these,

[1] Of course, one of the lessons that should have been taken from the previous chapters is that this kind of loose use of terms like 'isomorphic' is to be deplored. Consider this evidence that one should do as I say, not as I do.

we form the value spaces T (isomorphic to $\mathbb{R}$) and X (isomorphic to $\mathbb{R}^3$); they represent time and (position) space. We then introduce X-valued variables $\ulcorner x_1 \urcorner, \ldots, \ulcorner x_N \urcorner$, with x_n representing the position of the nth particle. The value-space for these variables will be denoted Q, and represents 'configuration space'. Q therefore consists of maps from $\{x_1, \ldots, x_N\}$ to X. Each such map is equivalent (via uncurrying) to a map from $\{x_1, \ldots, x_N\} \times \{x^1, x^2, x^3\} \to \mathbb{R}$: we abbreviate the pair $\langle x_n, x^i \rangle$ as x^i_n, so that Q can also be thought of as the value-space for $3N$ real-valued variables x^i_n (with $1 \leq i \leq 3$ and $1 \leq n \leq N$).

The *dynamics* of the theory are given by the following set of differential equations (where $1 \leq i \leq 3$ and $1 \leq n \leq N$):

$$m_n \frac{d^2 x^i_n}{dt} = F^i_n \tag{4.1}$$

Each equation in this set contains two new symbols, $\ulcorner m_n \urcorner$ and $\ulcorner F^i_n \urcorner$; these are also taken to be real-valued (in the case of m_n, to be *positive* real-valued). m_n represents the mass of the nth particle, and F^i_n the ith component of the force on the nth particle.

A *kinematically possible model* for this theory will specify (constant) values for the m_n, and the value of all x^i_n and F^i_n at each $t \in T$; so the data specified consists of N real numbers, and $6N$ functions from T to $\mathbb{R}$. Such a kinematically possible model is a *dynamically possible model* if it is a *solution* of (4.1): that is, if for all $t \in T$, and all $1 \leq i \leq 3$ and $1 \leq n \leq N$,

$$m_n \frac{d^2 x^i_n}{dt^2}(t) = F^i_n(t) \tag{4.2}$$

This theory is something of a framework; we can make it more specific by adding force-laws for the forces F^i_n. For example, if we suppose that each of our N particles has an electric charge q_n, and that they are in some electrical field given by $E^i : X \to \mathbb{R}^3$, then we have

$$F^i_n = q_n E^i(x_n) \tag{4.3}$$

On the other hand, if we are instead considering a theory where the N particles are mutually interacting through gravitation, then

$$F^i_n = \sum_{p \neq n} \frac{G m_p m_n}{|x_n - x_p|^2} \frac{x^i_n - x^i_p}{|x_n - x_p|} \tag{4.4}$$

where

$$|x_n - x_p| := \sqrt{\sum_j (x^j_n - x^j_p)^2} \tag{4.5}$$

Finally, in the trivial case of free particles, the force-functions are especially simple:

$$F_n^i = 0 \tag{4.6}$$

4.2. Changes of variables

The theory above is (by design) stated for a single coordinate system. It is taken as read that the theory describes phenomena which are independent of a coordinate system. However, simply because the *phenomena* are independent of a coordinate system, it does not follow that the *theory* is: in particular, just because the theory holds in our coordinate system it does not follow that it will hold in some other coordinate system. What follows is merely that some *other* theory, systematically related to this one, will be true in that coordinate system. In other words, in order to make use of another coordinate system, we must specify how to *translate* the theory into a theory appropriate to that coordinate system.

It will be best to illustrate this by an example. Let $N = 1$, and suppress the third dimension: then, writing $x_1^1 = x$ and $x_1^2 = y$, the theory reduces to the form

$$m\ddot{x} = F^x(x, y) \tag{4.7a}$$

$$m\ddot{y} = F^y(x, y) \tag{4.7b}$$

where dots indicate differentiation with respect to t. Writing the value-space of the set $\{\ulcorner x \urcorner, \ulcorner y \urcorner\}$ as $X \times Y$ (which is isomorphic to $\mathbb{R}^2$), we may consider solutions to this theory as functions from T to $X \times Y$.

We now introduce new variables (the *polar coordinates*) $\ulcorner r \urcorner$ and $\ulcorner \theta \urcorner$, and stipulate that $\ulcorner x \urcorner$ and $\ulcorner y \urcorner$ are translated into these new variables according to

$$x \equiv r \cos\theta \tag{4.8a}$$

$$y \equiv r \sin\theta \tag{4.8b}$$

where have used the notation introduced in Chapter 2 (anticipating that this translation will be invertible).

The value-space for the set of variables $\{\ulcorner r \urcorner, \ulcorner \theta \urcorner\}$ will be denoted $R \times \Theta$. We saw in Chapter 2 that any translation induces a dual map on models; here, this corresponds to the fact that a translation from the variables $\ulcorner x \urcorner, \ulcorner y \urcorner$ to the variables $\ulcorner r \urcorner, \ulcorner \theta \urcorner$ induces a

map from $R \times \Theta$ to $X \times Y$, namely:

$$(r, \theta) \mapsto (r \cos\theta, r \sin\theta) \tag{4.9}$$

In order that this map be a bijection, we stipulate that $r \geq 0$, that $(r, \theta) = (r, \theta + 2\pi)$, and that $(0, \theta) = (0, 0)$.[2] This means we can write down expressions for the inverse translations:

$$r \equiv \sqrt{x^2 + y^2} \tag{4.10a}$$

$$\theta \equiv \tan^{-1}\left(\frac{y}{x}\right) \tag{4.10b}$$

together with the requirement that if $x = 0$, then $\theta = \pi/2$.

We therefore have an invertible pair of translations, and so we can seek to translate the theory (4.7) into the new coordinate system. First, we observe that applying the derivative operator twice to each side of the translations (4.8) yields

$$\ddot{x} \equiv (\ddot{r} - r\dot{\theta}^2)\cos\theta - (2\dot{r}\dot{\theta} + r\ddot{\theta})\sin\theta \tag{4.11}$$

$$\ddot{y} \equiv (\ddot{r} - r\dot{\theta}^2)\sin\theta + (2\dot{r}\dot{\theta} + r\ddot{\theta})\cos\theta \tag{4.12}$$

If we substitute in these expressions, then we get a theory whose differential equations are

$$m(\ddot{r} - r\dot{\theta}^2)\cos\theta - (2\dot{r}\dot{\theta} + r\ddot{\theta})\sin\theta = F^x \tag{4.13a}$$

$$m(\ddot{r} - r\dot{\theta}^2)\sin\theta + (2\dot{r}\dot{\theta} + r\ddot{\theta})\cos\theta = F^y \tag{4.13b}$$

However, this theory is—in a certain sense—only a partial translation, since we are still expressing the forces in terms of their components in the original coordinate system. Provided that is understood, of course, the theory is still appropriate as a translation, and captures the same content as the original; indeed, there might even be situations where using one coordinate system to describe positions and another to describe forces might be appropriate (although it is not so easy to think of an example).

Nevertheless, one might argue that it is unsatisfactory, given that it makes use of two coordinate systems. For this reason, the label of 'Newtonian mechanics in polar

[2] Geometrically, this means that $R \times \Theta$ has the structure of a half-cylinder that has been 'pinched off' at one end.

coordinates' is typically reserved for the theory

$$m(\ddot{r} - r\dot{\theta}^2) = F^r \tag{4.14a}$$

$$m(2\dot{r}\dot{\theta} + r\ddot{\theta}) = F^\theta \tag{4.14b}$$

which may be obtained by extending the translation τ to the force expressions, according to

$$F^x \equiv F^r \cos\theta - F^\theta \sin\theta \tag{4.15}$$

$$F^y \equiv F^r \sin\theta + F^\theta \cos\theta \tag{4.16}$$

That is, if we substitute these expressions into (4.7), then after some algebraic manipulation we obtain the translated theory (4.14). As we would expect, if we do the reverse process, then we find that the translations (4.10) will convert the theory (4.14) into the theory (4.7)—provided, that is, that we translate the force-functions according to

$$F^r \equiv \frac{x}{\sqrt{x^2+y^2}} F^x + \frac{y}{\sqrt{x^2+y^2}} F^y \tag{4.17a}$$

$$F^\theta \equiv \frac{x}{\sqrt{x^2+y^2}} F^x - \frac{y}{\sqrt{x^2+y^2}} F^y \tag{4.17b}$$

Why are these the 'correct' ways of extending the coordinate translations to translations of force-expressions? The standard answer appeals to geometric arguments, in particular the fact that force is a *vector* quantity; thus, the components of forces in a new coordinate system are determined by computing how the 'coordinate vectors'—unit vectors directed along the coordinate axes—change under the move to a new coordinate system. However, this simply pushes the argument back, to the question of how it is we determine that forces are vector quantities. In Chapter 6 we will argue that at least part of the answer to that question is based on considerations due to symmetry, and yet (as we will see in the remainder of this chapter) those symmetry transformations depend upon fixing how it is that forces transform under coordinate changes. So circularity threatens; how to escape that threat (and whether it can be escaped) is left for the reader to ponder.

One might argue that this circularity is just a predictable consequence of being so foolish as to use coordinates in foundational enquiry.[3] Rather, we should present Newton's laws in Newton's terms, according to which positions are points in a three-dimensional

[3]See (Maudlin, 2012, chap. 2) for an especially trenchant expression of the point of view I have in mind here.

Euclidean space, and forces are quantities associated with both a magnitude and a direction in that space. For calculational convenience we can then *introduce* coordinates, and use them to arithmetise positions and forces; and, indeed, we'll find that those arithmetical expressions transform between coordinate systems in the ways given above. But (on this view) there's no need for mystery-mongering about 'what makes force a vector quantity'—that's something we put in at the *start* of describing the theory, not the output of sufficient philosophical chin-stroking!

Now, there is definitely something right about this. Once we know the geometric character of the entities involved in a given theory, there are indeed very good reasons to present the theory in terms of those geometric structures (rather than in terms of coordinates). But the topic we are interested in is exactly the question of how one comes to determine which geometric structures are the right ones. For Newtonian mechanics, it's tempting to think that it's just *obvious* what geometric structures the theory involves—isn't the position of a point particle just *manifestly* described by a point in a three-dimensional Euclidean space, and the notion of a force clearly a quantity that possesses both magnitude and direction? Unfortunately, even if we grant these examples,[4] that's no guarantee that finding the right kind of geometric object will always be so easy—whether in other theories, or even (as we shall see) in Newtonian mechanics.

4.3. Symmetries

We now turn our attention to *symmetries*, which we define as translations *between a theory and itself*. I remarked in the previous section that although the (presumed) coordinate-independence of the phenomena reassures us that we should be able to find *some* theory which can model those phenomena in another coordinate system and is a translation of our original theory, there was no *a priori* reason to think that this other theory should be similar to theory we started with.[5] However, it turns out that certain special translations *do* preserve the theory, in the sense that applying the translation yields a theory of the same form as—in the terminology of Chapter 2, which is a 'notational variant of'—the one with which we began. I remark that from now on we will refer to 'transformations' of variables rather than 'translations', in order to avoid a terminological clash with 'spatial translations'.

[4] Which is already somewhat dubious, given how much work is required to make (say) the notion of a point particle intellectually respectable.

[5] Again, I stress that the criterion for 'sameness of theory' that is being used here is that of logical equivalence—that is, identity of syntactic form (modulo logical manipulation), not of propositional content.

The symmetries in question are the so-called *Galilean transformations*:[6]

Definition 22. A *Galilean transformation* of the variables t, x^i_n to the variables $\tilde{t}, \tilde{x}^i_n$ is any transformation of the form

$$x^i_n \equiv R^i{}_j \tilde{x}^j_n + u^i \tilde{t} + a^i \tag{4.18a}$$

$$t \equiv \tilde{t} + b \tag{4.18b}$$

where $b \in \mathbb{R}$, $a^i \in \mathbb{R}^3$, $u^i \in \mathbb{R}^3$, and $R^i{}_j$ is an orthogonal matrix.[7] ♠

Geometrically, we interpret the group of Galilean transformations as comprising any combination of spatial translations (the a^i term), spatial rotations and reflections (the action of $R^i{}_j$), temporal translations (the b term), and Galilean boosts (the $u^i\tilde{t}$ term).[8]

Now, merely applying these transformations will not deliver a theory of the same form as (4.1). In order for that to happen, the force term must be transformed according to

$$F^i_n \mapsto R^i{}_j \tilde{F}^j_n \tag{4.19}$$

Again, the standard justification for why the force-terms should be transformed like this is that forces are vectors; for our purposes, though, we merely note that *if* this transformation is applied (together with the transformation (4.18)), then the theory obtained is

$$R^i{}_j m_n \frac{d^2 \tilde{x}^i_n}{dt} = R^i{}_j \tilde{F}^i_n \tag{4.20}$$

which is—once we've applied the inverse matrix to $R^i{}_j$ on both sides—a notational variant of (4.1).

Note that in order for this procedure to work, we don't actually need $R^i{}_j$ to be an orthogonal matrix: it will suffice that it be invertible. In this sense, the symmetry group of Newton's Second Law (alone) is *wider* than the Galilean group.[9] However, if we regard the force-term not just as a placeholder, but as some functional expression for the forces in terms of other physical quantities, then we can ask: are the new force-components $\tilde{F}^i_n$, when expressed as functions of the new coordinates $\tilde{x}^i_n$, *of the same functional form* as the old force-components F^i_n when those were expressed as functions of the old coordinates x^i_n?[10] In other words, suppose we supplement Newton's Second

[6] Here and throughout, we use the Einstein summation convention (see Appendix A).

[7] That is, a matrix whose transpose is its inverse: see Appendix A.

[8] I exclude time-reversal, although this is also a symmetry of these equations, simply because it's a rather tricky one to deal with.

[9] See (Wheeler, 2007, Appendix 1).

[10] cf. (Brown, 2005, §3.2).

Law (4.1) by a force-law of the schematic form

$$F_n^i = \Phi_n^i(t, x_p, v_p) \tag{4.21}$$

where $\Phi_n^i(t, x_p, v_p)$ is a functional expression featuring (in general) the time, particle positions, and particle velocities—such as found in (4.3), (4.4), or (4.6). Then we can ask whether a given transformation of the coordinates and force-components is a symmetry of this force-law; that is, whether substituting the expressions in (4.18) and (4.19) will enable us to derive

$$\tilde{F}_n^i = \Phi_n^i(\tilde{t}, \tilde{x}_p, \tilde{v}_p) \tag{4.22}$$

where $\tilde{v}_p^i = v_p^i + u^i$.

And the answer to this question is that *if* the forces are independent of the time and the particle velocities, and if they depend only on inter-particle displacements, and if they only depend on those either 'linear component-wise' or via distances—*then* the symmetry group is the Galilean group.[11] More precisely, the Galilean transformations will be symmetries if we suppose that Φ_n^i takes the form

$$\Phi_n^i(t, x_p, x'_p) = M^{km}{}_n(x_k^i - x_m^i) \tag{4.23}$$

where $M^{km}{}_n$ is an array of N^3 coefficients that depend only on the inter-particle distances (i.e. on expressions of the form $|x_n - x_p|$, defined as in Equation (4.5)). Thus, the force-law (4.4) satisfies this condition, as does the (trivial) force-law (4.6); but the force-law (4.3) does not.

As we discussed in §4.2, the transformation (4.18) will induce a map $\tilde{T} \times \tilde{X}$ to $T \times X$; this will, in turn, induce a (bijective) mapping from solutions over the latter space to solutions over the former space. However, if the transformation is a symmetry, then the same differential equations will hold over both spaces, and so we can interpret the transformation *actively* rather than *passively*: that is, we can identify X with $\tilde{X}$ and T with $\tilde{T}$, and regard this map as a bijection from the space of solutions over $T \times X$ to *itself*. This mapping will relate a given solution to (4.1) to a solution which is (relative to it) translated, rotated, and boosted. The question we now consider is what the relation is between these solutions: in the next chapter, we discuss some reasons for thinking that these solutions are *physically equivalent*, and in Chapter 6 we consider how to reformulate the theory in light of such a judgment.

[11] The fact that the symmetry group depends on the nature of the force-laws—in particular, on whether they are velocity-independent or not—is discussed in (Brown, 2005, §3.2) and Barbour (1989).

5. Symmetry and equivalence

In this chapter, we consider the idea that when two models of Newtonian mechanics are related by a Galilean symmetry, they are (or should be interpreted as) physically equivalent. This idea has a long history, going back (at least) to the famous correspondence of Leibniz and Clarke:

> To say that God can cause the whole universe to move forward in a right line, or in any other line, without making otherwise any alteration in it; is another chimerical supposition. For, two states indiscernible from each other, are the same state; and consequently, 'tis a change without any change.[1]

This quotation also points towards one of the key ideas underpinning this interpretational move, at least in the current literature: the idea that Galilean symmetries hold between models that are indiscernible from one another, or (as we would say now) that they are *empirically equivalent*.[2]

In this chapter, we get a handle on why we might think that such models are empirically equivalent. The argument for this goes, roughly, as follows:

- To say that two models are empirically discernible is to say that some quantity could be measured to have different values in the two models.
- To measure a physical quantity is to set up a dynamical process which reliably and systematically correlates the value of that quantity with some independent quantity belonging to the measuring device.
- Symmetries commute with the dynamics: applying a symmetry transformation and letting a system evolve delivers the same result as letting the system evolve then applying the symmetry transformation.
- So if a measuring device ends up in a certain state when the symmetry is not applied, and if the quantities characterising that state are independent of the symmetry, then it will end up in the same state when the symmetry is applied.

[1] (Alexander, 1956, p. 38)

[2] Exactly what role empirical equivalence plays in Leibniz's own argument(s) for this conclusion is a rather more vexed question, and not one that we will go into any deeper here.

- And so, there is no way of reliably correlating a symmetry-variant quantity with the end state of the measuring device.

In the next section, we make precise the sense in which 'symmetries commute with the dynamics', and how this means that the symmetry-invariant dynamics is (as we will say) autonomous from from the symmetry-variant dynamics; after that, we discuss how to draw a conclusion about measurability from this observation about dynamics.

5.1. Autonomy

To do the analysis, which will closely follow Wallace (ndc))'s treatment, we will take advantage of the determinism of the Newtonian theory: the fact that, under fairly mild assumptions, fixing the positions and velocities of all the particles at one time yields a unique evolution of the system thereafter.[3] For these purposes, it will be helpful to introduce *state space*. First, we introduce $3N$ real-valued variables $\ulcorner v_n^i \urcorner$ to represent the particle velocities, and then define the state space $\hat{Q}$ to be the value space associated with the set $\{\ulcorner x_n^i \urcorner, \ulcorner v_n^i \urcorner\}$, for $1 \leq i \leq 3$ and $1 \leq n \leq N$; thus, $\hat{Q}$ is isomorphic to $\mathbb{R}^{6N}$.

Then, let $x : T \to Q$ be a trajectory through configuration space. The *lift* of x is the trajectory $\hat{x} : T \to \hat{Q}$ defined by

$$\hat{x}_n^i(t) = \left(x_n^i(t), \frac{dx_n^i}{dt}(t) \right) \tag{5.1}$$

This is what lets us interpret the variables $\ulcorner v_n^i \urcorner$ as labelling velocities.[4] The value of this is that—as promised—although there can be more than one physically possible trajectory that passes through a given point of Q, there can be only one physically possible trajectory that passes through a given point of $\hat{Q}$. For any given point $\hat{x} \in \hat{Q}$, and any time period Δt, let $\Delta t(\hat{x})$ be the time-evolute of $\hat{x}$: that is, if the solution x is such that $\hat{x}(0) = \hat{x}$, then $\hat{x}(t) = \Delta t(\hat{x})$.

We now turn to the autonomy argument itself, which is based on Wallace (ndc). In the interests of simplicity, we will only give the argument for the *Euclidean* group of

[3] Mild, but not inviolable. For example, we must assume that the forces are describable by Lipschitz-continuous functions: see Norton (2008) for a model with non-Lipschitz force-functions, and Malament (2008), Wilson (2009), and Fletcher (2012) for discussion. For more ways in which classical determinism can break down, see Earman (1986).

[4] Technically speaking, $\hat{Q}$ is the *tangent space* to Q. The kind of construction here can be extended to higher derivatives and partial derivatives: this yields the concept of a *jet space*, which can be used to give a nice treatment of partial differential equations (since a partial differential equation is just an algebraic equation for an appropriate jet space). See (Olver, 1986, §2.3), Belot (2013).

symmetries E; extending this to the full Galilean group poses interesting issues, that we do not have the space to treat here.[5] First, we extend the action of the Euclidean group E on Q to an action of E on $\hat{Q}$, by stipulating that the action of some element $(R^i{}_j, a^i) \in E$ on (x^i_n, v^i_n) is

$$x^i_n \mapsto R^i{}_j x^j_n + a^i \tag{5.2a}$$

$$v^i_n \mapsto R^i{}_j v^j_n \tag{5.2b}$$

Strictly, we already anticipated that this is the appropriate action (at the end of the previous chapter), but now we can see why this is so. The reason is that it respects the lifting of trajectories from Q to $\hat{Q}$: that is, for any Euclidean transformation $g \in E$ and any trajectory $x : T \to Q$,

$$\widehat{gx} = g\hat{x} \tag{5.3}$$

Since the action of E on $\hat{Q}$ is independent of time, we have the following fact (a precise statement of our earlier claim that 'symmetries commute with the dynamics'):

Proposition 8. For any point $\hat{x}_0 \in \hat{Q}$ and any Euclidean transformation $g \in E$, $\Delta t(g(\hat{x})) = g(\Delta t(\hat{x}_0))$. In other words, the diagram below commutes:

$$\begin{array}{ccc} \hat{Q} & \xrightarrow{g} & \hat{Q} \\ {\scriptstyle \Delta t}\uparrow & & \uparrow{\scriptstyle \Delta t} \\ \hat{Q} & \xrightarrow{g} & \hat{Q} \end{array}$$

Proof. Take any $\hat{x}_0 \in \hat{Q}$, and consider the solution $\hat{x} : T \to \hat{Q}$ such that $\hat{x}(0) = \hat{x}_0$, and (hence) $\hat{x}(\Delta t) = \Delta t(\hat{x}_0)$. Since symmetries map solutions to solutions, $g\hat{x}$ is also a solution (for any $g \in E$). But $(g\hat{x})(0) = g(\hat{x}(0)) = g\hat{x}_0$, and $(g\hat{x})(\Delta t) = g(\hat{x}(\Delta t)) = g(\Delta t(\hat{x}_0))$; so $\Delta t(g\hat{x}_0) = g(\Delta t(\hat{x}_0))$. □

The action (5.2) partitions $\hat{Q}$ into E-orbits;[6] this means that if we choose some reference point $\hat{x}_0$ within each orbit, then points of $\hat{Q}$ can be (redundantly) labelled by pairs of the form $([\hat{x}_0], g)$ where $[\hat{x}_0]$ is an E-orbit and $g \in E$. But since the symmetries in E commute with the dynamics, if two points $\hat{x}, \hat{y} \in \hat{Q}$ are in the same E-orbit, then their time-evolutions $\Delta t(\hat{x})$ and $\Delta t(\hat{y})$ must be in the same E-orbit as well. Thus, if we know what E-orbit the system is in at one time, we can predict what E-orbit it will be in at any

[5] For discussion, see (Wallace, ndc, §8).
[6] See Appendix B.

later time: and it is in this sense that there is a dynamics for the E-invariant data which is *autonomous* from the E-variant data. It follows that if we alter the E-variant data, by moving the state of the system around within an orbit, this will have no effect on what orbit the system occupies at a later time.

What about the other direction? Will changing the invariant data alter the evolution of the variant data? More precisely, suppose that we take a given point (O, g) in $\hat{Q}$ (where O is an E-orbit), and consider the point (O', g): if $\Delta t(O, g) = (P, h)$, is it the case that $\Delta t(O', g) = (P', h)$? In general, the answer to this question is 'no'.[7] Let us take the case $N = 2$, and restrict our attention to the translation group T: a symmetry orbit of T in $\hat{Q}$ is then specified by the displacement $x_2^i - x_1^i$, and the (absolute) particle velocities v_1^i and v_2^i. Now suppose that we take the reference point within each orbit to the be the point such that the first particle is at the origin, i.e. $x_1^i = 0$. Then moving from a point (O, g) to a point (O', g) amounts to altering the location (or velocity) of particle 2 whilst leaving the location of particle 1 the same; and in general, this will affect the future evolution of particle 1's location. Suppose, for example, that the two particles are in a stable orbit; then moving particle 2 away from particle 1 will mean that particle 1's location will evolve very differently (compared to how it would have evolved had they remained in a stable orbit).

5.2. What makes a measurement?

Thus, we have found that structural features of a Euclidean symmetry mean that if we alter the Euclidean-variant data (i.e. absolute location and orientation) but not the Euclidean-invariant data (i.e. relative positions and orientations) then the future evolution of the invariant data is unaffected; and we have seen that this does not hold if we swap the terms 'variant' and 'invariant'. As already mentioned, these results extend to the full Galilean group. Consequently, insofar as we seek a dynamics for the Galilean-invariant data, we need not consider the variant data. In particular, this demonstrates that if we wanted to use the invariant degrees of freedom as a way of measuring the variant degrees of freedom, we would not be able to do so: there is no way to get a system of particles to measure their absolute velocity, in the sense of having some setup which will reliably correlate the absolute velocity of those particles with their relative positions and velocities.

[7] Indeed, in a sense this question isn't even well-posed: what it means to change the symmetry-invariant data whilst not changing the symmetry-variant data is dependent on the choice of reference point within each orbit. This argument shows that even if we ignore this problem, we obtain a negative answer.

However, this naturally invites the question: is there something wrong with using the *variant* degrees of freedom as a means of measuring the variant degrees of freedom? For example, the facts about an object's absolute position over time certainly encode facts about its absolute velocity. So suppose that someone proposed using absolute position as a means of measuring absolute velocity. Now, it certainly seems that there's something defective about this proposal; can we say anything enlightening about what that something is?

Broadly speaking, one can discern three proposed answers to this question in the literature. One answer points to the fact that absolute position and velocity are *unobservable*, in the sense that we as humans cannot perceive them directly; the above argument is then understood as showing that they cannot be indirectly detected, either. This answer then (typically) goes on to conjecture that the definition of symmetry should include the requirement that symmetry-related models are (in some appropriate sense) observationally equivalent. For examples of this answer, see Dasgupta (2016) and Ismael and van Fraassen (2003).

A second answer points toward considerations from philosophy of language. The idea here is that even if we were able to detect symmetry-variant quantities by recording the result in other symmetry-variant quantities, the knowledge that would thereby be gained would exhibit a peculiar form of *untransmissibility* or *unencodability*. We cannot (for example) encode the 'result' of this detection by such familiar means as writing it down, or weaving it into a tapestry, or sending it via email; it would therefore violate a principle that any reliably manipulable physical process can be used as a channel for communicating knowledge. Roberts (2008) outlines a concern of this kind.

Finally, Wallace (ndc) has argued that the problem with a 'measurement' of this kind is that the quantity being measured and the quantity being used to encode the measurement result are insufficiently independent of one another. To motivate this answer, consider the proposal that we use an object's absolute velocity as a measure of its own absolute velocity: certainly, these two quantities are guaranteed to covary with one another, but it seems wrong to think of this as a *measurement*. Paraphrasing somewhat, Wallace then argues that when we take the kind of abstract dynamical perspective laid out above, we end up recognising the symmetry-variant data as constituting a single quantity. It will then follow that the kind of procedure suggested at the start of this section is defective for the same reason.

There is not the space here to defend any of these answers in detail, or to discuss further nuances.[8] However, we will take it that one of these answers can be made to

[8]One very important such nuance is that the discussion here only concerns the case where we seek to

work; or at least, that there is clearly something defective about the proposal to record measurements of symmetry-variant data in other symmetry-variant data. Helping ourselves to that assumption, we conclude (on the basis of the autonomy of the symmetry-invariant data) that no non-defective measurement of symmetry-variant quantities is possible after all. Together with plausible Occamist norms about not having undetectable quantities in one's physics, this motivates the interpretation of symmetry-related models as physically equivalent; or in other words, the interpretation of symmetry-variant quantities as 'surplus structure'. In the next chapter, we examine some consequences of adopting such an interpretation.

record the result of measuring some quantity of the system in other quantities of that same system; thus, we have neglected any discussion of how things stand when we think about relationships between *subsystems*. Yet such cases are crucial to a proper understanding of the empirical significance of symmetries: see Kosso (2000), Brading and Brown (2004), Healey (2009), Greaves and Wallace (2014), Wallace (nda), and references therein.

6. Galilean spacetime

So suppose that we are persuaded that symmetry-related states of affairs should be regarded as physically equivalent. Our next step—plausibly—should be to find some way of presenting our theory so that the symmetry-related states of affairs are more manifestly equivalent; that is, to find a version of the theory such that symmetry-related models are mathematically equivalent.[1] In this chapter, we consider such a reformulation: that which sets the theory on *Galilean spacetime*. This amounts, in effect, to looking at the structure of $T \times X$ (which is, recall, isomorphic to $\mathbb{R}^4$) that is invariant under the action of the Galilean group.

6.1. Euclidean space and time

In this section, we consider the action of the Euclidean group E on X, and ask what substructure of X is invariant under this group action. It is best to begin with the action of the (three-dimensional) orthogonal group $O(3)$, i.e. the group of all rotations and reflections. So let $R^i{}_j$ be an orthogonal matrix, and consider the mapping from X to itself given by

$$x^i \mapsto R^i{}_j x^j \tag{6.1}$$

First, we observe that the *vector-space* structure of X is preserved, since the mapping (6.1) is a linear map: that is, for any $x^i, y^i \in X$ and $a \in \mathbb{R}$,

$$R^i{}_j(x^j + y^j) = R^i{}_j x^j + R^i{}_j y^j \tag{6.2}$$

$$R^i{}_j(a x^j) = a R^i{}_j x^j \tag{6.3}$$

[1] Indeed, on some analyses of symmetry, it is an error to—as we have done here—interpret symmetries as relating physically equivalent states of affairs before having such a redundancy-eliminating alternative to hand. In the terminology of Møller-Nielsen (2017), the viewpoint taken here (where one interprets symmetries as physical equivalences before reformulating the theory) is referred to as the *interpretationalist* approach; the alternative (where a symmetry may only be interpreted as a physical equivalence once such a reformulation has been found) is referred to as the *motivationalist* approach, since symmetries merely provide motivation for seeking an appropriate reformulation, not warrant for an interpretation. For defences of the motivationalist approach, see Møller-Nielsen (2017); Read and Møller-Nielsen (2018); for the interpretationalist approach, see Saunders (2003).

Second, we observe that the *Euclidean inner product* on X is preserved, since $R^i{}_j$ is an orthogonal matrix: that is, for any $x^i, y^i \in X$,

$$\delta_{ij} R^i{}_k x^k R^j{}_l y^l = \delta_{ij} x^i y^j \tag{6.4}$$

These results should be fairly intuitive: the first means that the linear structure of X is preserved under rotations and reflections, and the second means that the distance from any point to the origin is preserved.[2] Thus, the structure of X invariant under rotations and reflections includes, at least, the structure of a Euclidean vector space. Furthermore, it turns out that the structure of a Euclidean vector space $\mathbb{X}$ *exhausts* the invariant structure of X, in the following sense: the only automorphisms of a Euclidean vector space are the rotations and reflections. Thus, replacing X with its $O(3)$-invariant substructure means replacing it by a Euclidean vector space $\mathbb{X}$.

However, we don't want the theory to be invariant only under rotations and reflections: it should also be invariant under translations. Therefore, we need to pick out the substructure of $\mathbb{X}$ which is invariant under translations.[3] As discussed in Appendix A, the translation-invariant substructure of a vector space is an *affine* space; and the translation-invariant substructure of an inner-product space is a *metric* affine space. So the E-invariant substructure of X is a three-dimensional Euclidean affine space, $\mathcal{X}$.

Finally, consider the temporal translations. As with the spatial translations, these mean that we should replace T with a (one-dimensional) affine space $\mathcal{T}$. However, since the vector space associated to this affine space is T (which is isomorphic to $\mathbb{R}$), the affine space has not only a metric but also an *orientation*: in other words, a distinction between the past and future directions. This is a consequence of our cowardly decision to exclude time-reversal from the Galilean group by mere fiat; had we not done so, then the temporal vector space would be a non-oriented one-dimensional inner-product space, and the temporal affine space would be a non-oriented one-dimensional Euclidean affine space. We'll see in a moment how this makes our lives easier.

Thus, the spatial and temporal symmetries of our theory motivate us to move from $T \times X$ to $\mathcal{T} \times \mathcal{X}$. This latter structure is a product affine space, whose associated vector space is $T \oplus \mathbb{X}$.[4] It is (a very anachronistic rendering of) the kind of structure that Newton hypothesised for space and time; for this reason, the structure $\mathcal{T} \times \mathcal{X}$ is known

[2] This gloss—that preservation of the inner product is equivalent to preservation of the norm—exploits the fact that Euclidean inner products and Euclidean norms are interdefinable.

[3] The vector space $\mathbb{X}$, if regarded as a candidate for physical space, is referred to in the literature as *Aristotelian space*.

[4] See Appendix A.

in the literature on spacetime theories as *Newtonian spacetime*.[5]

This lets us state our theory in 'Euclidean-invariant' terms. Rather than having our particles' locations take values in $X \cong \mathbb{R}^3$, let us instead have them take values in our Euclidean space $\mathcal{X}$; as before, we will use the (non-indexed) variable x_n for the location of the nth particle. Furthermore, in light of the fact that forces transform like locations under rotations and reflections, but remain invariant under translations, let us have the forces take values in $\mathbb{X}$: this is the sense in which the symmetry structure of the theory indicates to us that force is a vectorial quantity. We replace the triple of variables F_n^i with the $\mathbb{X}$-valued variable $\vec{F}_n$.

We can now write down the following new version of (4.1):

$$m_n \frac{d^2 x_n}{dt} = \vec{F}_n \tag{6.5}$$

This equation is well-formed: given an $\mathcal{X}$-valued curve parameterised by $\mathcal{T}$, its future-directed derivative is an $\mathbb{X}$-valued curve parameterised by $\mathcal{T}$; and the future-directed derivative of *that* curve is another $\mathbb{X}$-valued curve parameterised by $\mathcal{T}$. Note that if $\mathcal{T}$ was not an *oriented* space, then we could not so straightforwardly encode derivatives with respect to $\mathcal{T}$ as vectors.[6] Thus, it makes sense to demand that (6.5) holds at all times in $\mathcal{T}$.

6.2. Galilean spacetime

However, Newtonian spacetime is not invariant under Galilean boosts; so our work is not yet done. First, though, we need to specify how it is that a Galilean boost acts on Newtonian spacetime (compared to coordinate space and time). First, its action on Newtonian *vector* space is as follows: given any $(t \oplus \vec{x}) \in T \oplus \mathbb{X}$, a boost along $\vec{u} \in \mathbb{X}$ acts according to

$$t \mapsto t \tag{6.6}$$

$$\vec{x} \mapsto \vec{x} + \vec{u}t \tag{6.7}$$

[5] Stein (1967)

[6] See Malament (2004) for discussion of what we could do instead.

To specify a boost's action on $\mathcal{T} \times \mathcal{X}$, we must choose some specific time $\tau_0 \in \mathcal{T}$; relative to this choice, a boost along $\vec{u}$ acts on $(\tau, x) \in \mathcal{T} \times \mathcal{X}$ as

$$\tau \mapsto \tau \tag{6.8}$$

$$x \mapsto x + \vec{u}(\tau - \tau_0) \tag{6.9}$$

We start by considering what structure of $T \oplus \mathbb{X}$ is invariant under boosts.[7] Let us refer to an element of $T \oplus \mathbb{X}$ as a *four-vector*. Both T and $\mathbb{X}$ are subspaces of $T \oplus \mathbb{X}$: the former corresponds to all four-vectors of the form $(t, 0)$, and the latter to all four-vectors of the form $(0, \vec{x})$. A boost preserves the latter subspace, but not the former, since a boost acts on elements of these subspaces as follows:

$$(t \oplus 0) = (t \oplus \vec{u}t) \tag{6.10}$$

$$(0 \oplus \vec{x}) = (0 \oplus \vec{x}) \tag{6.11}$$

However, although it does not preserve the subspace T, it does preserve the *quotient* space ${}^{T \oplus \mathbb{X}}\!/_{\mathbb{X}}$; and this quotient space is isomorphic (as a vector space) to T, via the isomorphism $t \mapsto \{(t, \vec{x}) : \vec{x} \in \mathbb{X}\}$. Thus, ${}^{T \oplus \mathbb{X}}\!/_{\mathbb{X}}$ is a one-dimensional oriented Euclidean space.

This motivates the conjecture that the substructure of $\mathbb{T} \oplus \mathbb{X}$ invariant under boosts is the structure of (what we shall call) a *Galilean vector space*:

Definition 23. A *Galilean vector space* is a four-dimensional vector space $\mathbb{G}$ with a privileged three-dimensional subspace $\mathbb{X}$, equipped with a Euclidean inner product on $\mathbb{X}$, and both a Euclidean inner product and an orientation on ${}^{\mathbb{G}}\!/_{\mathbb{X}}$. The quotient space ${}^{\mathbb{G}}\!/_{\mathbb{X}}$ may therefore be identified with the space T. ♠

We will say that a Galilean four-vector $\vec{\xi} \in \mathbb{G}$ is *purely spatial* if $\vec{\xi} \in \mathbb{X}$; note well that there is no analogous notion of a Galilean vector's being 'purely temporal'. The structure of $\mathbb{G}$ is that of a four-dimensional vector space that has been foliated into families related by purely spatial vectors; these families are the elements of the quotient space T. Unlike Newtonian four-vectors, Galilean four-vectors cannot be (uniquely) decomposed into spatial and temporal components. However, given a Galilean four-vector $\vec{\xi}$, we can take its *temporal projection* $\vec{\xi}_T$, by applying quotient map from $\mathbb{G}$ to T; that is, $\vec{\xi}_T$ is simply the family to which $\vec{\xi}_T$ belongs, and may be identified with a real number (in light of the inner product and orientation on the quotient space).

With this, we can define *Galilean spacetime* as follows:

[7] The following is based on Saunders (2013).

Definition 24. A *Galilean spacetime* is an affine space $\mathcal{G}$ whose associated vector space $\mathbb{G}$ is a Galilean vector space. ♠

And, indeed, we find that the automorphisms of Galilean spacetime—so defined—are precisely the Galilean transformations. The structure of Galilean spacetime, as with affine spaces generally, is that of Galilean vector space 'without the origin'. More specifically, it is a four-dimensional affine space with a foliation into three-dimensional subspaces, where each subspace is isomorphic to $\mathcal{X}$ (equivalently, where the elements of each three-dimensional subspace are related by purely spatial vectors); and where the quotient space $\mathcal{G}/\mathbb{X}$ is isomorphic to $\mathcal{T}$. Given a point $\xi \in \mathcal{G}$, we take its *temporal projection* to be the point $\xi_{\mathcal{T}} \in \mathcal{T}$ to which it is taken by the quotient map (in other words, to be the subspace to which it belongs).

The key difference from Newtonian spacetime, then, is that there is no 'persistence of space over time': since there is no notion of a vector being 'purely temporal', we cannot say of two points in $\mathcal{G}$ that they differ by a purely temporal vector, and hence correspond to the same point of space at two different times. (By contrast, since we do have a notion of purely spatial vectors, we can say of two points of $\mathcal{G}$ that they differ by such a vector and hence correspond to two different points of space at the same time; this is precisely the relation that foliates $\mathcal{G}$.)

Finally, then, we wish to state our theory of Newtonian mechanics in terms of these structures. To do so, we will take a kinematically possible model to consist of N smooth curves $\gamma_n : \mathcal{T} \to \mathcal{G}$, all of which are such that for any $t \in \mathcal{T}$, $(\gamma_n(t))_{\mathcal{T}} = t$. It follows that the derivative $d\gamma_n/dt$ is (at any time) a Galilean four-vector whose temporal projection is 1; and hence, that the second derivative $d^2\gamma_n/dt^2$ is (again, at any time) a purely spatial Galilean four-vector. So, letting forces take values (as before) in $\mathbb{X}$, our final formulation of Newtonian mechanics consists of the equation

$$m_n \frac{d^2\gamma_n}{dt^2} = \vec{F}_n \tag{6.12}$$

The upshot of all this is that if two solutions γ and γ' are related by a Galilean transformation, then they are *isomorphic* to one another. More precisely, there are isomorphisms $g : \mathcal{G} \to \mathcal{G}$ and $h : \mathcal{T} \to \mathcal{T}$ such that for all $t \in \mathcal{T}$, $g(\gamma(t)) = \gamma'(h(t))$; that is, such that the following diagram commutes:

$$\begin{array}{ccc} \mathcal{T} & \xrightarrow{\gamma} & \mathcal{G} \\ \downarrow{\scriptstyle h} & & \downarrow{\scriptstyle g} \\ \mathcal{T} & \xrightarrow{\gamma'} & \mathcal{G} \end{array}$$

As a result, γ and γ' are structurally indistinguishable. So, as desired, we have reformulated our theory in such a way that physical equivalence corresponds to structural equivalence.

There is a great deal of debate in the literature about whether this kind of structural equivalence is enough, or whether something stricter is required—such as a formulation of the theory in which symmetry-related models are *identical* (not just isomorphic). For the most part, this debate takes it as read that structurally equivalent models agree on the distribution of qualitative properties, and therefore concerns the acceptability of arguing on purely philosophical grounds (i.e. without further mathematical work) that there are no non-qualitative differences between possible worlds. However, since this debate (over so-called 'sophisticated substantivalism') mostly concerns metaphysical issues that are orthogonal to our main interests in this work, we leave it aside.[8]

[8]For discussion of this issue, see Pooley (2006) and references therein.

Part III.

Electromagnetism

7. Electromagnetism

Thus, we have seen that the symmetries of Newtonian mechanics provide a guide to which of its models should be regarded as physically equivalent—and hence, to what physical structure we should take the theory to be positing. In this part, we again consider how the symmetries of a physical theory provide a guide to its structure: this time, with the case study of *classical electromagnetism*. In this chapter, we introduce the theory and discuss its spacetime symmetries; this will bring out points of analogy and disanalogy with the spacetime symmetries of Newtonian mechanics. The following two chapters discuss the interpretation of a new kind of symmetry that we find in electromagnetism: its so-called *gauge* symmetry.

7.1. Electromagnetism on Newtonian spacetime

In discussing electromagnetism, we could have begun with the theory presented in terms of coordinates, analysed its spatial and temporal symmetries, and then passed to a formulation that takes those symmetries into account. However, doing so would be mostly duplicative of the analysis given in the previous Part. So instead, we start by just asserting to the reader that this coordinate-based formulation admits (at least) the following symmetries:

- Time translation
- Spatial translation
- Spatial rotation

Note that this time not only time-reversals, but also spatial reflections, are missing. Again, these are symmetries of the theory (at least on some analyses), but treating them raises some subtleties that we would rather not have to deal with.[1]

Skipping the details, taking account of these symmetries motivates employing a three-dimensional *oriented* Euclidean space $\mathcal{X}$ (with an oriented Euclidean vector space $\mathbb{X}$)

[1] What to say about time reversal in electromagnetism is especially controversial: for discussion, see Albert (2000), Malament (2004), Leeds (2006) and Arntzenius and Greaves (2009).

and a one-dimensional oriented Euclidean time $\mathcal{T}$ (with $T \cong \mathbb{R}$ as its associated vector space). We introduce two $\mathbb{X}$-valued variables $\ulcorner \vec{E} \urcorner$ and $\ulcorner \vec{B} \urcorner$, to represent (respectively) the *electric field* and the *magnetic field*; we also use an $\mathbb{X}$-valued variable $\ulcorner \vec{j} \urcorner$ to represent the current density, and a real-valued variable $\ulcorner \rho \urcorner$ to represent the current density.

A kinematically possible model of this theory will consist of time-dependent vector fields $\vec{E}(t, x) : \mathcal{T} \times \mathcal{X} \to \mathbb{X}$, $\vec{B}(t, x) : \mathcal{T} \times \mathcal{X} \to \mathbb{X}$, and $\vec{j} : \mathcal{T} \times \mathcal{X} \to \mathbb{X}$, and a time-dependent scalar field $\rho : \mathcal{T} \times \mathcal{X} \to \mathbb{R}$. A dynamically possible model consists of such a set of fields that satisfy *Maxwell's equations*:[2]

$$\mathrm{div}(\vec{E}) = \rho \tag{7.1a}$$

$$\mathrm{div}(\vec{B}) = 0 \tag{7.1b}$$

$$\mathrm{curl}(\vec{E}) + \frac{\partial \vec{B}}{\partial t} = 0 \tag{7.1c}$$

$$\mathrm{curl}(\vec{B}) - \frac{\partial \vec{E}}{\partial t} = \vec{j} \tag{7.1d}$$

Note that if we were not using an oriented Euclidean space, the curl operator would not be well-defined.[3]

Unfortunately, this is not the place for a full discussion of the physical significance of these equations, but *very* briefly: equation (7.1a) expresses the fact that electrical charge (represented by ρ) is the source for the electric field $\vec{E}$, in the sense that the total flux of $\vec{E}$ through a closed surface is proportional to the charge enclosed therein; equation (7.1b) expresses the fact that the magnetic field $\vec{B}$ does not have sources, in the sense that the total flux of $\vec{B}$ through a closed surface is always zero; equation (7.1c) expresses the fact that an electrical field may be 'induced' by a time-varying magnetic field; and equation (7.1d) expresses the fact that a magnetic field may be induced by a time-varying electric field or by an electrical current (represented by $\vec{j}$).[4]

However, this is not the only 'coordinate-free' presentation of electromagnetism on Newtonian spacetime—nor, arguably, the most perspicuous. An alternative presentation makes use of the language of *differential forms*.[5] In this presentation, we employ an $\mathbb{X}^*$-valued variable $\ulcorner \mathbf{E} \urcorner$ to represent the electric field, a $\Lambda^2(\mathbb{X}^*)$-valued variable $\ulcorner \mathbf{B} \urcorner$ to represent the magnetic field, an $\mathbb{X}^*$-valued variable $\ulcorner \mathbf{j} \urcorner$ to represent the current density and (again) a real-valued variable $\ulcorner \rho \urcorner$ to represent the charge density. A kine-

[2] Here and throughout, we use units in which $\mu_0 = \varepsilon_0 = c = 1$.

[3] See Appendix A.

[4] Any good textbook on electromagnetism will further discuss these laws; a classic treatment is (Feynman et al., 2011, Volume II).

[5] See Appendix C; my discussion follows Baez and Muniain (1994).

matically possible model consists of time-dependent 1-forms $\mathbf{E} : \mathcal{T} \times \mathcal{X} \to \mathbb{X}^*$ and $\mathbf{j} : \mathcal{T} \times \mathcal{X} \to \mathbb{X}^*$, a time-dependent 2-form $\mathbf{B} : \mathcal{T} \times \mathcal{X} \to \Lambda^2(\mathbb{X}^*)$, and a time-dependent scalar field $\rho : \mathcal{T} \times \mathcal{X} \to \mathbb{R}$, that obey the (restated) Maxwell equations

$$\star d \star \mathbf{E} = \rho \tag{7.2a}$$

$$d\mathbf{B} = 0 \tag{7.2b}$$

$$d\mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{7.2c}$$

$$\star d \star B - \frac{\partial \mathbf{E}}{\partial t} = \mathbf{j} \tag{7.2d}$$

where $\star$ is the Hodge star on (oriented) Euclidean space.

Thus, we have two theories of electromagnetism on (oriented) Newtonian spacetime, neither of which employ coordinates. The two are related to one another by the isomorphisms discussed in Appendix C, i.e. the musical isomorphism and Hodge duality: specifically, they are related by

$$\mathbf{E} \equiv \vec{E}^\flat \tag{7.3a}$$

$$\mathbf{B} \equiv \star(\vec{B}^\flat) \tag{7.3b}$$

$$\mathbf{j} \equiv \vec{j}^\flat \tag{7.3c}$$

7.2. Lorentz boosts

As with Newtonian mechanics, electromagnetism exhibits a boost symmetry (in addition to the symmetries of space and time separately). However, rather than Galilean boosts, the symmetry in question concerns *Lorentz* boosts. To describe such a boost, it is easiest to start with the Newtonian vector space $T \oplus \mathbb{X}$. Let $\vec{v} \in \mathbb{X}$, and let $(t \oplus \vec{x}) \in T \oplus \mathbb{X}$. We begin by decomposing $\vec{x}$ into components parallel and perpendicular to $\vec{v}$:

$$\vec{x}_\parallel := \frac{\vec{x} \cdot \vec{v}}{|\vec{v}|^2} \vec{v} \tag{7.4a}$$

$$\vec{x}_\perp := \vec{x} - \vec{x}_\parallel \tag{7.4b}$$

where $\vec{x} \cdot \vec{v}$ is the Euclidean inner product.

Then a Lorentz boost along $\vec{v}$ means that we transform $T \oplus \mathbb{X}$ according to

$$t' \equiv \gamma(t - \vec{v} \cdot \vec{x}_{\parallel}) \tag{7.5a}$$
$$\vec{x}'_{\parallel} \equiv \gamma(\vec{x}_{\parallel} - \vec{v}t) \tag{7.5b}$$
$$\vec{x}'_{\perp} \equiv \vec{x}_{\perp} \tag{7.5c}$$

with $\vec{x}'_{\parallel}$ and $\vec{x}'_{\perp}$ being the parallel and perpendicular components of $\vec{x}'$, and where the so-called *Lorentz factor* γ is given by

$$\gamma := \frac{1}{\sqrt{1 - |\vec{v}|^2}} \tag{7.6}$$

Now let $(\tau_0, x_0) \in \mathcal{T} \times \mathcal{X}$. A Lorentz boost along $\vec{v}$ centred on t_0 and x_0 acts on an arbitrary point $(\tau, x) \in \mathcal{T} \times \mathcal{X}$ as follows: we let $\vec{x} := x - x_0$ and $t = \tau - \tau_0$, and set

$$\tau' \equiv \tau_0 + t' \tag{7.7}$$
$$x' \equiv x_0 + \vec{x}' \tag{7.8}$$

with t' and $\vec{x}'$ defined as in (7.5).

A Lorentz boost is a symmetry of electromagnetism—provided, that is, that $\vec{E}$, $\vec{B}$, ρ and $\vec{j}$ are also transformed in a certain way. Given $\vec{v} \in \mathbb{X}$, decompose $\vec{E}$, $\vec{B}$ and $\vec{j}$ into parallel and perpendicular components (as in (7.4)); the transformations are then

$$\vec{E}'_{\parallel} \equiv \vec{E}_{\parallel} \tag{7.9a}$$
$$\vec{E}'_{\perp} \equiv \gamma(\vec{E}_{\perp} + \vec{v} \times \vec{B}) \tag{7.9b}$$
$$\vec{B}'_{\parallel} \equiv \vec{B}_{\parallel} \tag{7.9c}$$
$$\vec{B}'_{\perp} \equiv \gamma(\vec{B}_{\perp} - \vec{v} \times \vec{E}) \tag{7.9d}$$
$$\rho' \equiv \gamma(\rho - \vec{v} \cdot \vec{j}_{\parallel}) \tag{7.9e}$$
$$\vec{j}'_{\parallel} \equiv \gamma(\vec{j}_{\parallel} - \rho\vec{v}) \tag{7.9f}$$
$$\vec{j}'_{\perp} \equiv \vec{j}_{\perp} \tag{7.9g}$$

Note that the transformations for ρ and $\vec{j}$ resemble those for t and $\vec{x}$; this suggests that we might consider $(\rho \oplus \vec{j})$ as a $(T \oplus \mathbb{X})$-valued vector. However, $\vec{E}$ and $\vec{B}$ transform quite differently, suggesting that they cannot be so straightforwardly given a four-dimensional interpretation; we will return to this point below.

A composition of boosts with rotations is referred to as a *Lorentz transformation*; a

composition of a Lorentz transformation with a translation (of space and/or time) is known as a *Poincaré transformation*.[6] Thus, the symmetry group of electromagnetism includes the Poincaré group. Given this, we can apply much the same arguments we did in Chapter 5 to argue that the Poincaré-variant data will be autonomous from the Poincaré-invariant data, and hence that there is no empirical procedure for measuring such data that can be modelled within electromagnetism.[7] This, in turn, gives some reason for thinking that setting electromagnetism on Newtonian spacetime is inappropriate, since in this setting there are well-defined quantities that are not invariant under Lorentz transformations (such as absolute velocity).

We do need to be careful here, however. The Poincaré-invariance of electromagnetism means that we cannot expect such quantities to be detectable via any *purely electromagnetic* procedure; however, it does not mean that such quantities will not be detectable at all. In particular, suppose that we were to take our Newtonian theory from Part II and couple it to our electromagnetic theory, by introducing the Lorentz force law:

$$\vec{F}_n = e_n(\vec{E} + \vec{v} \times \vec{B}) \tag{7.10}$$

where e_n is the electrical charge on the nth particle. If we analyse the symmetries of this *combined* theory, we find that it is limited to the Euclidean symmetries of space and time.[8] So neither Lorentz boosts nor Galilean boosts are symmetries of this theory; as a result, electromechanical experiments representable by this theory are capable of detecting Newtonian quantities such as absolute rest.

Indeed, this is (more or less) the situation that physics took itself to be in toward the end of the nineteenth century: given a Galilei-invariant mechanical theory coupled to a Poincaré-invariant electromagnetic theory, it appeared that it should be possible to design experiments that would be capable of detecting different states of absolute motion (or rather, as it was interpreted, motion relative to the luminiferous aether). Of course, such experiments (culminating in the Michelson-Morley interferometer) only delivered null results, thereby suggesting a problem with this joint theory. This was resolved following Einstein's postulation of relativistic mechanics, i.e., of a mechanical theory which had the Poincaré group as its symmetry group: this theory could then be combined with classical electromagnetism into a Poincaré-invariant theory, hence explaining the failure of electromechanical experiments to detect Poincaré-variant quan-

[6] Strictly, these are *proper orthochronous* Lorentz and Poincaré transformations, since time-reversals and parity inversions are excluded.

[7] That said, how exactly to make those arguments relevant to a field theory (such as electromagnetism) rather than a particle theory isn't wholly straightfoward: see Wallace (ndb).

[8] Note, in particular, that the Lorentz force is velocity-dependent (cf. footnote 11).

tities such as the motion relative to the aether.[9] Under the further hypothesis that any future theories we might develop will also exhibit Poincaré symmetry, we have a motivation to suppose Lorentz boosts to relate physically equivalent models.

7.3. Minkowski spacetime

We therefore seek to determine what substructure of Newtonian spacetime is invariant under Lorentz boosts. As with Galilean boosts, Lorentz boosts do not preserve the decomposition of $\mathbb{T} \oplus \mathbb{X}$ into $\mathbb{T}$ and $\mathbb{X}$; unlike Galilean boosts, they also fail to preserve either of these subspaces individually. However, they are linear, and so they do preserve its structure as a vector space. Moreover, they preserve the *Minkowski inner product*, which we define from the temporal and spatial inner products as follows: given four-vectors $\xi = (t \oplus \vec{x})$ and $\xi' = (t' \oplus \vec{x}')$ in $\mathbb{T} \oplus \mathbb{X}$, then

$$\eta(\vec{\xi}, \vec{\xi}') := tt' - \vec{x} \cdot \vec{x}' \tag{7.11}$$

Moreover, the Minkowski inner product is exhaustive of the invariant structure, in the following sense. First, define a *Minkowski vector space* as a four-dimensional vector space equipped with a Minkowski inner product:

Definition 25. *Minkowski vector space* is a four-dimensional vector space $\mathbb{M}$, equipped with an inner product of signature $(1, 3)$.[10] ♠

Then the automorphisms of an (oriented) Minkowski vector space are exactly the (proper, orthochronous) Lorentz transformations.[11] Since we want our spacetime to also be invariant under translations (i.e. to be Poincaré-invariant, not just Lorentz-invariant), we set it not on a Minkowski vector space but on the associated affine space:

Definition 26. *Minkowski spacetime* is an affine space $\mathcal{M}$ whose associated vector space $\mathbb{M}$ is Minkowski vector space. ♠

It remains only to restate our theory as a theory set on (oriented) Minkowski spacetime; to do so, we employ the language of differential forms.[12] A kinematically possible

[9] In particular, we modify Newton's Second Law to replace the mass term m_n with (velocity-dependent) relativistic mass; this means that the theory can exhibit boosts as symmetries despite having velocity-dependent forces.

[10] See Appendix A.

[11] See footnote 6.

[12] There is also a Poincaré-invariant formulation of the theory in terms of tensor algebra, which has not been developed here: see e.g. (Malament, 2012, §2.6) for a presentation.

model of this theory consists of a 2-form $\mathbf{F}$, representing the *electromagnetic field*, and a 1-form $\mathbf{J}$ representing the *covariant current density*. A dynamically possible model is one which satisfies the following equations:

$$d\mathbf{F} = 0 \tag{7.12a}$$

$$\star d \star \mathbf{F} = \mathbf{J} \tag{7.12b}$$

where $\star$ is the Hodge star operator on oriented Minkowski spacetime.

This is related to the formulation on Newtonian spacetime as follows. First, we define $\mathbf{t}$ as the unique covector in $(T \oplus \mathbb{X})^*$ such that for any $(t \oplus \vec{x}) \in \mathbb{T} \oplus \mathbb{X}$,

$$\mathbf{t}(t \oplus \vec{x}) = t \tag{7.13}$$

Equivalently, $\mathbf{t}$ is the covector in $\mathbb{T}^*$ that is dual to the (positive) unit vector in $\mathbb{T}$, regarded as a covector in $(T \oplus \mathbb{X})^* = T^* \oplus \mathbb{X}^*$. We also use $\mathbf{t}$ to denote the 1-form on $\mathcal{T} \times \mathcal{X}$ that always takes the value $\mathbf{t}$.

Next, we think of $\mathbf{E}$, $\mathbf{B}$ and $\mathbf{j}$ as differential forms on $\mathcal{T} \times \mathcal{X}$, rather than as time-dependent forms on $\mathcal{X}$ (by thinking of them as taking values in $(T \oplus \mathbb{X})^*$ rather than $\mathbb{X}^*$). We then define a 2-form $\mathbf{F}$ and a 1-form $\mathbf{J}$ on $\mathcal{T} \times \mathcal{X}$ by

$$\mathbf{F} \equiv \mathbf{B} + \mathbf{E} \wedge \mathbf{t} \tag{7.14}$$

$$\mathbf{J} \equiv \rho \mathbf{t} - \mathbf{j} \tag{7.15}$$

It can then be shown that substituting these expressions into (7.12) yields the equations (7.2).[13] Moreover, under a Lorentz boost,

$$\mathbf{F}'(\tau', x') = \mathbf{F}(\tau, x) \tag{7.16}$$

$$\mathbf{J}'(\tau', x') = \mathbf{J}(\tau, x) \tag{7.17}$$

where $\mathbf{F}'$ is defined by putting together equations (7.3), (7.9), (7.14), and (7.13), and similarly for $\mathbf{J}'$.

Thus, in sum: we began this chapter with a theory whose models were set on Newtonian spacetime $\mathcal{T} \times \mathcal{X}$, and the data for which consisted of background fields ρ and $\vec{j}$, and dynamical fields $\vec{E}$ and $\vec{B}$ (equivalently, background fields ρ and $\mathbf{j}$, and dynamical fields $\mathbf{E}$ and $\mathbf{B}$). Spatial rotations and spatiotemporal translations related isomorphic models; however, a Lorentz boost produced a non-isomorphic model. However,

[13] See (Baez and Muniain, 1994, Chap. I.5).

we have now demonstrated how to define the structure of Minkowski spacetime from that of Newtonian spacetime, and how to define a 1-form **J** and a 2-form **F**, in such a way that for any model of Newtonian spacetime, the original and transformed currents and fields are isomorphic to one another. This concludes our study of the spacetime symmetries of electromagnetism; in the next chapter, we turn to discussing its gauge symmetry.

8. Gauge transformations of electromagnetism

In this chapter, we introduce the *gauge symmetry* of electromagnetism: this differs from the symmetries considered so far in being a non-spatiotemporal symmetry, and a local symmetry. First, we present an alternative formulation of electromagnetism in terms of the *electromagnetic potential*; we then discuss its gauge symmetry, and consider how it relates to the formulation in terms of the electromagnetic field.

8.1. The electromagnetic potential

Consider again the homogeneous Maxwell equation:

$$d\mathbf{F} = 0 \tag{8.1}$$

There is a reasonably basic fact about exterior derivatives, namely that the exterior derivative of an exterior derivative vanishes: for any differential form $\mathbf{K}$,

$$dd\mathbf{K} = 0 \tag{8.2}$$

This basic fact has a less basic converse: that for any differential form with vanishing exterior derivative on a *contractible* space, there is some differential form of which it is the exterior derivative. Intuitively, a contractible space is one which can be 'continuously deformed' into a point.[1] A differential form with vanishing exterior derivative is said to be *closed*, and a differential form which is the exterior derivative of another is said to be *exact*; so we can state the basic fact by saying that every exact form is closed, and can state the less basic converse by saying that every closed form on a contractible space is exact.

Hence, given a solution to (8.1) on a contractible space, there must exist a 1-form $\mathbf{A}$

[1] More precisely, a space is contractible if the identity map on that space is homotopic to a constant map.

such that

$$\mathbf{F} \equiv d\mathbf{A} \tag{8.3}$$

This 1-form is typically referred to as the *electromagnetic potential*. We can use this to translate the Lorentz-invariant Maxwell theory into the language of A: the homogeneous equation (8.1) becomes the triviality $dd\mathbf{A} = 0$, while the inhomogeneous equation (7.12b) becomes

$$\star\, d \star d\mathbf{A} = \mathbf{J} \tag{8.4}$$

Note that we do not describe this a *re-expression* of Maxwell's equations in terms of potentials; whether this equation captures exactly the same content as (7.12b) will be our question in much of this chapter.

If we are working with Newtonian spacetime, then just as the electromagnetic field can be 'decomposed' into electric and magnetic fields, so the electromagnetic potential can be decomposed into an electric potential ϕ (a scalar field) and a magnetic potential $\vec{A}$ (a vector field), according to

$$\mathbf{A} = \phi\mathbf{t} - \vec{A}^\flat \tag{8.5}$$

These are related to the electric and magnetic fields via

$$\vec{B} \equiv \mathrm{curl}(\mathbf{A}) \tag{8.6a}$$

$$\vec{E} \equiv -\,\mathrm{grad}(\phi) - \frac{\partial \vec{A}}{\partial t} \tag{8.6b}$$

In effect, these are the Newtonian expression of equation (8.3).[2]

8.2. Gauge symmetry

We now turn to the fact that this theory admits an *internal gauge symmetry*: for any smooth function $\lambda : \mathcal{M} \to \mathbb{R}$, the transformation

$$\mathbf{A}' \equiv \mathbf{A} + d\lambda \tag{8.7}$$

is a symmetry of (8.4); this becomes clear when we recognise that the electromagnetic field $\mathbf{F}$ is invariant under this transformation.

This symmetry is importantly different to the symmetries we have discussed so far. For one thing, it is our first example of a non-spatiotemporal symmetry. For another,

[2] This comes about as a result of the relationship between the exterior derivative and the vector-calculus operators: see Appendix A.

it is a *local* symmetry. Whereas a given Euclidean, Galilean, Lorentz or Poincaré transformation is given by specifying the values of some finite number of parameters (e.g. by specifying a given rotation matrix and the velocity vector $\vec{v}$ of a Galilean boost), an electromagnetic gauge transformation is given by specifying a certain *scalar field*.

Mathematically, this means that the space of gauge transformations is far larger and richer than the space of Euclidean (etc.) transformations. Physically, it has the immediate consequence that, in a certain sense, the above theories of electromagnetism are *indeterministic*. For, since we can choose *any* smooth function λ, we can choose λ to be zero (or constant) at all times before some specified time t_0, but non-constant for at least some period of time thereafter. So the gauge transformation induced by this λ will be trivial before t_0, and non-trivial on at least some part of spacetime later than t_0. Now take any solution $\mathbf{A}$ for some fixed $\mathbf{J}$. By the definition of symmetry, $\mathbf{A} + d\lambda$ is also a solution, for that same $\mathbf{J}$; but by construction, both solutions agree at all times before t_0. So, a solution is not uniquely determined by its data prior to some given time, and hence the theory is indeterministic. Indeed, extending this argument shows that the theory is indeterministic in an even stronger sense than this: we could specify $\mathbf{A}$ at all times other than a brief window Δt, and at all places other than a small spatial region R, and we would still not be able to uniquely determine the value of $\mathbf{A}$ within R during the time period Δt.[3]

This indeterminism means that the argument for the empirical redundancy of symmetry-variant data cannot be carried out quite as straightforwardly as in Chapter 5, since states of this theory no longer have unique time-evolutes. However, since the indeterminism in question arises from the symmetry itself, if we have two Δt-evolutes S_1 and S_2 of a given state S_0, then we can guarantee that S_1 and S_2 are themselves related by a symmetry transformation. It follows that if S_0' is the result of applying a gauge transformation to S_0, then any Δt-evolute of S_0' is gauge-equivalent to any Δt-evolute of S_0. This is all we need for the argument of Chapter 5 to go through, and so we can conclude that anything which varies under a gauge transformation—such as the value of the gauge potentials—is empirically otiose.[4]

[3] This argument is modelled on Earman and Norton (1987)'s presentation of the 'Hole Argument': the argument that the diffeomorphism symmetry of General Relativity (another species of local symmetry) leads to, as they put it, 'radical local indeterminism'. Indeed, radical local indeterminism of this kind is a generic feature of theories with local symmetries; see Wallace (2003) for further discussion.

[4] Indeed, in a certain sense the gauge potentials are even less empirically accessible than the other symmetry-variant quantities we have considered, since a gauge transformation on a subregion of spacetime (that vanishes at the boundary) is also a symmetry of any larger region of spacetime, which precludes measuring gauge potentials by experiments conducted *outside* the system. See the references in footnote 8, as well as Teh (2016), Gomes and Butterfield (nd), and Wallace (ndb).

8.3. Fields and potentials

So, we have introduced the electromagnetic potential; demonstrated that it exhibits gauge symmetry; and concluded that we would like a formulation of the theory which is appropriately gauge-invariant. At this point, one might make the observation that the electromagnetic field is a gauge-invariant object—indeed, this was the observation that led us to recognise the gauge-symmetry of the potential-theory in the first place. So have we not come full circle: introducing the potential-theory only to conclude that we were better off with the field-theory?

One response to this is to note the various ways in which the potential-theory can be useful, even if we think that the field-theory does a better job of capturing the physical content of electromagnetic phenomena: in various situations, it is easier to solve the differential equations governing the potential (often by imposing a particular choice of gauge) than those governing the field. This idea is worth exploring, but there is a prior question we should ask: can we treat the field-theory and the potential-theory as interchangeable? That is, should we regard these as equivalent theories? We obtained the potential-theory by substituting certain expressions into the field-theory, i.e. by applying a certain kind of translation to the field-theory; but as we discussed in Chapter 2, there is no guarantee (in general) that the image of a theory under a translation will be equivalent to the original theory.

And, in fact, there are good reasons for thinking that these two theories should not be regarded as equivalent—at least, not without certain important qualifications. In Chapter 2, we saw that one of the marks of equivalent theories was the existence of an appropriate bijection between the models of those theories. We can use the definition (8.3) to induce a map from models of the potential-theory to models of the field-theory. If this map is not bijective, then that is at least some grounds for thinking that the two theories in fact have somewhat different contents.

So first: is this map surjective? This will be so if, given any model $\mathbf{F}$ of the field-theory, there is some model $\mathbf{A}$ of the potential-theory (for the same source $\mathbf{J}$) such that $\mathbf{F} = d\mathbf{A}$. When we introduced the notion of the vector potential we gave a partial answer to this question, by noting that every closed form on a contractible space is exact. So as long as we are exclusively doing electromagnetism on contractible spaces, this map is indeed surjective.

If we relax that condition, however, then we can find electromagnetic fields to which no potential corresponds. For example (using Newtonian spacetime), consider a 'mag-

netic monohole':[5] a static vacuum solution, set on Euclidean space with a point removed, where $\vec{E} = 0$ and $\vec{B}$ is radially directed away from the hole at inverse-square magnitude. (In spherical coordinates centred on the hole, this means that $\vec{B} = \frac{1}{r^2}\hat{r}$, $\hat{r}$ being the unit radial vector.)

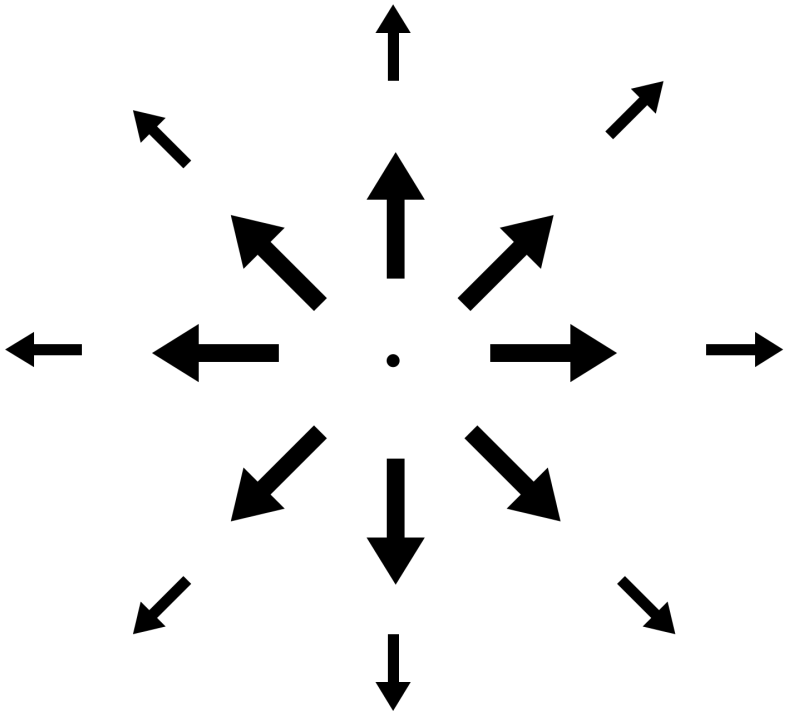


Figure 8.1.: A magnetic monohole.

This is a solution to Maxwell's equations. Yet there can be no vector potential $\vec{A}$ such that $\vec{B} = \nabla \times \vec{A}$: for any vector field $\vec{A}$, by Stokes' theorem the integral of $\nabla \times A$ over any closed surface must vanish; yet the integral of $\vec{B}$ over the unit sphere (centred on the missing point) is 4π. Hence, it cannot be the case that $\vec{B} = \nabla \times \vec{A}$.

Of course, the physical significance of this observation is highly dubious—if Maxwell's equations are correct, and if there are not random punctures in space, then an example like the above could not arise. Nevertheless, it does indicate one sense in which the potential-based theory might be thought to more strongly preclude magnetic monopoles than the field-based theory: if a magnetic monopole were to exist, then we could still use the field-based version of Maxwell's equations to represent the physical situation

[5] This is a bad pun: a magnetic monopole is a 'source' for the magnetic field, in the sense of being a point where $\mathrm{div}(\vec{B}) \neq 0$; this example is obtained by taking a magnetic monopole and removing the point where the monopole is located, hence 'monohole'.

around the monopole, whilst the potential-theory would break down more thoroughly. So insofar as we are seeking to get a handle on what these two theories say, this points to one sense in which they come apart.[6]

Second: is this map injective? We know right away that the answer must be negative, since gauge-equivalent potentials yield the same electromagnetic field. But this a bit of a trivial answer. Clearly, if gauge-equivalent potentials are regarded as physically distinct, then we shouldn't expect the two theories to have the same content. The more interesting question, therefore, is: is this map injective *as a map from gauge-equivalence classes of potentials* to fields? That is, could we have a pair of potentials $\mathbf{A}$ and $\mathbf{A}'$ which generate the same field (i.e. which are such that $d\mathbf{A} = d\mathbf{A}'$), but which are not gauge-related (i.e. for which there is no λ such that $\mathbf{A}' = \mathbf{A} + d\lambda$)? Again, if we confine attention to contractible spaces the answer to this question is 'yes': for on such a space, $d(\mathbf{A}' - \mathbf{A}) = 0$ entails that $\mathbf{A}' - \mathbf{A} = d\lambda$. Together with the observation above, this implies that over contractible spaces, the map is a bijection between gauge-equivalence classes of potentials and fields; in Chapter 12, we'll see how to strengthen this observation.

But on spaces which are not contractible, the argument just given breaks down; this opens the prospect that one might have empirically distinguishable potentials that nevertheless are associated with the same field. The best-known example of a phenomenon of this kind—the *Aharonov-Bohm effect*—is the topic of the next chapter.

[6] That said, this divergence between the two theories can be corrected if we move to the fibre-bundle formulation of the potential-based theory (discussed briefly in the following chapter), and admit bundles which are not 'globally trivialisable'.

9. The Aharonov-Bohm effect

9.1. Potentials around a solenoid

We finished the last chapter with the claim that we can have a pair of electromagnetic potentials that give rise to the same electromagnetic field, but which are *not* related by a gauge transformation: that is, which are (let us say) field-equivalent but not gauge-equivalent. We now construct an example of such a case.[1] We work in Newtonian spacetime, where a (Cartesian) coordinate system for space has been chosen.

A *solenoid* consists of a coil that is tightly wound into a helix; passing wire through the coil generates a magnetic field inside the coil (see Figure 9.1). Let I be the current

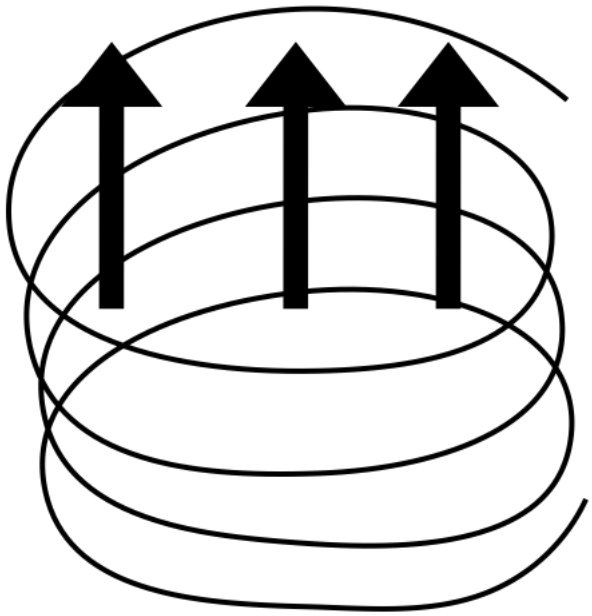


Figure 9.1.: A solenoid, showing the magnetic field.

through the coil, and let n be the number of times the coil wraps around per unit length. If we idealise such a solenoid as infinitely long, and suppose it to be oriented along the

[1] For more detailed derivations of the results discussed here, see (Feynman et al., 2011, §§13–14).

z-axis, then we find that the magnetic field inside the solenoid is $B_x = B_y = 0$, $B_z = nI$, and that the magnetic field outside the solenoid vanishes. On the other hand, the vector potential outside the solenoid does not vanish: if the radius of the solenoid is R, then the vector potential $\vec{A}$ at any point (x, y, z) is given by

$$A_x = -\frac{nIR^2 y}{2r^2} \tag{9.1}$$

$$A_y = \frac{nIR^2 x}{2r^2} \tag{9.2}$$

$$A_z = 0 \tag{9.3}$$

where $r = \sqrt{x^2 + y^2}$ (i.e. the radial distance from the solenoid).

Thus, if we confine our attention to the region outside the solenoid,

$$U := \{(r, z) : r > R, -\infty < z < \infty\} \tag{9.4}$$

then the vector potential $\vec{A}_I$ when the solenoid is switched on is field-equivalent to the vector potential $\vec{A}_0 (= \vec{0})$ when the solenoid is switched off (since in both cases, the magnetic field outside the solenoid vanishes). However, $\vec{A}_I$ is *not* gauge-equivalent to $\vec{A}_0$: that is, it is not the case that there is some scalar field λ such that $\vec{A}_I = \vec{\nabla}\lambda$. As with the magnetic monohole, we can see this by using Stokes' theorem: if it were the case that $\vec{A}_I = \vec{\nabla}\lambda$, then the line integral of $\vec{A}_I$ around a closed loop enclosing the solenoid would have to vanish (since it would be equal to the integral of $\text{curl}(\vec{\nabla}\lambda) = 0$ over the surface of the loop). However, the integral of $\vec{A}_I$ around such a loop does *not* vanish, as we can see by either direct calculation or by noting that it is equal to the flux of $\vec{B}_I$ through that loop (again, using Stokes' theorem).

As discussed in the previous chapter, this coming-apart of gauge-equivalence and field-equivalence is only possible in spaces that are not simply connected: $\vec{A}_0$ and $\vec{A}_I$ are only field-equivalent over U, not over the whole region including the interior of the solenoid. This also means that since $\vec{A}_0$ and $\vec{A}_I$ are field-equivalent over any subregion of U, they must also be gauge-equivalent over any such subregion which is simply connected. For example, consider the regions U_1 and U_2, defined as follows:

$$U_1 = \{(x, y, z) : R < r < \infty, x \geq 0, -\infty < z < \infty\} \tag{9.5}$$

$$U_2 = \{(x, y, z) : R < r < \infty, x \leq 0, -\infty < z < \infty\} \tag{9.6}$$

U_1 and U_2 are illustrated in Figure 9.1; note that $U_1 \cup U_2 = U$. We define two functions

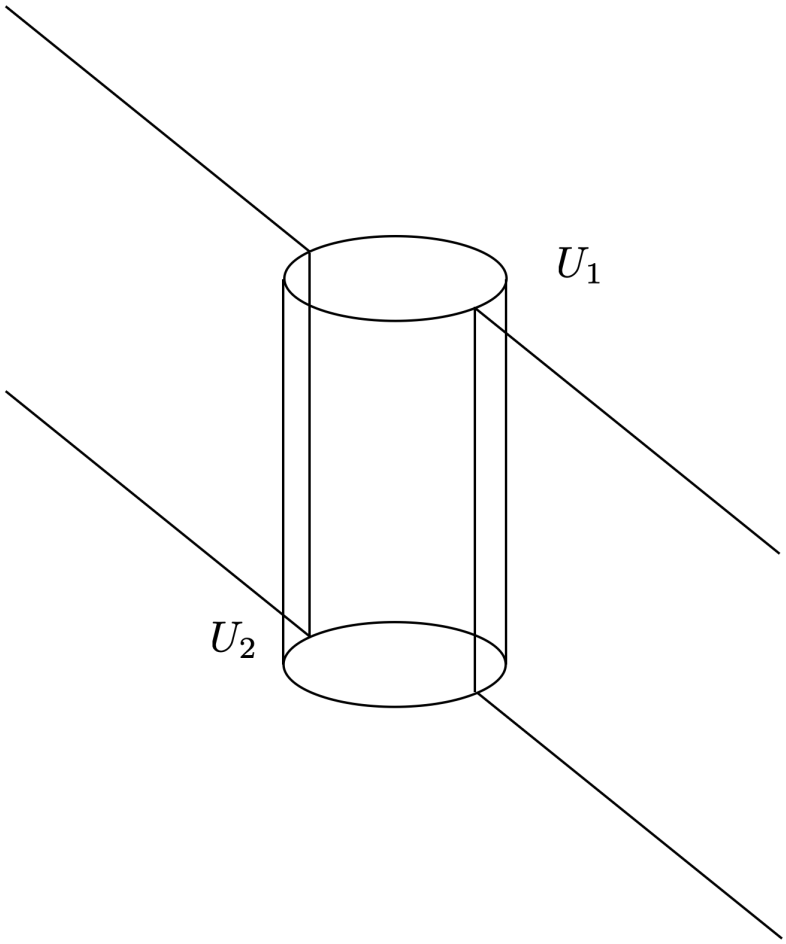


Figure 9.2.: The regions U_1 and U_2 (the solenoid and the plane both continue in the positive and negative z-direction).

$\theta_1 : U_1 \to \mathbb{R}$ and $\theta_2 : U_2 \to \mathbb{R}$, by the conditions

$$\tan(\theta_1) = \frac{y}{x}, \quad 0 \leq \theta_1 \leq \pi \tag{9.7}$$

$$\tan(\theta_2) = \frac{y}{x}, \quad \pi \leq \theta_1 \leq 2\pi \tag{9.8}$$

We have to specify the ranges for θ_1 and θ_2 in order to uniquely fix them as functions (since tan is a periodic function).

Then on U_1, $\vec{A}_I$ is given by

$$\vec{A}_I = \frac{nIR^2}{2} \operatorname{grad}(\theta_1) \tag{9.9}$$

and similarly on U_2. Hence, on either of U_1 or U_2 individually, $\vec{A}_I$ is gauge-equivalent to $\vec{A}_0$—as we expected, given that U_1 and U_2 are simply-connected subregions of U. The reason we cannot parlay this into a demonstration that $\vec{A}_I$ is gauge-equivalent to $\vec{A}_0$ over *all* of U is that the two functions θ_1 and θ_2 cannot be combined into a single smooth function: they come apart along the positive x-axis, where $\theta_1 = 0$ and $\theta_2 = 2\pi$. It follows that the definition (8.3) of the electromagnetic field in terms of the potential does *not* induce an injective map from gauge-equivalence classes of potential-models to field-models.

Hence, if gauge-equivalence is our condition of physical equivalence, then we reach the following conclusion: switching the solenoid on should not produce any empirically detectable difference in the region U_1, nor in the region U_2, but it might produce such a difference in the region U. We don't have any *guarantee* that such a difference will be detectable—but the possibility of such differences is not ruled out by our stance on gauge transformations. In fact, it turns out that although there is no known classical experiment that is capable of detecting this difference, there is a quantum experiment that can do so. Let us see how that goes.

9.2. Quantum charges in classical electromagnetism

The Aharonov-Bohm effect, as standardly derived, takes place in the context of what we might call 'semi-classical electromagnetism': the theory of a quantum point charge moving against the backdrop of a classical electromagnetic field. We are, by now, familiar with the classical electromagnetic field; in this section, I give a thumbnail sketch of what we need to know about the quantum particle. The theory governing that particle is non-relativistic, so we take the background spacetime to be Newtonian rather than Minkowskian.

For our purposes, we will take such a particle to be represented by its *wavefunction*: a function $\psi : \mathcal{T} \times \mathcal{X} \rightarrow \mathbb{C}$, where $\mathbb{C}$ is the complex plane. The dynamics of this wavefunction is given by the *Schrödinger equation*, which for a particle of charge e and mass m moving in a magnetic vector potential $\vec{A}$ (and zero electric field) is

$$\frac{\partial \psi}{\partial t} = \frac{i}{2m}\left(\nabla - ie\vec{A}\right)^2 \psi \tag{9.10}$$

In turn, the wavefunction yields a probabilistic guide to the measured location of the electron:[2] if the wavefunction at a given time t is $\psi(t)$, then the electron's probability

[2]Exactly *how* the wavefunction does this is a topic we deliberately pass over in silence.

density over possible locations is given by $|\psi|^2$.

We won't go into the details of this equation, but instead just note two things about it. First, it directly features the magnetic vector potential, not the magnetic field. Our first impression, therefore, might be that the magnetic vector potential enjoys a more significant status in this theory than in the theory of classical electromagnetism. This impression is undermined, however, by our second observation: that the Schrödinger equation (9.10) admits a local gauge symmetry of the form

$$\psi \mapsto \mathrm{e}^{ie\lambda}\psi \tag{9.11}$$

$$\vec{A} \mapsto \vec{A} + \nabla\lambda \tag{9.12}$$

where λ is an arbitrary smooth scalar field. Noting that $|\mathrm{e}^{ie\lambda}\psi|^2 = |\psi|^2$, we make the assumption that this symmetry implies empirical undetectability in the same way the other symmetries we have examined do.[3]

However, as we have just seen, there is a difference between the claim that there is no empirical difference between *gauge-equivalent* potentials, and the claim that there is no empirical difference between *field-equivalent* potentials: the latter claim is strictly stronger, at least insofar as we consider non-simply-connected spaces. And in fact, we can use this quantum particle to pry open the difference between these two claims, via a variant of the famous double-slit experiment. A beam of charged particles is separated; the two component beams pass around opposite sides of a solenoid (which we shield from the beams) before being recombined as they meet a detector screen. When the solenoid is turned off, this corresponds to the standard double-slit experiment, and we get interference fringes on the screen. When the solenoid is turned on, however, we find that there is a shift in the interference fringes, which the above theory can explain as a consequence of the phase-shift induced in the electron's wavefunction by $\vec{A}_I$ (since the wavefunction dynamics depend on $\vec{A}$ via (9.10)). The magnitude of the shift is proportional to the result of integrating $\vec{A}_I$ along the two 'arms' of the experiment and taking the difference; that is, it is proportional to the integral of $\vec{A}_I$ around a loop enclosing the solenoid (which is, as we've already discussed, equal to the total magnetic flux through the solenoid). This shift, first predicted by Aharonov and Bohm (1959)—and subsequently experimentally verified to a high degree of precision—is the *Aharonov-Bohm effect*. Note that in order to manifest the effect, we need to make use of the whole region U: that is, the effect cannot arise if our experiments are limited to one side of the solenoid or the other (i.e., to the regions U_1 and U_2).

[3]For a more detailed account of the empirical implications of quantum symmetries, see Wallace (ndc).

9.3. The reality of the fields

In most discussions of the Aharonov-Bohm effect, it is framed as a problem for a view which takes only the electromagnetic field to have 'physical reality': such a view is problematic, goes the thought, because it is committed to an unpalatable form of non-locality (whereby the magnetic field within the solenoid is able to exert an 'action at a distance' on the particle).[4] Our interests are slightly different, and hence so is our framing of the effect: for us, the effect is primarily to be understood as a vivid demonstration of the problems with regarding the field-theory and the potential-theory as equivalent, *even if* the latter is interpreted with gauge-equivalence as sufficient for physical equivalence.

That said, if these two theories are not to be considered equivalent, then we should consider them to postulate different structures; hence, there is a choice to be made between them. The fact that the fields theory is seemingly committed to this kind of non-locality seems like a good reason to prefer the potential-theory, but then we have the question: if we would like to regard gauge-equivalent states as physically equivalent, and committing to the field theory is not a good way of doing so, what should we do instead?

One alternative is to look for some other set of invariants, around which a theory can be constructed that *is* plausibly regarded as equivalent to the potentials theory (when gauge-equivalent models of the latter are interpreted as equivalent). For example, one proposal that has received a lot of attention in the philosophical literature is the so-called *holonomy* interpretation.[5] Given a magnetic vector potential $\vec{A}$, the *holonomy* of any loop in $\mathcal{X}$ is the integral of $\vec{A}$ around that loop.[6] Like the magnetic field, the holonomy of a loop is gauge-invariant; but unlike the magnetic field, specifying all the holonomies in a given region fixes the vector potential in that region to within gauge equivalence. However, although the holonomies do therefore capture the gauge-invariant data, it is not so clear how to construct a theory out of them: that is, how the equations of motion (for either the purely electromagnetic degrees of freedom, or for a charged particle coupling to them) are to be written down.

Another alternative is to resist the call to rewrite electromagnetism in terms of strict *invariants*, rather than in terms of *covariants*. That is: when discussing spacetime sym-

[4] See e.g. Healey (1997), Maudlin (1998), Belot (1998), Leeds (1999), Nounou (2003), Maudlin (2018), and references therein.

[5] See Healey (2007).

[6] There is a sense in which the magnetic field can be thought of as giving a strict subset of the holonomy data: the magnetic field at a point $x \in \mathcal{X}$ is the holonomy around an infinitesimal loop centred on x.

metries (whether of Newtonian mechanics or electromagnetism), we took those symmetries to motivate the move to a weaker spacetime setting—Newtonian spacetime rather than coordinate space, then Galilean or Minkowski spacetime rather than Newtonian spacetime. However, in a certain sense the structures that arise in Galilean spacetime are not *invariants* of the Galilean group. When we apply a Galilean transformation to a model set on Galilean spacetime, it is not that the model remains fully invariant under that transformation; rather, the model is *covariant* under it, in the sense that we obtain a new model *which nevertheless is isomorphic to the first*. So, what if we take a similar attitude toward electromagnetism?[7]

This would mean looking for a formalism in which gauge transformations, although definable, are isomorphisms. An example of such a formalism would be the so-called *fibre bundles* formalism. Unfortunately, that formalism is sufficiently technical that we won't have space to present it in detail, but *very* roughly: rather than treating fields as functions from spacetime to some fixed value-space, a fibre bundle equips each spacetime point with its *own* value-space and treats a field as mapping each spacetime point into that spacetime point's value-space; because the value-spaces associated to different spacetime points are not identified with one another, a question such as 'is the value of the field at x the same as the value of the field at y?' ceases to make sense. This is necessary, since a gauge transformation acts differently at different points of spacetime: if this is to count as an isomorphism, then we can't be allowed to ask such questions (since the answer to such a question could change, if we apply a gauge transformation that vanishes at x but not at y).

As a final observation, we note that both the holonomy and fibre-bundle formalisms are *non-separable*, in the sense that specifying the physical state in certain regions does not (in general) suffice to specify the state on the union of those regions. For example, fixing the holonomies in U_1 and U_2 does not determine the holonomies throughout U—indeed, the Aharonov-Bohm effect is based on the fact that the holonomy of a loop around the solenoid is not determined by the holonomies of loops that do not enclose the solenoid. Although we haven't said enough to really explain why, a similar thing is true for fibre bundles: fixing a model of the fibre-bundle formalism over U_1 and over U_2 does not uniquely determine a model over U.

The reason for this is that non-separability is an immediate consequence of regarding gauge equivalence as necessary and sufficient for physical equivalence (once we recognise that gauge-equivalence over subregions need not entail gauge-equivalence over the whole region). As discussed above, the potentials $\vec{A}_0$ and $\vec{A}_I$ are gauge-equivalent

[7] The difference between these two attitudes is discussed in more detail in Dewar (2019b).

over U_1 and U_2, but not over U—even though $U = U_1 \cup U_2$. Hence, we can say that the physical state of U_1 is the same regardless of whether the solenoid is on or not, and the same for the physical state of U_2; and yet, that the physical state of U *does* change depending on the solenoid. So this interpretational stance carries a commitment to non-separability with it; and that commitment will be reflected in any formalism that seeks to implement that stance.[8]

[8] That said, Wallace (2014) argues that we can have both nonlocality and gauge-invariance; how this squares with the argument here is not entirely clear to me.

Part IV.

Categories

10. Introduction to category theory

In this final part of the book, we look at some of the ways in which *category theory* offers a means of formalising some of the concepts of structure and equivalence that we have encountered so far; in particular, at how we can use categorical notions to bring together the logical and physical examples. In this chapter, we introduce the notion of a category, and look at some examples of categories.[1]

10.1. Motivation and definition

In our investigations, we have encountered various definitions of the form 'a wotsit is a set, equipped with such-and-such bells and whistles': for example, Tarski models, vector spaces, or groups. The result of such definitions is that we get a collection of 'structured sets' (e.g. groups), which have 'structure-preserving mappings' (e.g. group homomorphisms) between them. One way of approaching category theory is to note that a lot of the interesting mathematical information gets encoded in these mappings. For example, a subgroup N of a group G is a normal subgroup (i.e. is invariant under conjugation) iff there is some group homomorphism $\phi : G \to H$ such that N is the kernel of ϕ; so, roughly speaking, if you knew all the facts about group homomorphisms, then you could figure out which subgroups are normal. This motivates the study of the networks of structure-preserving mappings, and the development of a theory of such mappings: i.e., the postulation of axioms that any such collection of mappings should obey. We refer to such a network as a *category*, and axiomatise this notion as follows.

Definition 27. A *category* $\mathcal{C}$ consists of a class $|\mathcal{C}|$ of *objects* and a class $\mathrm{Hom}(\mathcal{C})$ of *arrows* (also often referred to as *morphisms*), such that:

- Every arrow $f \in \mathrm{Hom}(\mathcal{C})$ is associated with a pair of objects A and B of $\mathcal{C}$, referred to as its *domain* and *codomain*: we denote this by writing $f : A \to B$.
- For any two arrows $f : A \to B$ and $g : B \to C$, there is a third arrow from A to C called the *composition* of f and g, and denoted $g \circ f$.

[1] For more on category theory, see Awodey (2010) or Halvorson (2019).

- Composition is *associative*: given three arrows $f : A \to B$, $g : B \to C$, and $h : C \to D$,
$$h \circ (g \circ f) = (h \circ g) \circ f \tag{10.1}$$
- Associated with every object A in $\mathcal{C}$ there is an *identity arrow* $\mathrm{Id}_A : A \to A$: this arrow has the property that for any arrow $f : A \to B$, $\mathrm{Id}_B \circ f = f = f \circ \mathrm{Id}_A$.

♠

Given any two objects A and B in a category $\mathcal{C}$, the set of all arrows with domain A and codomain B will be denoted $\mathcal{C}(A, B)$.

You should satisfy yourself that these axioms seem plausible conditions for the general notion of structure-preserving mappings between structured sets. Indeed, the following are all examples of categories:

- The category **Grp**, with groups as objects and group homomorphisms as arrows
- The category **Vec**, with vector spaces as objects and linear maps as arrows whose objects are vector spaces and whose arrows are linear maps
- The category **Set**, with sets as objects and functions as arrows

(Exercise: demonstrate that the above examples are indeed categories.)

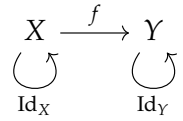
However, having postulated those axioms, we can then study the structures that obey them without regard for whether all those structures are interpretable as collections of mappings or not; this is analogous to the way that group theory postulates some axioms intended to capture the notion of a set of transformations, but then goes on to study anything satisfying those axioms without necessarily thinking of it as a set of transformations. In other words, we abstract away from 'concrete categories' such as the above, to study all algebraic structures satisfying the axioms.[2] To illustrate this, here are some further examples of categories.

10.2. Examples

For categories with finitely many objects and arrows, we can explicitly describe the structure of the category. We start with a couple of examples of categories of this kind.

[2]This is not intended as a claim about the actual history of the development of category theory; for that history, see (Marquis, 2020, §2).

Example 2. The category **2** contains two objects, which we will label X and Y; it contains the identity arrows Id_X and Id_Y (as it must) and one non-identity arrow $f : X \to Y$. We can depict this category, as follows:



2 is quite a boring category, but not the most boring: that honour belongs (arguably) to its little cousin **1**.

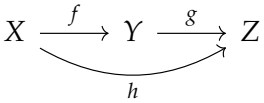
Example 3. The category **1** contains a single object X, and only the single arrow Id_X:

Let's go wild, and consider a category with *three* objects. (Incidentally, these names—**1**, **2** and **3**—are not especially canonical, so don't be surprised if other books use different names for these categories, or use these names for different categories.)

Example 4. The category **3** has three objects X, Y, Z, and contains non-identity arrows $f : X \to Y$, $g : Y \to Z$, and $h : X \to Z$, where

$$h = g \circ f \tag{10.2}$$

This category has the following diagram, where (as is usual) we no longer bother to draw the identity arrows:



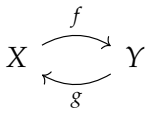
Note that if we hadn't specified the compositional relation (10.2), it would have been ambiguous how many arrows there are in the category in total; in general, specifying a category requires specifying the compositional structure among the arrows. This means that one has to be a little careful in using diagrams like the above to depict categories; in most instances, there is a (perhaps implicit) convention that the diagram commutes. Just to make this point clear, consider the following category:

Example 5. The category $\mathbf{1}'$ (whose name will be explained in the next chapter) has two objects X and Y, and non-identity arrows $f : X \to Y$ and $g : Y \to X$, where

$$g \circ f = \mathrm{Id}_X \tag{10.3a}$$

$$f \circ g = \mathrm{Id}_Y \tag{10.3b}$$

We give the diagram of this category as follows:



Here, if we did not require that the compositional relations (10.3) held, then the category would have an infinite number of arrows: all the results of composing f and g together arbitrarily many times!

In the category $\mathbf{1}'$, g and f are said to be *inverse* to one another. More generally:

Definition 28. For any arrow $f : X \to Y$ in a category $\mathcal{C}$, an arrow $g : Y \to X$ is said to be *inverse* to f if $g \circ f = \mathrm{Id}_X$ and $f \circ g = \mathrm{Id}_Y$. ♠

This leads to the category-theoretic definition of *isomorphism*: an isomorphism is just an invertible arrow.

Definition 29. An arrow $f : X \to Y$ in a category $\mathcal{C}$ is an *isomorphism* if there exists an inverse arrow $f^{-1} : Y \to X$. ♠

Note that identity arrows are self-inverse, and hence are always isomorphisms. A category like $\mathbf{1}'$ in which every arrow is an isomorphism is referred to as a *groupoid*.

We've seen already that classes of mathematical objects with structure-preserving mappings between them often constitute a category; certain mathematical objects can themselves be regarded as categories, as we now discuss.

Example 6. A *partial order* is a set X equipped with a binary relation $\leq$ that is reflexive, transitive, and antisymmetric: that is, for any $x, y \in X$,

$$x \leq x \tag{10.4}$$

$$x \leq y, y \leq z \Rightarrow x \leq z \tag{10.5}$$

$$x \leq y, y \leq x \Rightarrow x = y \tag{10.6}$$

Any partial order can be considered to be a category, with the objects being the elements of X, and with (exactly one) arrow between any pair of objects that stand in the relation $\leq$. The composition of an arrow from x to y with an arrow from y to z is defined as the arrow from x to z, which is guaranteed to exist by transitivity. For any x, we take Id_x to be the arrow from x to itself (whose existence is guaranteed by reflexivity).

The category $\mathbf{2}$ can be regarded as a category of this kind, arising from the two-element partial order where one element is greater than the other.

Example 7. A *preorder* is a set X equipped with a binary relation $\lesssim$ that is reflexive and transitive, but not (necessarily) antisymmetric: so there can be $x \neq y$ in X such that $x \lesssim y$ and $y \lesssim x$. As with a partial order, any preorder can be regarded as a category with exactly one arrow between any pair of objects that stand in the relation $\lesssim$.

Example 8. Any set X can be regarded as a category: one with no arrows other than the identity arrows. (Such a category is referred to as a *discrete* category.)

To be clear, the fact that any set can be regarded as a (discrete) category is distinct from the fact that there is a category **Set** of sets: it is both the case that any *individual* set can be regarded as a category, and that the collection of *all* sets forms a category.[3]

Example 9. A group G can be considered to be a category, with exactly one object and the group elements being the arrows (all of them arrows from that one object to itself). Composition of arrows is identified with group multiplication, and the identity arrow is identified with the group identity element; the group axioms then guarantee that the categorical axioms are satisfied.

Since every element of a group has in inverse, a group (regarded as a category) is one where every arrow is an isomorphism—i.e., a groupoid.

As with sets, the fact that every group can be regarded as a category is distinct from the fact that there is a category **Grp** of groups; rather, the point is that an individual group (say, the Lorentz group) may be regarded as a category. Note that when we regard a set as a category, the elements of the set get represented as objects; by contrast, when we regard a group as a category, the elements of the group get represented as arrows. This reinforces the point that in many (arguably, most) categories, it is the arrows that carry the interesting structure. Our final example also demonstrates this point (and will be discussed further in the next chapter).

Example 10. We define a category **Mat** as follows. The objects of **Mat** are the natural numbers, and the arrows in **Mat** from m to n are (all) the real $n \times m$ matrices; composition of arrows is given by matrix multiplication (i.e. given $M : m \to n$ and $N : n \to p$, $N \circ M = NM$) and the identity arrows are the identity matrices.[4] This is a category: if M is an $n \times m$ matrix and N is a $p \times n$ matrix, then NM is a $p \times m$ matrix; and matrix multiplication is an associative operation, for which the identity matrices act as identities.

[3] Similarly, there is a category **Pos** which has all partially ordered sets as its objects, and order-preserving maps as its arrows; this should not be confused with the fact that any partially ordered set can be regarded as a category.

[4] See Appendix A.

11. Functors between categories

11.1. Functors

In considering mathematical objects of a certain kind, one often wants to know how they relate to other kinds of mathematical object. In the context of category theory, the standard tool for making these kinds of comparison is the *functor*. A functor is a 'homomorphism of categories'; its precise definition is as follows.

Definition 30. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. A *functor* F from $\mathcal{C}$ to $\mathcal{D}$ consists of

1. a map $F_{\text{obj}} : |\mathcal{C}| \to |\mathcal{D}|$; and
2. for every $A, B \in |\mathcal{C}|$, a map $F_{AB} : \mathcal{C}(A, B) \to \mathcal{D}(F_{\text{obj}}(A), F_{\text{obj}}(B))$

such that the following two conditions hold:

- For any $f : A \to B$ and $g : B \to C$ in $\mathcal{C}$,

$$F_{AC}(g \circ f) = F_{BC}(g) \circ F_{AB}(f) \tag{11.1}$$

- For any $A \in |\mathcal{C}|$,

$$F_{AA}(\mathrm{Id}_A) = \mathrm{Id}_{F_{\text{obj}}(A)} \tag{11.2}$$

♠

In the interests of reducing notational clutter, we will typically just use F to denote the maps F_{obj} and F_{AB} (for any $A, B \in |\mathcal{C}|$); the argument of the map will make it clear which one is meant. Hence, for example, the equations (11.1) and (11.2) can be written as

$$F(g \circ f) = F(g) \circ F(f) \tag{11.3}$$

$$F(\mathrm{Id}_A) = \mathrm{Id}_{F(A)} \tag{11.4}$$

To illustrate the idea, let's look at some examples of functors. Here is a functor $F : \mathbf{2} \to \mathbf{1}'$:

$$F(A) = X \tag{11.5}$$
$$F(B) = Y \tag{11.6}$$
$$F(f) = j \tag{11.7}$$

Note that we don't need to specify what F does to the identity arrows: once we've specified F's action on objects, it must sent Id_X to $\mathrm{Id}_{F(X)}$, by condition (11.2). In the other direction, here's a functor $G : \mathbf{1}' \to \mathbf{2}$:

$$G(X) = A \tag{11.8}$$
$$G(Y) = A \tag{11.9}$$
$$G(j) = G(k) = \mathrm{Id}_A \tag{11.10}$$

Naturally, these examples are a bit trivial. Here is a less trivial example: as discussed in Appendix B, every vector space may be regarded as a group (with vector addition as the group operation). This means that there is a functor from the category **Vec** of vector spaces to the category **Grp** of groups: it maps any vector space to itself, or—perhaps better—to the 'copy' of itself that lives in the category of groups.

As already mentioned, functors are structure-preserving mappings between categories. As we discussed in the last chapter, collections of objects with structure-preserving mappings between them are paradigm cases of categories, and this example is no different. Indeed, if we compose two functors $F : \mathcal{C} \to \mathcal{D}$ and $G : \mathcal{D} \to \mathcal{E}$, then we find that the result is a functor $G \circ F : \mathcal{C} \to \mathcal{D}$. Moreover, this composition operation is associative ($H \circ (G \circ F) = (H \circ G) \circ F$), and for every category $\mathcal{C}$, there is an *identity functor* $\mathrm{Id}_{\mathcal{C}} : \mathcal{C} \to \mathcal{C}$ which maps every object and arrow in $\mathcal{C}$ to itself (you should convince yourself that this does, indeed, satisfy the definition of a functor). So there is a *category of categories*, denoted **Cat**, whose objects are categories and whose arrows are functors.

11.2. Equivalence functors

However, this does not mean that any two categories related by a functor should be regarded as equivalent, any more than the existence of a homomorphism between groups means those should be regarded as equivalent. So when should we regard two categories as structurally equivalent? One natural proposal is to do so when they are *iso-*

morphic, which we can cash out in category-theoretic terms (given that, as we've just observed, there is a category of categories):

Definition 31. Given categories $\mathcal{C}$ and $\mathcal{D}$, a functor $F : \mathcal{C} \to \mathcal{D}$ is an *isomorphism of categories* if there is a functor $G : \mathcal{D} \to \mathcal{C}$ such that $G \circ F = \mathrm{Id}_{\mathcal{C}}$ and $F \circ G = \mathrm{Id}_{\mathcal{D}}$. ♠

However, for many purposes a somewhat weaker notion is useful: that of *categorical equivalence*. Indeed, as we will discuss in the next chapter, the fact that categories admit this interestingly weaker notion accounts for much of the interest in category theory as a way of making precise certain ideas about equivalence in philosophy of science. Essentially, an equivalence functor is an 'isomorphism up to isomorphism'; more precisely, the definition is as follows.[1]

Definition 32. Given categories $\mathcal{C}$ and $\mathcal{D}$, a functor $F : \mathcal{C} \to \mathcal{D}$ is an *equivalence of categories* if F is *full*, *faithful*, and *essentially surjective*, where:

(i) F is *full* if for any objects A and B of $\mathcal{C}$, $F_{AB} : \mathcal{C}(A, B) \to \mathcal{D}(F(A), F(B))$ is surjective.

(ii) F is *faithful* if for any objects A and B of $\mathcal{C}$, $F_{AB} : \mathcal{C}(A, B) \to \mathcal{D}(F(A), F(B))$ is injective.

(iii) F is *essentially surjective* if for any object X of $\mathcal{D}$, there is some object A of $\mathcal{C}$ such that $F(A)$ is isomorphic (in $\mathcal{D}0$ to X.

♠

Perhaps the most significant difference between isomorphism and equivalence of categories is that two categories can be equivalent *even if they have different numbers of objects*. For example, the following functor is an equivalence from $\mathbf{1}'$ to $\mathbf{1}$:

$$F(X) = F(Y) = 1 \tag{11.11}$$

$$F(j) = F(k) = \mathrm{Id}_1 \tag{11.12}$$

This functor is surjective, hence essentially surjective; and although F is not bijective on arrows overall, it *is* bijective on the arrows between any chosen pair of objects. In the

[1] A more conceptually revealing (but less readily applicable) definition is that a functor $F : \mathcal{C} \to \mathcal{D}$ is an equivalence if there is an 'almost inverse' functor $G : \mathcal{D} \to \mathcal{C}$: a functor such that $G \circ F$ is 'naturally isomorphic' to $\mathrm{Id}_{\mathcal{C}}$ and $F \circ G$ is 'naturally isomorphic' to $\mathrm{Id}_{\mathcal{D}}$. This makes the relationship to categorical isomorphism clearer, but requires introducing natural transformations, which we don't have the space to do here. For discussion (and a proof that these two notions coincide), see (Awodey, 2010, chap. 7).

other direction, the following functor is an equivalence from **1** to **1**′:

$$G(1) = X \tag{11.13}$$

(The fact that $G(\mathrm{Id}_1) = \mathrm{Id}_X$ is implied by functoriality.) This functor is not surjective; but it is *essentially* surjective, since Y is isomorphic to X.

Here are two further examples of equivalences, which again illustrate the fact that equivalence can be weaker than isomorphism.

Proposition 9. Any preorder X is categorically equivalent to some poset (where both are regarded as categories).

Proof. Two objects x and y of X are isomorphic if $x \lesssim y$ and $y \lesssim x$. We define the corresponding poset by first quotienting X by isomorphism, to obtain a set Y: that is, elements of Y are equivalence classes of isomorphic elements of X. Let us denote the equivalence class containing $x \in X$ as $[x]$. We then define a binary relation $\leq$ on Y, according to

$$[x] \leq [y] \Leftrightarrow x \lesssim y \tag{11.14}$$

The transitivity of $\lesssim$ guarantees that this definition is well-posed, in the sense of being independent of the choice of x and y from within an equivalence class. Furthermore, the reflexivity and transitivity of $\lesssim$ entail the reflexivity and transitivity of $\leq$. It remains only to confirm that $\leq$ is anti-symmetric. Indeed, if $[x] \leq [y]$ and $[y] \leq [x]$, then $x \lesssim y$ and $y \lesssim x$; hence x and y are isomorphic, and so $[x] = [y]$. □

Proposition 10. The category **FinVect** of finite-dimensional vector spaces (with linear maps as arrows) is equivalent to the category **Mat** (of natural numbers with matrices as arrows).

Proof. First, equip every vector space $\mathbb{V}$ in **FinVect** with an (arbitrary) ordered basis $e_i^{\mathbb{V}}$.[2] Now define a functor F : **FinVect** $\rightarrow$ **Mat** as follows. For any finite-dimensional vector space $\mathbb{V}$, $F(\mathbb{V})$ is the dimension of $\mathbb{V}$. For any linear transformation $f : \mathbb{V} \rightarrow \mathbb{W}$, where $\dim(\mathbb{V}) = m$ and $\dim(\mathbb{W}) = n$, $F(f)$ is the $n \times m$ matrix representing f relative to those bases:[3] thus,

$$f(e_i^{\mathbb{V}}) = F(f)^j{}_i e_j^{\mathbb{W}} \tag{11.15}$$

[2]Strictly, I ought to write $\vec{e}_i^{\mathbb{V}}$, but that looks terrible; I trust the reader to remember that these objects are vectors.

[3]See Appendix A.

First, we show that F is indeed a functor. If we have two linear transformations $f : \mathbb{U} \to \mathbb{V}$ and $g : \mathbb{V} \to \mathbb{W}$, then

$$\begin{aligned} F(g \circ f)^k{}_i e_k^{\mathbb{W}} &= g(f(e_i^{\mathbb{U}})) \\ &= g(F(f)^j{}_i e_j^{\mathbb{V}}) \\ &= F(g)^k{}_j F(f)^j{}_i e_i^{\mathbb{W}} \end{aligned}$$

from which it follows that $F(g \circ f)^k{}_i = F(g)^k{}_j F(f)^j{}_i = (F(g)F(f))^k{}_i$, i.e. that F preserves composition of arrows. Furthermore, if I is the identity transformation on $\mathbb{V}$, then

$$e_j^{\mathbb{V}} = F(I)^i{}_j e_i^{\mathbb{V}} \tag{11.16}$$

for which the only solution is $F(I)^i{}_j = \delta^i{}_j$; thus, F preserves identity arrows.

We now show that F is an equivalence functor. First, for any natural number n, there is some vector space $\mathbb{V}$ in **FinVect** of dimension n; so F is essentially surjective (indeed, surjective).

Second, consider any vector spaces $\mathbb{V}$ and $\mathbb{W}$ in **FinVect**. If f and g are arrows from $\mathbb{V}$ to $\mathbb{W}$ in **FinVect** such that $F(f) = F(g)$, then for every basis vector $e_i^{\mathbb{V}}$ in $\mathbb{V}$,

$$\begin{aligned} f(e_i^{\mathbb{V}}) &= F(f)^j{}_i e_j^{\mathbb{W}} \\ &= F(g)^j{}_i e_j^{\mathbb{W}} \\ &= g(e_i^{\mathbb{V}}) \end{aligned}$$

Thus, f and g agree on the basis vectors, and hence on all vectors: that is, $f = g$. So F is faithful.

Finally, again consider any vector spaces $\mathbb{V}$ and $\mathbb{W}$ in **FinVect**; suppose that their dimensions are m and n respectively. Then for any $n \times m$ matrix $M^j{}_i$, define a linear transformation $f : \mathbb{V} \to \mathbb{W}$ be defined by the condition that for any basis vector $e_i^{\mathbb{V}}$ in $\mathbb{V}$,

$$f(e_i^{\mathbb{V}}) = M^j{}_i e_j^{\mathbb{W}} \tag{11.17}$$

It follows immediately that $F(f) = M$. So F is full.

□

One might feel somewhat disquieted by this example: on the face of it, it appears to suggest that a vector space has the same structure as a natural number, which seems

either false or nonsensical.[4] However, we need to be careful, since (as was discussed in Chapter 10) categories will often encode their most interesting structure in their networks of arrows. And in fact, in this case, we can recover the internal vector-space structure from the categorical structure: the vectors in an n-dimensional space $\mathbb{V}$ may be identified with the set of linear maps from a 1-dimensional vector space to $\mathbb{V}$, and the vector-space operations (addition and scalar multiplication) can be defined using categorical data.[5] This enables us to recover vectors from the category **FinVect**, or—what comes to the same thing—the category **Mat**.

11.3. Forgetful functors

A functor which is *not* an equivalence functor is referred to as a *forgetful functor*.[6] A nice way to be precise about the concept of a forgetful functor, and about what it is that they are forgetting, is the so-called "stuff, structure, properties" perspective. The intuition here is that there are three ways in which a mathematical object can be interesting, and hence three ways of making it less interesting:

> In math we're often interested in equipping things with extra structure, stuff, or properties .... For example, a group is a set (*stuff*) with operations (*structure*) such that a bunch of equations hold (*properties*).[7]

In other words, the *stuff* is the raw materials out of which the object is 'made'; the *structure* comprises the relations and properties which organise the stuff into an interesting mathematical object; and the *properties* are the properties that the object exhibits, in virtue of having its stuff organised thus-and-so by its structure. Philosophers would identify the stuff as the 'ontology', the structure as the 'ideology', and the properties as the 'facts'. Given a Σ-model $\mathfrak{A}$, the stuff would be the domain $|\mathfrak{A}|$, the structure the signature Σ (or perhaps the set of formulae $\mathrm{Form}(\Sigma)$), and the properties would be the sentences that $\mathfrak{A}$ satisfies (or perhaps all the facts about which formulae the tuples of $\mathfrak{A}$ satisfy).

Category theory offers a nice way of formalising this distinction (that's somewhat more general than the model-theoretic formalisation). Given a functor $F : \mathcal{C} \to \mathcal{D}$, we say that F:

[4] See Hudetz (2019) for an elaboration of this concern.
[5] Dewar (nd)
[6] At least, from a certain perspective. In the literature, the term "forgetful functor" is frequently used with a more vague denotation.
[7] (Baez and Shulman, 2010, p. 15)

- forgets *at most properties* if it is full and faithful;
- forgets *at most structure* if it is faithful;
- forgets *at most stuff* if it is arbitrary.

Note that the stuff-structure-properties distinction is somewhat hierarchical in nature. Without structure, an object could not satisfy any (interesting) properties; and without stuff, an object would not be able to exhibit any (interesting) structure. Correspondingly, each of these levels of forgetfulness subsumes the one above: if a functor forgets stuff, then it might also forget structure and properties; and if it forgets structure, then it might also forget properties. To see how this classification of functors fits with the intuitive stuff-structure-properties distinction, let's consider some examples.[8]

First, suppose that a functor $F : \mathcal{C} \to \mathcal{D}$ forgets properties: i.e., that it is full and faithful, but not essentially surjective. That means that there are objects in $\mathcal{D}$ that lie outside the image of F, even up to isomorphism: that is, which are not isomorphic to any object in the image of F. For example, let **AbGrp** be the category of *Abelian* groups, i.e., groups whose group multiplication operation is commutative, with group homomorphisms as arrows. Every Abelian group is, of course, a group: so there is a functor $F : \mathbf{AbGrp} \to \mathbf{Grp}$, which simply maps every Abelian group to itself (or, if you prefer, to its "copy" in the category **Grp**). This functor is full and faithful: since F simply embeds **AbGrp** within **Grp**, for any Abelian groups G and H, $\mathbf{AbGrp}(G, H) = \mathbf{Grp}(F(G), F(H))$. But it is not essentially surjective: no non-Abelian group is isomorphic to any Abelian group, and only Abelian groups lie in the image of F. Thus, F forgets properties: it forgets the property of being Abelian.

Second, consider a functor $F : \mathcal{C} \to \mathcal{D}$ which forgets structure (and properties) but not stuff: that is, which is faithful but not full. That means that there are objects X and Y of $\mathcal{C}$ such that the induced function $F : \mathcal{C}(X, Y) \to \mathcal{D}(F(X), F(Y))$ is injective but not surjective. So (intuitively) there are more arrows between $F(X)$ and $F(Y)$ in $\mathcal{D}$ than there are between X and Y in $\mathcal{C}$. Insofar as we are taking the arrows to represent structure-preserving maps, a pair of objects will admit more arrows by virtue of being less structured: hence, the sense in which F forgets structure. As an example, consider the functor $F : \mathbf{Grp} \to \mathbf{Set}$ which maps any group to its underlying set, and any group homomorphism to its corresponding function. If two homomorphisms $f, g : G \to H$ are distinct from one another, then they must correspond to different functions between the underlying sets $|G|$ and $|H|$; hence, F is faithful. But (in general) there are many

[8] The below sequence of examples is taken from `https://ncatlab.org/nlab/show/stuff,+structure,+property`.

functions between $|G|$ and $|H|$ that do *not* correspond to homomorphisms, and so F is not full. Thus, F forgets structure: it forgets group structure.

Finally, consider a functor $F : \mathcal{C} \to \mathcal{D}$ which forgets stuff: that is, which is not faithful. Then there are objects X and Y of $\mathcal{C}$ such that the induced function $F : \mathcal{C}(X, Y) \to \mathcal{D}(F(X), F(Y))$ is not injective: that is, that for some pair of arrows $f, g : X \to Y$ such that $f \neq g$, $F(f) \neq F(g)$. Intuitively, then, the arrows between $F(X)$ and $F(Y)$ are less "fine-grained" than those between X and Y: there are more ways of mapping X to Y than there are ways of mapping $F(X)$ to $F(Y)$. Given that more stuff provides more "raw material" for a mapping, getting rid of stuff reduces the number of ways such a mapping could be performed. As a (somewhat trivial) example, consider the (unique) functor $F : \mathbf{Set} \to \mathbf{1}$; this functor sends every set to 1, and every function to Id_1. Thus, F forgets stuff: it forgets the stuff that makes up the sets.

12. Categories of theories

There are many things that one can do with category theory. Our interest in the topic is rather more specific: we are interested in using category-theoretic tools to illuminate the concepts of structure and equivalence that we have been discussing so far. As such, this final chapter considers how some of what we have already done can be put into a category-theoretic form. The first section considers how to apply category theory to the model-theoretic considerations of Part I, and the second looks at how we can use category theory in the context of the physical theories in Parts II and III.

12.1. Categories of Tarski-models

In Part I, we considered the concept of a class of models of a first-order theory. Now that we have category-theoretic resources to hand, we can work with the richer concept of a *category* of models.[1]

Definition 33. Given a theory T, the *category of models* of T is a category whose objects are models of T and whose arrows are elementary embeddings. ♠

As we discussed in Chapter 1, there are various kinds of mappings between models that one might work with in model theory. So why do we choose elementary embeddings to be the arrows in our category of models? The reason is that if we do so, then translations between theories induce functors between their categories of models:

Proposition 11. Given theories T_1 and T_2, let $\mathbf{Mod}(T_1)$ and $\mathbf{Mod}(T_2)$ be their categories of models. If $\tau : T_1 \to T_2$ is a translation, then τ^* is extendable to a functor, by stipulating that for any elementary embedding $h : \mathfrak{A} \to \mathfrak{B}$ (where $\mathfrak{A}, \mathfrak{B} \in \mathrm{Mod}(T_2)$), $\tau^*(h) = h$.

Proof. Suppose that $\tau^*(h)$, i.e. h, is not an elementary embedding of $\tau^*(\mathfrak{A})$ into $\tau^*(\mathfrak{B})$: that is, that there is some Σ_1-formula $\phi(x_1, \ldots, x_n)$ and some $a_1, \ldots, a_n \in |\mathfrak{A}|$ such that $\tau^*(\mathfrak{A}) \models \phi[a_1, \ldots, a_n]$ but $\tau^*(\mathfrak{B}) \not\models \phi[h(a_1), \ldots, h(a_n)$. It follows that $\mathfrak{A} \models \tau(\phi)[a_1, \ldots, a_n]$,

[1] The ideas in this section are covered in much greater detail in Halvorson (2019).

and $\mathfrak{B} \not\models \tau(\phi)[h(a_1), \ldots, a_n]$; so h is not an elementary embedding of $\mathfrak{A}$ into $\mathfrak{B}$. It follows that given an elementary embedding h in $\mathbf{Mod}(T_2)$, $\tau^*(h)$ is an elementary embedding in $\mathbf{Mod}(T_1)$. □

Had we taken homomorphisms or embeddings as arrows, this would not hold true, as the following examples demonstrate.

Example 11. Consider the following theories, T_1 and T_2. T_1 has signature $\{P\}$, and axioms

$$\exists x \exists y (x \neq y \wedge \forall z (z = x \vee z = y)) \tag{12.1a}$$

$$\exists x P x \tag{12.1b}$$

whilst T_2 has signature $\{Q\}$, and axioms

$$\exists x \exists y (x \neq y \wedge \forall z (z = x \vee z = y)) \tag{12.2a}$$

$$\exists x \neg Q x \tag{12.2b}$$

Figure 12.1 displays the models of these theories: the models of T_1 are $\mathfrak{A}_1$ and $\mathfrak{B}_1$, and the models of T_2 are $\mathfrak{A}_2$ and $\mathfrak{B}_2$.

The map

$$Qx \mapsto \neg Px \tag{12.3}$$

is a translation from T_2 to T_1; the associated semantic map will map $\mathfrak{A}_1$ to $\mathfrak{A}_2$ and $\mathfrak{B}_1$ to $\mathfrak{B}_2$. But there is a homomorphism from $\mathfrak{A}_1$ to $\mathfrak{B}_1$ yet no homomorphisms from $\mathfrak{A}_2$ to $\mathfrak{B}_2$; so there can be no functor from the category of models of T_1 with homomorphisms as arrows to the category of models of T_2 with homomorphisms as arrows.

Example 12. Let T_1 be the theory in signature $\{P^{(1)}\}$ with the following axioms:

$$\exists x \exists y (\forall z (z = x \vee z = y)) \tag{12.4}$$

$$\exists x (Px \wedge \forall y (Py \to y = x)) \tag{12.5}$$

In English: there are at most two things, of which exactly one is P. Let T_2 be the theory in signature $\{R^{(2)}\}$ with the following axioms:

$$\exists x \exists y (\forall z (z = x \vee z = y)) \tag{12.6}$$

$$(\forall x \forall y (Rxy \leftrightarrow x = y) \leftrightarrow \exists y \exists z (y \neq z)) \tag{12.7}$$

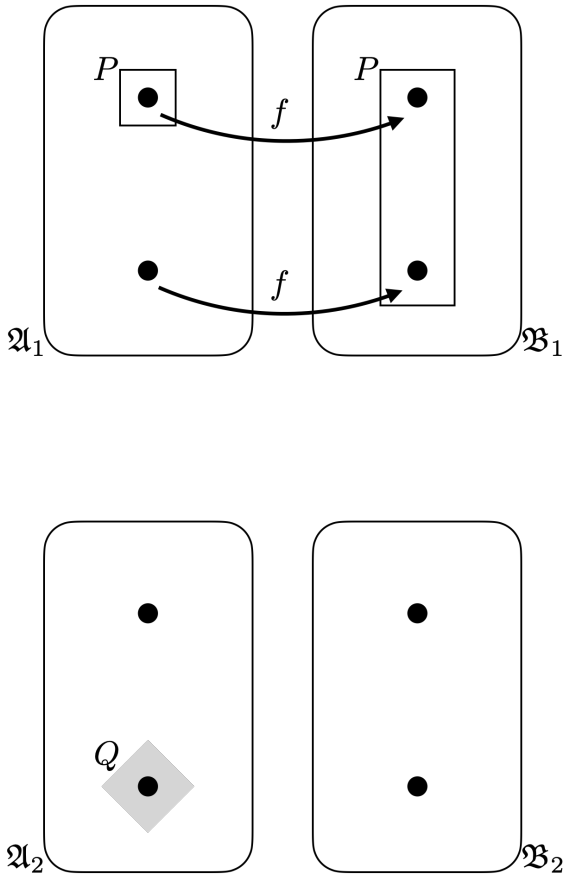


Figure 12.1.: The models of the theories in Example 1.

In English: there are at most two things, and R is reflexive just in case there are two things. The models of T_1 and T_2 are depicted and labelled in Figure 12.2.

The map

$$Rxy \mapsto (x = y \wedge \exists z \neg Pz) \tag{12.8}$$

is a translation from T_2 to T_1. However, although there is an embedding h from $\mathfrak{A}_1$ to $\mathfrak{B}_1$, there are no embeddings from $\mathfrak{A}_2$ to $\mathfrak{B}_2$; hence, there can be no functor from the category of models of T_1 with embeddings as arrows to the category of models of T_2 with embeddings as arrows.

Thus, we take the category of models of a theory to be one with elementary embeddings as arrows. We have already seen that invertible translations induce bijective maps on models; since the extension of such a map to a functor just acts as the identity on elementary embeddings, it follows more or less immediately that the functor is an isomorphism between the categories of models—and hence, that it is an equivalence.

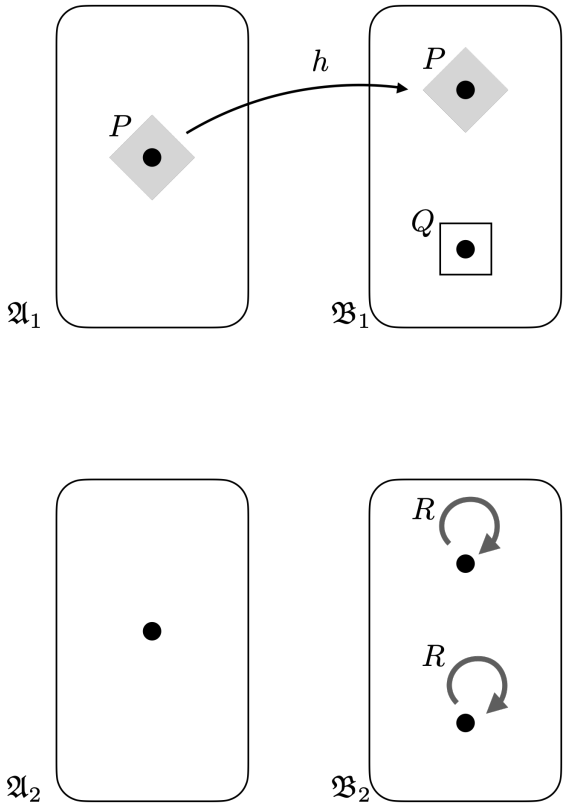


Figure 12.2.: The models of the theories in Example 2.

Thus, intertranslatability entails categorical equivalence.

In the other direction, categorical equivalence does *not* entail intertranslatability, as the following example (due to Barrett and Halvorson (2016b)) demonstrates.

Example 13. T_1 is in the signature $\Sigma_1 = \{P_0^{(1)}, P_1^{(1)}, \dots\}$, and consists of the sole axiom

$$\exists x \forall y (y = x) \tag{12.9}$$

T_2 is in the signature $\Sigma_2 = \{Q_0^{(1)}, Q_1^{(1)}, \dots\}$, and consists of the axioms

$$\exists x \forall y (y = x) \tag{12.10}$$
$$\forall y (Q_0 y \to Q_1 y) \tag{12.11}$$
$$\forall y (Q_0 y \to Q_2 y) \tag{12.12}$$
$$\vdots$$

That is, both T_1 and T_2 assert that there is exactly one thing; but T_2 asserts that the

predicate Q_0 is such that if the unique thing satisfies the predicate Q_0, then it satisfies every other predicate Q_i.

If we identify isomorphic models then the categories of both T_1 and T_2 are *discrete*, i.e., contain no arrows that are not identity arrows. Moreover, any model of of T_1 or T_2 can be specified by a subset of $\mathbb{N}$: we include n in the subset just in case the model satisfies $\exists x P_n x$ (for models of T_1), or $\exists x Q_n x$ (for models of T_2). In the other direction, any subset of $\mathbb{N}$ determines a model of T_1, whilst $\mathbb{N}$ itself and any subset of $\mathbb{N} \setminus \{0\}$ determines a model of T_2. There are $\aleph_0$ members of $\mathbb{N}$, and $\aleph_0$ members of $\mathbb{N} \setminus \{0\}$; so up to isomorphism, there are $2^{\aleph_0}$ models of T_1 and $2^{\aleph_0}$ models of T_2. Since $\mathbf{Mod}(T_1)$ and $\mathbf{Mod}(T_2)$ are discrete, any bijection between them is an equivalence of categories.

However, T_1 and T_2 are not intertranslatable. Suppose that they were, with inverse translations $\tau : T_1 \to T_2$ and $\sigma : T_2 \to T_1$. Let $\mathfrak{B}$ be the model of T_2 that corresponds to $\mathbb{N}$, i.e. that satisfies $\exists x Q_i x$ for all $i \in \mathbb{N}$. Now consider the sentence $\exists x Q_0 x$. We know that $\mathfrak{B} \models \exists x Q_0 x$, and that (up to isomorphism) it is the only model of T_2 to do so. So $\tau^* \mathfrak{B} \models \sigma(\exists x Q_0 x)$, since τ and σ are inverse to one another. Since $\sigma(\exists x Q_0 x)$ is a Σ_1-formula of finite length, there must exist some $P_i \in \Sigma_1$ which does not occur in it. So now let $\mathfrak{A}$ be the model obtained from $\tau^* \mathfrak{B}$ by 'switching' the value of P_i: if the sole object satisfies P_i in $\tau^* \mathfrak{B}$ then it does not in $\mathfrak{A}$, and vice versa (and otherwise, $\tau^* \mathfrak{B}$ and $\mathfrak{A}$ are identical). Since P_i does not occur in $\sigma(\exists x Q_0 x)$, it follows that $\mathfrak{A} \models \sigma(\exists x Q_0 x)$. So $\sigma^* \mathfrak{A} \models \exists x Q_0 x$; and hence, $\sigma^* \mathfrak{A} = \mathfrak{B}$. But since $\sigma^*(\tau^* \mathfrak{B}) = \mathfrak{B}$, it follows that σ^* is not injective, and hence not bijective; so τ and σ cannot be a pair of inverse translations after all.

Therefore, categorical equivalence of theories is a strictly weaker notion than intertranslatability. Whether this is a feature or a bug is a thorny question: it will depend on whether the theories T_1 and T_2 in the above example 'ought' to be regarded as equivalent or not, and the answer to that is not obvious. In the meantime, there is interesting work to be done in further clarifying the relationship between definability or translatability on the one hand, and categorical structure on the other. For example, Hudetz (2019) outlines a way to strengthen categorical equivalence so as to make it sensitive to the definability of the models of one theory in terms of the other, and shows that this is equivalent to intertranslatability 'up to surplus structure' (in a sense that is made precise). On the other hand, Barrett (nd) shows that if an equivalence functor is induced from a translation, then the models of the two theories are are codeterminate with one another (although in a sense weaker than that arising from full intertranslatability).

12.2. Categories of electromagnetic models

Let us now consider categories of models in physics; in the interests of space I will only discuss electromagnetism, but the main ideas here could also be applied to Newtonian mechanics. To the theory of electromagnetic fields we associate a category $\mathcal{F}$, and to the theory of electromagnetic potentials we associate a category $\mathcal{A}$, where

- An object of $\mathcal{F}$ is a 2-form $\mathbf{F}F$ on Minkowski spacetime $\mathcal{M}$ that satisfies Maxwell's equations (7.12); and
- An arrow in $\mathcal{F}$ between the 2-forms $\mathbf{F}$ and $\mathbf{F}'$ is an isometry $\psi : \mathcal{M} \to \mathcal{M}$, such that $\mathbf{F}' = \psi^* \mathbf{F}$.
- An object in $\mathcal{A}$ is a 1-form $\mathbf{A}$ on Minkowski spacetime that satisfies equation (8.4); and
- An arrow in $\mathcal{A}$ between the 1-forms $\mathbf{A}$ and $\mathbf{A}'$ consists of an isometry $\psi : M \to M$ and an exact one-form $\mathbf{G}$ on M, such that

$$\mathbf{A}' = \psi^* \mathbf{A} + \mathbf{G} \tag{12.13}$$

In other words, the category $\mathcal{F}$ represents field-theoretic solutions of Maxwell's equations, with isometries regarded as isomorphisms; and the category $\mathcal{A}$ represents potential-theoretic solutions of Maxwell's equations, with isometries and gauge transformations regarded as isomorphisms. In Chapter 8 I suggested that so long as we confined our attention to contractible spaces—such as Minkowski spacetime—there is a sense in which the theory of electromagnetic fields is equivalent to the gauge-invariant content of the theory of potentials. The following proposition, due to Weatherall (2016), can be thought of as a way of making this idea precise.

Proposition 12 (Weatherall (2016), Proposition 5.5)**.** The categories $\mathcal{F}$ and $\mathcal{A}$ are equivalent.

Proof. We define a functor $W : \mathcal{A} \to \mathcal{F}$ as follows. First, given any object $\mathbf{A}$ in $\mathcal{A}$, we define

$$W(\mathbf{A}) := d\mathbf{A} \tag{12.14}$$

Second, given any arrow $(\psi, \mathbf{G})$ in $\mathcal{A}$, we define

$$W(\psi, \mathbf{G}) = \psi \tag{12.15}$$

We need to verify that W is indeed a functor. Given a pair of objects $\mathbf{A}$ and $\mathbf{A}'$ in $\mathcal{A}$ such that

$$\mathbf{A}' = \psi^* \mathbf{A} + \mathbf{G} \tag{12.16}$$

then they are mapped by W to, respectively

$$\begin{aligned} \mathbf{F} &= d\mathbf{A} \\ \mathbf{F}' &= d\mathbf{A}' \\ &= d(\psi^* \mathbf{A} + \mathbf{G}) \end{aligned}$$

Since ψ is an isometry, $d(\psi^* \mathbf{A}) = \psi^*(d\mathbf{A})$; and since $\mathbf{G}$ is exact, $d\mathbf{G} = 0$. Hence,

$$\begin{aligned} \mathbf{F}' &= \psi^*(d\mathbf{A}) \\ &= \psi^* \mathbf{F} \end{aligned}$$

Hence, if $(\psi, \mathbf{G})$ is an arrow from $\mathbf{A}$ to $\mathbf{A}'$, then $W(\psi, \mathbf{G}) = \psi$ is an arrow from $W(\mathbf{A})$ to $W(\mathbf{A}')$. It is immediate from the definition of W that $W(\mathrm{Id}_M, 0) = \mathrm{Id}_M$ and that $W(\Phi \circ \Psi) = W(\Phi) \circ W(\Psi)$; so W is a functor.

It remains to show that W is full, faithful, and essentially surjective. First, let $\mathbf{F}$ be any object in $\mathcal{F}$. By Stokes' Theorem (given that $d\mathbf{F} = 0$ is the first of Maxwell's equations), there exists some $\mathbf{A}$ on M such that $\mathbf{F} = d\mathbf{A} = W(\mathbf{A})$. So W is surjective, and hence essentially surjective.

Second, consider any pair of arrows $(\psi, \mathbf{G})$ and $(\chi, \mathbf{K})$ from $\mathbf{A}$ to $\mathbf{A}'$. If $W(\psi, \mathbf{G}) = W(\chi, \mathbf{K})$, then $\psi = \chi$. Since $\mathbf{A}' = \psi^* \mathbf{A} + \mathbf{G} = \chi^* \mathbf{A} + \mathbf{K}$, it follows that $\psi^* \mathbf{A} + \mathbf{G} = \psi^* \mathbf{A} + \mathbf{K}$, and hence that $\mathbf{G} = \mathbf{K}$. So $(\psi, \mathbf{G}) = (\chi, \mathbf{K})$, and hence, W is faithful.

Finally, consider any arrow $\psi : W(\mathbf{A}) \to W(\mathbf{A}')$ in $\mathcal{F}$. By definition, this is an isometry such that

$$\psi^*(d\mathbf{A}) = d\mathbf{A}' \tag{12.17}$$

Consider the 1-form $\mathbf{G} := \mathbf{A}' - \psi^* \mathbf{A}$. $\mathbf{G}$ is closed, i.e. $d\mathbf{G} = 0$:

$$\begin{aligned} d\mathbf{G} &= d(\mathbf{A}' - \psi^* \mathbf{A}) \\ &= d\mathbf{A}' - \psi^*(d\mathbf{A}) \\ &= 0 \end{aligned}$$

where we have used equation (12.17) and the fact that ψ is an isometry. Since M is

contractible, it follows that $\mathbf{G}$ is exact. Since

$$\mathbf{A}' = \psi^* \mathbf{A} + \mathbf{G} \tag{12.18}$$

it follows that $(\psi, \mathbf{G})$ is an arrow in $\mathcal{A}$, and evidently $W(\psi, \mathbf{G}) = \psi$. Thus, W is full. □

This result demonstrates one of the most useful features of categorical equivalence: it can elicit the sense in which two theories can be regarded as equivalent even if an individual model in one theory corresponds to a set of models in the other—provided that the models in that set are equivalent to one another (formally, that they are regarded as isomorphic in the ambient category). However, there are some limitations to the above.

One obvious limitation is that it will not hold if we expand our category to include non-contractible spaces: for, as discussed in Chapter 9, there are pairs of non-gauge-equivalent potentials $\mathbf{A}, \mathbf{A}'$ (over such spaces) that give rise to the same field $\mathbf{F}$. The functor W (or rather, the extension of W to this enlarged category) would therefore map $\mathbf{A}$ and $\mathbf{A}'$ to $\mathbf{F}$; and clearly, $W_{\mathbf{A}\mathbf{A}'}$ would not be surjective, since there are no arrows between $\mathbf{A}$ and $\mathbf{A}'$,[2] but there is (at least) the identity arrow between $\mathbf{F}$ and itself. That said, this seems like a feature rather than a bug: as we have already discussed, the fact that gauge-equivalence is a stricter criterion than field-equivalence seems to indicate that the potentials theory should *not* be considered equivalent to the fields theory over non-contractible spaces, even if gauge-equivalent models are taken to be physically equivalent.

A more troubling limitation is that Proposition 2 seems to depend quite sensitively on the question of how we represent gauge transformations. In the above, we took each gauge transformation to be specified by an exact 1-form $\mathbf{G}$; i.e. by a 1-form such that for some scalar field λ, $\mathbf{G} = d\lambda$. What if we instead take a gauge transformation to be specified by the scalar field λ itself? In particular, recall that when we come to couple the Maxwell theory to the quantum theory of a charged particle, a gauge transformation will be specified by such a scalar field λ—it is just that the action of this gauge transformation on $\mathbf{A}$ is given by $d\lambda$.

Thus, let us define a category $\mathcal{A}'$. In this category, each object is a four-potential $\mathbf{A}$ on Minkowski spacetime (as before); but an arrow from $\mathbf{A}$ to $\mathbf{A}'$ is a pair (ψ, λ), where $\psi : M \to M$ is an isometry and $\lambda : M \to \mathbb{R}$ is a smooth scalar field such that

$$\mathbf{A}' = \psi^* \mathbf{A} + d\lambda \tag{12.19}$$

[2] Assuming, without loss of generality, that there is no isometry ψ such that $\mathbf{A}' = \psi^* \mathbf{A}$.

This is indeed a category, with composition of arrows given by $(\chi, \mu) \circ (\psi, \lambda) = (\chi \circ \psi, \lambda + \mu)$ and the identity arrow on any $\mathbf{A}$ given by $(\mathrm{Id}_M, 0)$. However, it is not equivalent to the category $\mathcal{A}$.

Proposition 13. There is no full functor from $\mathcal{A}$ to $\mathcal{A}'$.

Proof. There exist solutions to Maxwell's equations that lack any non-trivial isometries; let $\mathbf{A}$ be such a solution. Then the only arrow from $\mathbf{A}$ to itself in $\mathcal{A}$ is the identity arrow, i.e. the arrow specified by $(\mathrm{Id}_M, 0)$. Suppose that J is a functor from $\mathcal{A}$ to $\mathcal{A}'$, and consider $J(\mathbf{A})$. At a minimum, any pair of the form (Id_M, λ) for λ a constant scalar field is an arrow from $J(\mathbf{A})$ to itself. But $J(\mathrm{Id}_M, 0) = (\mathrm{Id}_M, 0)$ (by the fact that J is a functor); so if $\lambda \neq 0$, then there is no arrow from $\mathbf{A}$ to itself that J sends to (Id_M, λ). So J is not full. □

This suggests that we may be able to think of the two categories $\mathcal{A}$ and $\mathcal{A}'$ as encoding different interpretations of the theory of electromagnetic potentials. The difference between these two interpretations, though, is rather subtle. The two interpretations agree on what the models of the theory are, and they agree on the relations of physical equivalence between models; what they disagree on are the relations of physical equivalence *between gauge transformations* (in a sense, they disagree on the relations of physical equivalence between relations of physical equivalence). According to $\mathcal{A}'$, two gauge transformations—say, those represented by scalar fields λ and μ—are physically inequivalent if $\lambda \neq \mu$; according to $\mathcal{A}$, if $d\lambda = d\mu$, then we should regard these two gauge transformations as equivalent.

This indicates that we should not expect merely moving to a more abstract, category-theoretic perspective will free us from troublesome questions of interpretation. Instead, that perspective gives us new tools with which to articulate those questions. As with most things in life, the category-theoretic resources giveth, by showing how to make precise a sense in which (for example) the field-theoretic and potential-theoretic formulations of electromagnetism can be seen as equivalent (under the right circumstances); and it taketh away, by showing how to draw even more fine-grained distinctions between different potential-theoretic formulations than we could before.

A. Vector and affine spaces

This appendix reviews basic facts about vector and affine spaces, including the notions of metric and orientation. To a large extent, the treatment follows Malament (2009).

A.1. Matrices

Definition 34. An $m \times n$ *matrix* is a rectangular table of mn real numbers A_{ij}, where $1 \leq i \leq m$ and $1 \leq j \leq n$, arrayed as follows:

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \tag{A.1}$$

♠

The operation of matrix multiplication is defined as follows.

Definition 35. Given an $m \times n$ matrix A and an $n \times p$ matrix B, their *matrix product* AB is the $m \times p$ matrix with entries $(AB)^i{}_k$, where

$$(AB)^i{}_k = A^i{}_j B^j{}_k \tag{A.2}$$

♠

Note that this uses the *Einstein summation convention*: repeated indices are summed over.

Definition 36. Given an $m \times n$ matrix $A^i{}_j$, its *transpose* is an $n \times m$ matrix denoted by $A_i{}^j$, and defined by the condition that for all $1 \leq i \leq m$ and $1 \leq j \leq n$, $A_i{}^j = A^i{}_j$, ♠

An element of $\mathbb{R}^n$ can be identified with an $n \times 1$ matrix; as such, an $m \times n$ matrix A encodes a map $A : \mathbb{R}^m \to \mathbb{R}^n$. Such a map is *linear*, in the sense that for any $x^i, y^i \in \mathbb{R}^m$

and any $a, b \in \mathbb{R}$,

$$A^i{}_j(ax^j + by^j) = a(A^i{}_j x^j) + b(A^i{}_j y^j) \tag{A.3}$$

And, in fact, any linear map from $\mathbb{R}^m$ to $\mathbb{R}^n$ corresponds to some matrix.

It follows that a *square* matrix—one that is an $n \times n$ matrix, for some $n \in \mathbb{N}$—can be identified with a linear map $\mathbb{R}^n \to \mathbb{R}^n$.

Definition 37. Given a square $n \times n$ matrix $A^i{}_j$, the *determinant* of $A^i{}_j$ is

$$\det(A) = \sum_{\sigma \in S_n} \left(\operatorname{sgn}(\sigma) \prod_{i=1}^{n} A^i{}_{\sigma(i)} \right) \tag{A.4}$$

where S_n is the permutation group and $\operatorname{sgn}(\sigma)$ is the sign of the permutation σ (see Appendix B). ♠

Definition 38. Given an $m \times n$ matrix $A^i{}_r$, an $n \times m$ matrix $B^r{}_i$ is its *inverse* if multiplying them together in either order yields an identity matrix: that is, if

$$A^i{}_r B^r{}_j = \delta^i_j \tag{A.5}$$

$$B^r{}_i A^i{}_s = \delta^r_s \tag{A.6}$$

where δ^i_j is the $m \times m$ identity matrix and δ^r_s is the $n \times n$ identity matrix. ♠

Definition 39. A square matrix $A^i{}_j$ is *orthogonal* if its transpose is its inverse: that is, if

$$A^i{}_j A_i{}^k = \delta^k_j \tag{A.7}$$

♠

A.2. Vector spaces

Definition 40. A *(real) vector space* $\mathbb{V}$ consists of a set $|\mathbb{V}|$, equipped with a binary operation $+ : \mathbb{V} \times \mathbb{V} \to \mathbb{V}$ (addition), a unary operation $-$ (additive inversion), an operation $\cdot : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ (scalar multiplication), and a privileged element $0 \in \mathbb{V}$ (the zero vector),

such that the following conditions are obeyed (for any $\vec{u}, \vec{v}, \vec{w} \in \mathbb{V}$ and $a, b \in \mathbb{R}$):

$$\vec{u} + \vec{v} = \vec{v} + \vec{u} \tag{A.8}$$

$$\vec{u} + (\vec{v} + w) = (\vec{u} + \vec{v}) + w \tag{A.9}$$

$$\vec{u} + 0 = \vec{u} \tag{A.10}$$

$$(-\vec{u}) + \vec{u} = 0 \tag{A.11}$$

$$a \cdot (b \cdot \vec{u}) = (ab) \cdot \vec{u} \tag{A.12}$$

$$1 \cdot \vec{u} = \vec{u} \tag{A.13}$$

$$a \cdot (\vec{u} + \vec{v}) = a \cdot \vec{u} + a \cdot \vec{v} \tag{A.14}$$

$$(a + b) \cdot \vec{u} = a \cdot \vec{u} + b \cdot \vec{u} \tag{A.15}$$

♠

We will often write $a\vec{u}$ instead of $a \cdot \vec{u}$.

Given a set of vectors $S \subseteq |\mathbb{V}|$, the vectors in S are *linearly dependent* if there exist $\vec{v}_1, \ldots, \vec{v}_k \in S$ and $a_1, \ldots, a_k \in \mathbb{R}$ such that $a_1\vec{u}_1 + \cdots a_k\vec{u}_k = 0$; otherwise, they are *linearly independent*.

Definition 41. A *basis* for $\mathbb{V}$ is a set B of linearly independent vectors such that for every $\vec{v} \in \mathbb{V}$, there exist $\vec{v}_1, \ldots, \vec{v}_k \in B$ and $a_1, \ldots, a_k \in \mathbb{R}$ such that $a_1\vec{v}_1 + \cdots + a_k\vec{v}_k = \vec{v}$. ♠

From now on, we assume that any vector space $\mathbb{V}$ we consider is *finite-dimensional*: that is, that there exists a finite basis B of $\mathbb{V}$. For such a vector space, there is some natural number n such that every basis of $\mathbb{V}$ contain n elements; we say that n is the *dimension* of $\mathbb{V}$, and denote it by $\dim(V)$.

Definition 42. Let $\mathbb{V}$ be an n-dimensional vector space, and let $\{\vec{e}_1, \ldots, \vec{e}_n\}$ be a basis for $\mathbb{V}$. Given any $\vec{v} \in \mathbb{V}$, the *components* of $\vec{v}$ are the (unique) numbers $v^1, \ldots, v^n$ such that

$$\vec{v} = v^i\vec{e}_i \tag{A.16}$$

♠

It follows that relative to a choice of basis, an n-dimensional space may be identified with $\mathbb{R}^n$, and so with $n \times 1$ matrices.

Definition 43. Given a vector space $\mathbb{V}$, a *subspace* of $\mathbb{V}$ is a vector space $\mathbb{W}$ such that $|\mathbb{W}| \subseteq |\mathbb{V}|$ and the vector-space structure on $\mathbb{W}$ is the restriction of the vector-space on $\mathbb{V}$ to $\mathbb{W}$. ♠

Definition 44. Let $\mathbb{W}$ be a subspace of $\mathbb{V}$. Two vectors $\vec{v}_1, \vec{v}_2 \in \mathbb{V}$ are *equivalent modulo* $\mathbb{W}$ if $(\vec{v}_2 - \vec{v}_1) \in \mathbb{W}$: that is, if there is some $\vec{w} \in \mathbb{W}$ such that $\vec{v}_2 = \vec{v}_1 + \vec{w}$. Let the equivalence class of $\vec{v}$ be denoted $[\vec{v}]$. ♠

Definition 45. Let $\mathbb{W}$ be a subspace of $\mathbb{V}$. The *quotient* of $\mathbb{V}$ by $\mathbb{W}$ is a vector space $\mathbb{V}/\mathbb{W}$, defined as follows. The underlying set is the partition of $\mathbb{V}$ by equivalence modulo $\mathbb{W}$: i.e. the elements of $\mathbb{V}/\mathbb{W}$ are equivalence classes $[v]$. Addition and scalar multiplication are defined as follows:

$$[\vec{v}_1] + [\vec{v}_2] = [\vec{v}_1 + \vec{v}_2] \tag{A.17}$$

$$a[\vec{v}] = [a\vec{v}] \tag{A.18}$$

It is straightforward to verify that these definitions do not depend on the choice of representative. ♠

Definition 46. Let V and W be two vector spaces. The *direct sum* of V and W, denoted $V \oplus W$, is the vector space defined as follows: its underlying set is $V \times W$, and addition and scalar multiplication are defined pointwise:

$$(\vec{v}, \vec{w}) + (\vec{v}', \vec{w}') = (\vec{v} + \vec{v}', \vec{w} + \vec{w}') \tag{A.19}$$

$$a(\vec{v}, \vec{w}) = (a\vec{v}, a\vec{w}) \tag{A.20}$$

It is straightforward to show that $V \oplus W$ is a vector space (and that if V and W are finite-dimensional, that $\dim(V \oplus W) = \dim(V) + \dim(W)$). The element $(\vec{v}, \vec{w})$ of $V \oplus W$ will be denoted by $\vec{v} \oplus \vec{w}$. ♠

Proposition 14. For any vector spaces $\mathbb{V}$ and $\mathbb{W}$, $\mathbb{V}$ and $\mathbb{W}$ are both subspaces of $\mathbb{V} \oplus \mathbb{W}$.

Definition 47. Let $\mathbb{V}$ and $\mathbb{W}$ be vector spaces. A *linear map* is a map $f : \mathbb{V} \to \mathbb{W}$ such that for any $\vec{u}, \vec{v} \in \mathbb{V}$ and any $x \in \mathbb{R}$,

$$f(\vec{u} + \vec{v}) = f(\vec{u}) + f(\vec{v}) \tag{A.21}$$

$$f(x \cdot \vec{u}) = x \cdot f(\vec{u}) \tag{A.22}$$

♠

Given bases on $\mathbb{V}$ and $\mathbb{W}$, and hence an identification of $\mathbb{V}$ with $\mathbb{R}^m$ and $\mathbb{W}$ with $\mathbb{R}^n$ (where $m = \dim(\mathbb{V})$ and $n = \dim(\mathbb{W})$), a linear map $f : \mathbb{V} \to \mathbb{W}$ may be identified with an $n \times m$ matrix $F^i{}_j$.

Definition 48. A linear map $f : \mathbb{V} \to \mathbb{W}$ is a *linear isomorphism* if it is invertible. ♠

Definition 49. Given an invertible linear map $f : \mathbb{V} \to \mathbb{V}$, the *determinant* of f is the determinant of the matrix $F^i{}_j$ that represents f relative to any basis B of $\mathbb{V}$. ♠

It can be shown that the determinant is independent of the choice of basis, so this definition is well-formed.

Proposition 15. Given a basis B of $\mathbb{V}$, if linear maps $f, g : \mathbb{V} \to \mathbb{W}$ agree on B (i.e. if $f(\vec{v}) = g(\vec{v})$ for all $\vec{v} \in B$), then $f = g$.

Proposition 16. Given two ordered bases $B = \langle \vec{e}_1, \ldots, \vec{e}_n \rangle$ and $B' = \langle \vec{e}'_1, \ldots, \vec{e}'_n \rangle$ of $\mathbb{V}$, there is a unique linear map $f : \mathbb{V} \to \mathbb{V}$ such that $\vec{e}'_i = f(\vec{e}_i)$ for all $1 \leq i \leq n$.

Definition 50. Given a vector space $\mathbb{V}$, an *inner product* on $\mathbb{V}$ is a non-degenerate, bilinear, symmetric map $\langle \, , \, \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$: that is, a map such that

$$\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle \tag{A.23}$$

$$\langle \vec{u}, a\vec{v} \rangle = a \langle \vec{u}, \vec{v} \rangle \tag{A.24}$$

$$\langle \vec{u}, \vec{v} + w \rangle = \langle \vec{u}, \vec{v} \rangle + \langle \vec{u}, w \rangle \tag{A.25}$$

$$\text{If } \vec{u} \neq 0, \text{ then for some } \vec{v} \in \mathbb{V}, \langle \vec{u}, \vec{v} \rangle \neq 0 \tag{A.26}$$

♠

A vector space equipped with an inner product will be referred to as an *inner product space*.

Definition 51. Given an inner product space $\mathbb{V}$, two vectors $\vec{u}, \vec{v} \in \mathbb{V}$ are *orthogonal* if $\langle \vec{u}, \vec{v} \rangle = 0$. ♠

Definition 52. Given an inner product space $\mathbb{V}$, a basis B of $\mathbb{V}$ is *orthonormal* if for all $\vec{u}, \vec{v} \in B$,

$$\langle \vec{u}, \vec{v} \rangle^2 = \begin{cases} 0 & \text{if } \vec{u} \neq \vec{v} \\ 1 & \text{if } \vec{u} = \vec{v} \end{cases} \tag{A.27}$$

♠

Given an orthonormal basis B of $\mathbb{V}$, the *signature* of $\mathbb{V}$ is the pair (n^+, n^-), where $n^+, n^- \in \mathbb{N}$, such that there are n^+ elements $u \in B$ such that $\langle u, u \rangle = 1$, and n^- elements $u \in B$ such that $\langle u, u \rangle = -1$. Evidently, $n^+ + n^- = \dim(\mathbb{V})$; moreover, one can show that the signature of $\mathbb{V}$ is independent of what orthonormal basis is chosen.

Definition 53. An inner product on a vector space $\mathbb{V}$ is *positive definite* if $\langle \vec{v}, \vec{v} \rangle \geq 0$ for all $\vec{v} \in \mathbb{V}$; equivalently, if its signature is $(\dim(\mathbb{V}), 0)$. ♠

Definition 54. Given an inner product space $\mathbb{V}$, a linear automorphism $f : \mathbb{V} \to \mathbb{V}$ is an *orthogonal* map if it preserves the inner product: that is, if

$$\langle f(\vec{u}), f(\vec{v}) \rangle = \langle \vec{u}, \vec{v} \rangle \tag{A.28}$$

♠

Definition 55. Let $\mathbb{V}$ be a vector space, and let B and B' be two ordered bases of $\mathbb{V}$. B and B' are *co-oriented* if the linear automorphism of $\mathbb{V}$ taking B into B' (see Proposition 16) has positive determinant. ♠

Proposition 17. Co-orientation is an equivalence relation on the set of ordered bases of $\mathbb{V}$, with exactly two equivalence classes (if $\mathbb{V}$ is non-empty).

Definition 56. An *orientation* on $\mathbb{V}$ is a choice of equivalence class of co-oriented ordered bases on $\mathbb{V}$ as the set of *right-handed* ordered bases; the other equivalence class is referred to as the set of *left-handed* ordered bases. A vector space equipped with an orientation is said to be an *oriented* vector space. ♠

A.3. Affine spaces

Since a vector space is a group, we can form the principal homogeneous space of a vector space. Such a space is known as an *affine space*.

Definition 57. Let $\mathbb{V}$ be a vector space. An *affine space* $\mathcal{V}$ is a set $|\mathcal{V}|$ equipped with a free and transitive action $a \mapsto a + \vec{v}$ of $\mathbb{V}$: that is, for any $a, b \in \mathcal{V}$ there is a unique $\vec{v} \in \mathbb{V}$ such that

$$b = a + \vec{v} \tag{A.29}$$

We will use $(b - a)$ to denote this unique vector. ♠

Proposition 18. If $\mathbb{W}$ is a proper subspace of $\mathbb{V}$, then the action of $\mathbb{W}$ on $\mathcal{V}$ is free but not transitive.

Proposition 19. If $\mathbb{W}$ is a subspace of $\mathbb{V}$, then the quotient $\mathcal{U} = {}^{\mathcal{V}}\!/_{\mathbb{W}}$ is an affine space with associated vector space $\mathbb{U} = {}^{\mathbb{V}}\!/_{\mathbb{W}}$.

Proof. We define the action of $\mathbb{U}$ on $\mathcal{U}$ as follows. Let $\vec{u} \in \mathbb{U}$ and $x \in \mathcal{U}$: thus $y = [x]$, for some $y \in \mathcal{V}$, and $\vec{u} = [\vec{v}]$, for some $\vec{v} \in \mathbb{V}$. Then define

$$x + \vec{u} := [y + \vec{v}] \tag{A.30}$$

First, we need to check that this is well-defined, i.e. that it is independent of the choice of y and $\vec{v}$. So let y' and $\vec{v}'$ be such that $y' = y + \vec{w}$ and $\vec{v}' = \vec{v} + \vec{w}'$, for $\vec{w}, \vec{w}' \in \mathbb{W}$. Then

$$y' + \vec{v}' = y + \vec{v} + (\vec{w} + \vec{w}') \tag{A.31}$$

and so (since $(\vec{w} + \vec{w}') \in \mathbb{W}$) $[y' + \vec{v}'] = [y + \vec{v}]$, and so our definition is indeed well-defined.

Now suppose that $x_1, x_2 \in \mathcal{U}$, with $x_1 = [y_1]$ and $x_2 = [y_2]$ for $y_1, y_2 \in \mathcal{V}$. Since the action of $\mathbb{V}$ on $\mathcal{V}$ is free and transitive, there is a unique $\vec{v} \in \mathbb{V}$ such that $y_2 = y_1 + \vec{v}$. Let $\vec{u} = [\vec{v}]$. Then:

$$\begin{aligned} x_1 + \vec{u} &= [y_1 + \vec{v}] \\ &= x_2 \end{aligned}$$

So the action of $\mathbb{U}$ on $\mathcal{U}$ is transitive. Furthermore, if $x_2 = x_1 + \vec{u}'$, i.e., $[y_2] = [y_1 + \vec{v}']$ (for some $\vec{v}' \in \mathbb{V}$ such that $[\vec{v}'] = \vec{u}'$), then $y_2 = y_1 + \vec{v}' + \vec{w}$ for some $\vec{w} \in \mathbb{W}$; so $\vec{v} = \vec{v}' + \vec{w}$, and hence $\vec{u}' = [\vec{v}'] = [\vec{v}] = \vec{u}$. So the action of $\mathbb{U}$ on $\mathcal{U}$ is free. □

Definition 58. Let $\mathcal{V}$ and $\mathcal{W}$ be affine spaces, with (respective) underlying vector spaces $\mathbb{V}$ and $\mathbb{W}$. The *product affine space* $\mathcal{V} \times \mathcal{W}$ is the affine space whose underlying set is $|\mathcal{V}| \times |\mathcal{W}|$ and whose associated vector space is $\mathbb{V} \oplus \mathbb{W}$, where the action of $\mathbb{V} \oplus \mathbb{W}$ on $|\mathcal{V}| \times |\mathcal{W}|$ is given by

$$(\vec{v} + \vec{w})(x, y) = (x + \vec{v}, y + \vec{w}) \tag{A.32}$$

♠

Structures on an affine space's associated vector space can be 'transferred' to the affine space, as the following two definitions indicate.

Definition 59. A *metric affine space* is an affine space $\mathcal{V}$ whose associated vector space $\mathbb{V}$ is an inner product space. ♠

A metric affine space carries a notion of distance: given any two points $x, y \in \mathcal{V}$, the distance between them is $|y - x|$.

Definition 60. An *oriented affine space* is an affine space $\mathcal{V}$ whose associated vector space $\mathbb{V}$ has an orientation. ♠

A.4. Vector calculus on Euclidean space

Throughout this section, let $\mathcal{X}$ be an oriented Euclidean space: that is, a three-dimensional affine space whose associated vector space $\mathbb{X}$ is equipped with a positive-definite metric and an orientation.

Definition 61. A *vector field* is a smooth map $\vec{V} : \mathcal{X} \to \mathbb{X}$. ♠

Definition 62. Given a scalar field $\phi : \mathcal{X} \to \mathbb{R}$ and a vector $\vec{v} \in \mathbb{X}$, the *directional derivative* (of λ along $\vec{v}$) is a scalar field whose value at any point $x \in \mathcal{X}$ is given by

$$\nabla_{\vec{v}}\phi = \lim_{\varepsilon \to 0} \frac{\phi(x + \varepsilon\vec{v}) - \phi(x)}{\varepsilon} \tag{A.33}$$

♠

As a notational special case, suppose that we have introduced a right-handed, orthonormal basis $\langle \vec{e}_1, \vec{e}_2, \vec{e}_3 \rangle$ on $\mathbb{X}$. Then the directional derivative $\nabla_{\vec{e}_i}\phi$ will be denoted by $\nabla_i\phi$. We use this basis to define the operators grad and div; however, these definitions will pick out the same operators if we use any other right-handed, orthonormal basis.

Given a vector field $\vec{V} : \mathcal{X} \to \mathbb{X}$, the *components* of $\vec{V}$ (relative to this basis) are the three scalar fields V^1, V^2, V^3 such that $\vec{V} = V^i\vec{e}_i\rangle$.

Definition 63. Given a scalar field $\phi : \mathcal{V} \to \mathbb{R}$, the *gradient* of ϕ is a vector field $\mathrm{grad}(\phi)$ whose components are

$$\mathrm{grad}(\phi)^i = \nabla_i\phi \tag{A.34}$$

♠

Geometrically, the gradient is a vector field whose direction is the direction in which ϕ is most strongly changing, and whose magnitude is the rate at which ϕ is changing along that direction.

Definition 64. Given a vector field $\vec{V} : \mathcal{X} \to \mathbb{X}$, the *divergence* of $\vec{V}$ is a scalar field $\mathrm{div}(\vec{V})$ given by

$$\mathrm{div}(\vec{V}) = \sum_i \nabla_i V^i \tag{A.35}$$

♠

Gometrically, the divergence of $\vec{V}$ at a point $x \in \mathcal{X}$ expresses the extent to which x is a source or a sink for $\vec{V}$: if the 'outflow' of $\vec{V}$ around x exceeds the 'inflow', then $\text{div}(\vec{V})$ is positive; if the inflow exceeds the outflow, it is negative; and if inflow is equal to outflow, then it is zero.

Definition 65. Given a vector field $\vec{V} : \mathcal{X} \to \mathbb{X}$, the *curl* of $\vec{V}$ is a vector field $\text{curl}(\vec{V})$ whose components are

$$(\text{curl}(\vec{V}))^1 = \nabla_2 V^3 - \nabla_3 V^2 \tag{A.36}$$

$$(\text{curl}(\vec{V}))^2 = \nabla_3 V^1 - \nabla_1 V^3 \tag{A.37}$$

$$(\text{curl}(\vec{V}))^3 = \nabla_1 V^2 - \nabla_2 V^1 \tag{A.38}$$

♠

Geometrically, the curl of $\vec{V}$ at a point $x \in \mathcal{X}$ expresses the 'rotation' of $\vec{V}$ at x: the direction of $\text{curl}(\vec{V})$ is the axis of rotation (using the right-hand-rule), and its magnitude expresses the amount of rotation.

B. Group theory

B.1. Groups

Definition 66. A *group* consists of a set G, equipped with a binary operation $*$ (of *group multiplication*), a unary operation $^{-1}$ (of *inversion*), and a privileged element e (the *identity*), such that for any $g, h, k \in G$:

$$g * (h * k) = (g * h) * k \tag{B.1}$$

$$g * e = g = e * g \tag{B.2}$$

$$g^{-1} * g = g = g * g^{-1} \tag{B.3}$$

♠

We will frequently abbreviate $g * h$ by gh.

Example 14. Any vector space is a group, with addition as the group operation (and the zero vector as the identity, and additive inverse as group inverse).

Example 15. The real numbers are a group with respect to addition (with 0 as the identity and $-x$ as the inverse of x) and with respect to multiplication (with 1 as the identity and $1/x$ as the inverse of x).

Example 16. Given a set A, a *permutation* of A is a bijection $f : A \to A$. The *symmmetric group* of A is the group $\mathrm{Sym}(A)$ consisting of all permutations of A, with composition as the group operation.

Example 17. If a finite set A has n elements, then $\mathrm{Sym}(A)$ is denoted by S_n. A *transposition* is a permutation $\tau \in S_n$ that just exchanges two elements: i.e. is such that for some $a, b \in A$ where $a \neq b$, $\tau(a) = b$ and $\tau(b) = a$, and for all other $c \in A$, $\tau(c) = c$.

Any $\sigma \in S_n$ can be expressed as a finite composition of transpositions. It can be shown that if σ is expressible as an even number of transpositions, then it is *only* expressible as an even number of transpositions; and similarly in case σ is expressible as an odd number of transpositions. Accordingly, σ is given a *sign* $\mathrm{sgn}(\sigma)$: if it is an even number then $\mathrm{sgn}(\sigma) = +1$, and if it is an odd number then $\mathrm{sgn}(\sigma) = -1$.

Definition 67. Given groups G and H, a *(group) homomorphism* is a map $\phi : G \to H$ such that for any $g, g' \in G$,

$$\phi(g * g') = \phi(g) * \phi(g') \tag{B.4}$$

♠

Definition 68. Given a group G, a subset $H \subseteq G$ is a *subgroup* of G if H is closed under group multiplication and inversion: that is, if for all $g, h \in H$, $g * h \in H$ and $g^{-1} \in H$. ♠

We can also state this as follows: a subset H of G is a subgroup if $e \in H$ and H is a group under the restriction of the operations $*$ and $^{-1}$ to H.

B.2. Group actions

Definition 69. Given a group G and set X, an *action* of G on X assigns every $g \in G$ to some bijection $g\bullet : X \to X$, such that for any $g, h \in G$ and $x \in X$,

$$(gh)x = g(hx) \tag{B.5}$$

♠

Definition 70. Given an action of G on X, two points $x, y \in X$ are *G-related* if the one can be mapped to the other by G: that is, if there is some $g \in G$ such that $y = gx$. (The proof that this is an equivalence relation is left as an exercise.) A *G-orbit* in X is an equivalence class of G-equivalent points of X. ♠

Definition 71. Given an action of a group G on some set Ω, the *quotient* of Ω by G is the set Ω/G consisting of G-orbits in Ω. ♠

Definition 72. An action of G on X is *transitive* if for any $x, y \in X$, there is some $g \in G$ such that $y = gx$. ♠

In other words, a transitive group action is one such that any two elements of X are G-related; hence, a transitive group action is one for which there only exists a single orbit.

Definition 73. An action of G on X is *free* if for any $x \in X$ and any $g, h \in G$: if $gx = hx$ then $g = h$ (or, equivalently: if $g \neq h$ then $gx \neq hx$). ♠

Thus, a free action is one where distinct elements of G have distinct effects on *every* element of X.

Definition 74. Let X and Y be G-sets. A map $f : X \to Y$ is *G-equivariant* if for any $g \in G$ and $x \in X$,

$$f(gx) = g(f(x)) \tag{B.6}$$

♠

Actions which are both free and transitive are said to be *regular*, and have the following important feature:

Proposition 20. Suppose that X and Y are G-sets where the action of G is free and transitive. Then there exist G-equivariant bijections $f : X \to Y$ and $f^{-1} : Y \to X$.

Proof. Pick any points $x_0 \in X$ and $y_0 \in Y$, and let $f(x_0) = y_0$. For any other $x \in X$, we know that $x = g_x x_0$ for some unique element g_x of G; now set $f(x) = g_x y_0$. This suffices to determine f; we now show that f is a G-equivariant bijection. First, for any $x \in X$ and any $g \in G$,

$$f(gx) = f(gg_x x_0) = gg_x y_0 = g(f(x))$$

So f is G-equivariant.

Next, consider any $x_1 = g_1 x_0$ and $x_2 = g_2 x_0$ in X. If $f(x_1) = f(x_2)$, i.e. $f(g_1 x_0) = f(g_2 x_0)$, then $g_1 y_0 = g_2 y_0$. But since G acts freely on Y, it follows that $g_1 = g_2$ and so $x_1 = x_2$. So f is injective.

Finally, consider any $y \in Y$, and (again using the fact that G's action on Y is regular), express it in the form $g_y y_0$. Then

$$f(g_y x_0) = g_y(f(x_0)) = g_y y_0 = y$$

So f is surjective, and hence a bijection.

Showing that f^{-1} is G-equivariant is left as an exercise. □

A G-set for which the action of G is regular is said to be a *principal homogeneous space for G*, or alternatively a *G-torsor*. Taking bijective G-equivariant maps as the appropriate notion of isomorphism for G-sets, we see that G has, up to isomorphism, a unique principal homogeneous space.[1]

[1] If that's the case, why don't we speak instead of *the* principal homogeneous space for G, just as we speak of *the* real numbers as the unique (up to isomorphism) complete ordered field? That's a good question; one reason not to do so is that there will typically be multiple isomorphisms between two principal homogeneous spaces for G, so there is no canonical way to identify one principal homogeneous space with another (whereas there is a *unique* isomorphism between two complete ordered fields).

Example 18. Given a vector space $\mathbb{V}$, the principal homogeneous space for $\mathbb{V}$ (regarded as a group) is the affine space $\mathcal{V}$.

C. Differential forms

C.1. Multi-covectors

Definition 75. Let $\mathbb{V}$ be a vector space. A *covector* (over $\mathbb{V}$) is a linear map $\mathbf{p} : \mathbb{V} \to \mathbb{R}$. We refer to the set of covectors over $\mathbb{V}$ as the *dual vector space*, and denote it by $\mathbb{V}^*$. ♠

It is not hard to show that $\mathbb{V}^*$ is also a vector space (by defining addition and scalar multiplication pointwise), of the same dimension as $\mathbb{V}$. The dual vector space to a direct sum of vector spaces is the direct sum of the duals: that is, $(\mathbb{V} \oplus \mathbb{W})^* = \mathbb{V}^* \oplus \mathbb{W}^*$.

Definition 76. Let $\mathbb{V}$ be a vector space. For any $k \in \mathbb{N}$, a *k-covector* is an alternating multilinear map $\mathbf{q} : \mathbb{V}^k \to \mathbb{R}$; that is, a map which is linear in each argument, and which has the property that swapping any two arguments changes the sign of the result. ♠

Thus, a 1-covector is a covector; a 2-covector is an antisymmetric bilinear map $f : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$; and so on. We consider real numbers to be 0-covectors. The set of k-covectors over a vector space $\mathbb{V}$ will be denoted $\Lambda^k(\mathbb{V}^*)$.

If the arguments fed to a k-covector are linearly dependent, then the result will vanish: for example, given a 2-covector $\mathbf{p}$, if we feed it $\vec{u}$ and $a\vec{u}$ (where $a \in \mathbb{R}$),

$$\mathbf{p}(\vec{u}, a\vec{u}) = a\mathbf{p}(\vec{u}, \vec{u}) = 0 \tag{C.1}$$

If $\mathbb{V}$ is n-dimensional, then there can be at most n linearly independent vectors, and so any k-covector for $k > n$ will be trivial; for this reason, we typically treat the n-covectors as the end of the line.

We can form new multi-covectors out of old ones by using the *exterior product*:

Definition 77. Given a k-covector f and an l-covector g, their *exterior product* $f \wedge g$ is the $(k+l)$-covector whose result, for any $\vec{v}_1, \ldots, \vec{v}_{k+l} \in \mathbb{V}$, is given by

$$(f \wedge g)(\vec{v}_1, \ldots, \vec{v}_{k+l}) = \frac{1}{k!l!} \sum_{\sigma \in S_{k+l}} \operatorname{sgn}(\sigma) f(\vec{v}_{\sigma(1)}, \ldots, f_{\sigma(k)})\, g(\vec{v}_{\sigma(k+1)}, \ldots, \vec{v}_{\sigma(k+l)}) \tag{C.2}$$

where S_{k+l} is the permutation group for $k+l$ elements, and $\text{sgn}(\sigma)$ is the sign of the permutation σ (see Appendix B). ♠

For example, the exterior product of two covectors $\mathbf{p}$ and $\mathbf{q}$ is a 2-covector $\mathbf{p} \wedge \mathbf{q}$, defined by the condition that for any $\vec{u}, \vec{v} \in \mathbb{V}$,

$$(\mathbf{p} \wedge \mathbf{q})(\vec{u}, \vec{v}) := \mathbf{p}(\vec{u})\mathbf{q}(\vec{v}) - \mathbf{p}(\vec{v})\mathbf{q}(\vec{u}) \tag{C.3}$$

Similarly, the exterior product of a covector $\mathbf{p}$ with a 2-covector $\mathbf{r}$ is a 3-covector $\mathbf{p} \wedge \mathbf{r}$, such that for any $\vec{u}, \vec{v}, \vec{w} \in \mathbb{V}$,

$$(\mathbf{p} \wedge \mathbf{r})(\vec{u}, \vec{v}, \vec{w}) = \mathbf{p}(\vec{u})\mathbf{r}(\vec{v}, \vec{w}) + \mathbf{p}(\vec{v})\mathbf{r}(\vec{w}, \vec{u}) + \mathbf{p}(\vec{w})\mathbf{r}(\vec{u}, \vec{v}) \tag{C.4}$$

C.2. Euclidean multi-covectors

Let $\mathbb{X}$ be oriented Euclidean vector space, i.e. a three-dimensional vector space equipped with a positive-definite inner product and an orientation. The inner product induces a very useful isomorphism between vectors and covectors (i.e. between $\mathbb{X}$ and $\mathbb{X}^*$), known as the *musical isomorphism*. On the one hand, given any vector $\vec{v} \in \mathbb{X}$, its associated covector is the linear map $\vec{v}^\flat : \mathbb{X} \to \mathbb{R}$ such that for any $w \in \mathbb{X}$,

$$\vec{v}^\flat(\vec{w}) = \langle \vec{v}, \vec{w} \rangle \tag{C.5}$$

In the other direction, given a covector $\mathbf{p}$, its associated vector $\mathbf{p}^\sharp$ is defined as the vector such that for any vector $v \in \mathbb{X}$,

$$\langle \mathbf{p}^\sharp, \vec{v} \rangle = \mathbf{p}(\vec{v}) \tag{C.6}$$

In the interests of space, we skip over proving that this condition does pick out a unique vector.

Moreover, since $\mathbb{X}$ carries both an inner product and an orientation, it exhibits *Hodge duality*. That is, for any $1 \leq k \leq 3$, there is an isomorphism between $\Lambda^k(\mathbb{X}^*)$ and $\Lambda^{3-k}(\mathbb{X}^*)$: i.e., between the scalars and the 3-covectors, and between the covectors and the 2-covectors. These isomorphisms are given by the *Hodge star* operator, which we define as follows. Let $\langle \vec{e}_1, \vec{e}_2, \vec{e}_3 \rangle$ be an (arbitrarily chosen) right-handed and orthonormal basis of $\mathbb{X}$; and let $\mathbf{e}^i := (\vec{e}_i)^\flat$ (resulting in a basis $\langle \mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3 \rangle$ of $\Lambda^1(\mathbb{X})$). Then the isomorphism $\star : \mathbb{R} \to \Lambda^3(\mathbb{X})$ is defined by

$$\star 1 = \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3 \tag{C.7}$$

and the isomorphism $\star : \Lambda^3(\mathbb{X}) \to \mathbb{R}$ by

$$\star\left(\mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3\right) = 1 \tag{C.8}$$

The isomorphism $\star : \Lambda^1(\mathbb{X}) \to \Lambda^2(\mathbb{X})$ is defined by

$$\star\mathbf{e}^1 = \mathbf{e}^2 \wedge \mathbf{e}^3 \tag{C.9}$$
$$\star\mathbf{e}^2 = \mathbf{e}^3 \wedge \mathbf{e}^1 \tag{C.10}$$
$$\star\mathbf{e}^3 = \mathbf{e}^1 \wedge \mathbf{e}^2 \tag{C.11}$$

and, finally, the isomorphism $\star : \Lambda^2(\mathbb{X}) \to \Lambda^1(\mathbb{X})$ by

$$\star(\mathbf{e}^1 \wedge \mathbf{e}^2) = \mathbf{e}^3 \tag{C.12}$$
$$\star(\mathbf{e}^2 \wedge \mathbf{e}^3) = \mathbf{e}^1 \tag{C.13}$$
$$\star(\mathbf{e}^3 \wedge \mathbf{e}^1) = \mathbf{e}^2 \tag{C.14}$$

Since the Hodge star is required to be linear, these conditions fix its action uniquely. Moreover, it can be shown that the Hodge star (so defined) is independent of which right-handed orthonormal basis of $\mathbb{X}$ is chosen.

We can use the Hodge star and wedge product to express the cross product in a more geometrical fashion: given a pair of vectors $\vec{v}, \vec{w} \in \mathbb{X}$,

$$\vec{v} \times \vec{w} = (\star(\vec{v}^\flat \wedge \vec{w}^\flat))^\sharp \tag{C.15}$$

In other words, we take our vectors, flatten them to a pair of covectors, take their wedge product (a 2-covector), apply the Hodge star to that 2-covector to get a covector back again, and then sharpen that to make a vector. Easy![1]

C.3. Minkowski multi-covectors

Let $\mathbb{M}$ be an oriented Minkowski vector space, i.e. a four-dimensional vector space equipped with a Lorentzian inner product and an orientation. Again, the inner product means that we can establish a musical isomorphism between $\mathbb{M}$ and $\mathbb{M}^*$, again by the

[1] In fact, one can simplify this a bit by defining a wedge product directly on $\mathbb{X}$—thereby constructing an exterior algebra of *multivectors*—and then introducing a Hodge duality between vectors and 2-vectors. With that duality, we can write this expression as $\vec{v} \times \vec{w} = \star(\vec{v} \wedge \vec{w})$. However, such a construction still requires both a metric and an orientation, since those are required to (uniquely) define the Hodge star operator.

conditions that for any $\vec{\xi}, \vec{\eta} \in \mathbb{M}$ and $\mathbf{p} \in \mathbb{M}^*$,

$$\vec{\xi}^\flat(\vec{\eta}) = \langle \vec{\xi}, \vec{\eta} \rangle \tag{C.16}$$

$$\langle \mathbf{p}^\sharp, \vec{\xi} \rangle = \mathbf{p}(\vec{\xi}) \tag{C.17}$$

And again, since $\mathbb{M}$ carries both an inner product and an orientation, it exhibits Hodge duality. In this case, Hodge duality holds between $\Lambda^k(\mathbb{M}^*)$ and $\Lambda^{4-k}(\mathbb{M}^*)$, for each $0 \leq k \leq 4$: that is, between scalars and 4-covectors, covectors and 3-covectors, and between 2-covectors and 2-covectors. Again, take an arbitrary right-handed orthonormal basis $\langle \vec{e}_0, \vec{e}_1, \vec{e}_2, \vec{e}_3 \rangle$, set $\mathbf{e}^\mu = (\vec{e}_\mu)^\flat$ to obtain the dual basis $\langle \mathbf{e}^0, \mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3 \rangle$, and define:

$$\star 1 = \mathbf{e}^0 \wedge \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3 \tag{C.18}$$

$$\star \mathbf{e}^0 = \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3 \tag{C.19}$$

$$\star \mathbf{e}^1 = \mathbf{e}^0 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3 \tag{C.20}$$

$$\star \mathbf{e}^2 = \mathbf{e}^0 \wedge \mathbf{e}^3 \wedge \mathbf{e}^1 \tag{C.21}$$

$$\star \mathbf{e}^3 = \mathbf{e}^0 \wedge \mathbf{e}^1 \wedge \mathbf{e}^2 \tag{C.22}$$

$$\star(\mathbf{e}^0 \wedge \mathbf{e}^1) = \mathbf{e}^3 \wedge \mathbf{e}^2 \tag{C.23}$$

$$\star(\mathbf{e}^0 \wedge \mathbf{e}^2) = \mathbf{e}^1 \wedge \mathbf{e}^3 \tag{C.24}$$

$$\star(\mathbf{e}^0 \wedge \mathbf{e}^3) = \mathbf{e}^2 \wedge \mathbf{e}^1 \tag{C.25}$$

$$\star(\mathbf{e}^1 \wedge \mathbf{e}^2) = \mathbf{e}^0 \wedge \mathbf{e}^3 \tag{C.26}$$

$$\star(\mathbf{e}^1 \wedge \mathbf{e}^3) = \mathbf{e}^2 \wedge \mathbf{e}^0 \tag{C.27}$$

$$\star(\mathbf{e}^2 \wedge \mathbf{e}^3) = \mathbf{e}^0 \wedge \mathbf{e}^1 \tag{C.28}$$

$$\star(\mathbf{e}^0 \wedge \mathbf{e}^1 \wedge \mathbf{e}^2) = -\mathbf{e}^3 \tag{C.29}$$

$$\star(\mathbf{e}^0 \wedge \mathbf{e}^3 \wedge \mathbf{e}^1) = -\mathbf{e}^2 \tag{C.30}$$

$$\star(\mathbf{e}^0 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3) = -\mathbf{e}^1 \tag{C.31}$$

$$\star(\mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3) = -\mathbf{e}^0 \tag{C.32}$$

$$\star(\mathbf{e}^0 \wedge \mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3) = -1 \tag{C.33}$$

C.4. Differential forms

Finally, we introduce differential forms: just as a vector field is a vector-valued field, so a differential form is a multicovector-valued field.

Definition 78. Let $\mathcal{V}$ be an affine space with vector space $\mathbb{V}$. A *k-form on* $\mathcal{V}$ is a smooth map $\mathbf{p} : \mathcal{V} \to \Lambda^k(\mathbb{V}^*)$. ♠

Addition, scalar multiplication, and exterior multiplication of differential forms are defined pointwise. As with multi-covectors, the only non-trivial k-forms on an n-dimensional space are those for $k \leq n$. The set of k-forms on $\mathcal{V}$ is denoted by $\Omega^k(\mathcal{V})$, and the set of all differential forms on $\mathcal{V}$ by $\Omega(\mathcal{V})$.

For oriented Euclidean space and oriented Minkowski spacetime, there is a musical isomorphism between the set of 1-forms and the set of vector fields, and Hodge dualities between the appropriate sets of k-forms; again, these are defined pointwise. Thus, on Euclidean space Hodge duality relates 1-forms to 2-forms, and 3-forms to scalar fields; while on Minkowski spacetime Hodge duality relates 1-forms to 3-forms, 2-forms to 2-forms, and 4-forms to scalar fields.

However, in addition to this, differential forms also exhibit a very natural kind of differential calculus.[2] First, given any scalar field f, we define the *differential* of f to be the 1-form df such that for any vector $\vec{v} \in \mathbb{V}$,

$$df(\vec{v}) = \nabla_{\vec{v}} f \tag{C.34}$$

(where $\nabla_{\vec{v}}$ is the directional derivative with respect to $\vec{v}$; see Appendix A). The extension of this concept to arbitrary differential forms is known as the *exterior derivative*.

Definition 79. Let $\mathcal{V}$ be an n-dimensional affine space. The *exterior derivative* is the unique map $d : \Omega(\mathcal{V}) \to \Omega(\mathcal{V})$ such that for any $k < n$, $d : \Omega^k(\mathcal{V}) \to \Omega^{k+1}(\mathcal{V})$, and which has the following properties:

- for any scalar field (0-form) f,

$$df(V) = \nabla_V f \tag{C.35}$$

- for any k-form $\mathbf{p}$,

$$d(d\mathbf{p}) = 0 \tag{C.36}$$

[2] Although the perspicacious reader might have suspected this would be true, given the name.

- for any $x, y \in \mathbb{R}$ and $\mathbf{p}, \mathbf{q} \in \Omega^k(\mathcal{V})$,

$$d(x\mathbf{p} + y\mathbf{q}) = xd\mathbf{p} + yd\mathbf{q} \tag{C.37}$$

- and for any $\mathbf{p} \in \Omega^k(\mathcal{V})$, $\mathbf{q} \in \Omega(\mathcal{V})$,

$$d(\mathbf{p} \wedge \mathbf{q}) = d\mathbf{p} \wedge \mathbf{q} + (-1)^k \mathbf{p} \wedge d\mathbf{q} \tag{C.38}$$

♠

In fancy lingo, the exterior derivative is a linear and idempotent antiderivation on the exterior algebra of differential forms, which extends the differential on scalar fields. It is non-trivial to show that there exists an operator with these properties, and that it is unique; however, we will just take that fact as given.

C.5. Differential forms and Euclidean vector calculus

We can use differential forms on oriented Euclidean space $\mathcal{X}$ to better understand the vector-calculus operators discussed in Appendix A. First, as discussed above, the differential of a scalar field is a 1-form. The gradient is the vector field obtained from the differential by application of the musical isomorphism, that is:

$$\text{grad}(f) = (df)^\sharp \tag{C.39}$$

Thus, the gradient corresponds to the exterior derivative of a scalar field.

The exterior derivative of a 1-form $\mathbf{P}$ is a 2-form $d\mathbf{P}$, whose components (with respect to some orthonormal basis $\mathbf{e}^i$ on $\mathbb{X}^*$) are

$$(d\mathbf{P})_{ij} = \nabla_i P_j - \nabla_j P_i \tag{C.40}$$

where P_i are the components of $\mathbf{P}$ with respect to that same basis (i.e. $\mathbf{P} = P_i \mathbf{e}^i$), and ∇_i is the directional derivative with respect to the dual basis $\vec{e}_i$. As a result, if we take the Hodge dual then we obtain a 1-form

$$(\star d\mathbf{P})_1 = \nabla_2 P_3 - \nabla_3 P_2 \tag{C.41}$$

$$(\star d\mathbf{P})_2 = \nabla_3 P_1 - \nabla_1 P_3 \tag{C.42}$$

$$(\star d\mathbf{P})_3 = \nabla_1 P_2 - \nabla_2 P_1 \tag{C.43}$$

which we recognise as the same pattern of components as the curl; in more intrinsic geometric language, for any vector field $\vec{V}$,

$$\text{curl}(\vec{V})^\flat = \star d(\vec{V}^\flat) \tag{C.44}$$

So we can take the curl of a vector field by flattening it (to a 1-form), taking its exterior derivative (2-form), applying the Hodge dual (1-form) and sharpening it (vector field). Hence, the curl operator corresponds to the exterior derivative of a 1-form.

Finally, the exterior derivative of a 2-form $\mathbf{T}$ is a 3-form $d\mathbf{T}$, which we can express as

$$d\mathbf{T} = (\nabla_1 T_{23} + \nabla_2 T_{31} + \nabla_3 T_{12})\mathbf{e}^1 \wedge \mathbf{e}^2 \wedge \mathbf{e}^3 \tag{C.45}$$

Thus, when we apply the Hodge star we obtain a scalar field

$$\star\, d\mathbf{T} = \nabla_1 T_{23} + \nabla_2 T_{31} + \nabla_3 T_{12} \tag{C.46}$$

It follows that given a vector field $\vec{V}$, if we first flatten it to a 1-form, then turn into a 2-form (via the Hodge star), then take the exterior derivative (to get a 3-form) and finally convert it into a scalar field (by the Hodge star again), we have obtained the divergence; that is,

$$\text{div}(\vec{V}) = \star d \star (\vec{V}^\flat) \tag{C.47}$$

So the divergence operator corresponds to taking the exterior derivative of a 2-form.

Bibliography

Aharonov, Y. and Bohm, D. (1959). Significance of Electromagnetic Potentials in the Quantum Theory. *Physical Review*, 115(3):485–491.

Albert, D. Z. (2000). *Time and Chance*. Harvard University Press.

Alexander, H. G., editor (1956). *The Leibniz-Clarke Correspondence*. Manchester University Press, Manchester.

Andreas, H. (2017). Theoretical Terms in Science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.

Arntzenius, F. and Greaves, H. (2009). Time Reversal in Classical Electromagnetism. *The British Journal for the Philosophy of Science*, 60(3):557–584.

Awodey, S. (2010). *Category Theory*. Oxford University Press, Oxford; New York.

Baez, J. C. and Muniain, J. P. (1994). *Gauge Fields, Knots, and Gravity*. World Scientific, Singapore.

Baez, J. C. and Shulman, M. (2010). Lectures on *n*-categories and cohomology. In Baez, J. C. and May, J. P., editors, *Towards Higher Categories*, number v. 152 in The IMA Volumes in Mathematics and Its Applications, pages 1–68. Springer, New York.

Barbour, J. B. (1989). *Absolute or Relative Motion: A Study from a Machian Point of View of the Discovery and the Structure of Dynamical Theories*. Cambridge University Press, Cambridge.

Barrett, T. W. (n.d.). Structure and Equivalence. *Philosophy of Science*.

Barrett, T. W. and Halvorson, H. (2016a). Glymour and Quine on Theoretical Equivalence. *Journal of Philosophical Logic*, 45(5):467–483.

Barrett, T. W. and Halvorson, H. (2016b). Morita Equivalence. *The Review of Symbolic Logic*, 9(3):556–582.

Belot, G. (1998). Understanding Electromagnetism. *The British Journal for the Philosophy of Science*, 49(4):531–555.

Belot, G. (2013). Symmetry and equivalence. In Batterman, R. W., editor, *The Oxford Handbook of Philosophy of Physics*. Oxford University Press, New York.

Brading, K. and Brown, H. R. (2004). Are Gauge Symmetry Transformations Observable? *The British Journal for the Philosophy of Science*, 55(4):645–665.

Brown, H. R. (2005). *Physical Relativity: Space-Time Structure from a Dynamical Perspective*. Oxford University Press, Oxford.

Carnap, V. R. (1958). Beobachtungssprache Und Theoretische Sprache. *Dialectica*, 12(3-4):236–248.

Dasgupta, S. (2016). Symmetry as an Epistemic Notion (Twice Over). *The British Journal for the Philosophy of Science*, 67(3):837–878.

Dewar, N. (2019a). Ramsey Equivalence. *Erkenntnis*, 84(1):77–99.

Dewar, N. (2019b). Sophistication about Symmetries. *The British Journal for the Philosophy of Science*, 70(2):485–521.

Dewar, N. (n.d.). On recovering internal structure from categorical structure.

Earman, J. (1986). *A Primer on Determinism*. The Western Ontario Series in Philosophy of Science. Springer Netherlands.

Earman, J. and Norton, J. (1987). What Price Spacetime Substantivalism? The Hole Story. *The British Journal for the Philosophy of Science*, 38(4):515–525.

Feynman, R. P., Leighton, R. B., and Sands, M. (2011). *The Feynman Lectures on Physics, Vol. II: The New Millennium Edition: Mainly Electromagnetism and Matter*. Basic Books.

Fletcher, S. C. (2012). What counts as a Newtonian system? The view from Norton's dome. *European Journal for Philosophy of Science*, 2(3):275–297.

Gomes, H. and Butterfield, J. (n.d.). On the empirical significance of symmetries. page 48.

Greaves, H. and Wallace, D. (2014). Empirical Consequences of Symmetries. *The British Journal for the Philosophy of Science*, 65(1):59–89.

Halvorson, H. (2012). What Scientific Theories Could Not Be. *Philosophy of Science*, 79(2):183–206.

Halvorson, H. (2019). *The Logic in Philosophy of Science*. Cambridge University Press, Cambridge.

Healey, R. (1997). Nonlocality and the Aharonov-Bohm Effect. *Philosophy of Science*, 64(1):18–41.

Healey, R. (2007). *Gauging What's Real*. Oxford University Press, Oxford.

Healey, R. (2009). Perfect Symmetries. *The British Journal for the Philosophy of Science*, 60(4):697–720.

Hodges, W. (1993). *Model Theory*. Number 42 in Encyclopedia of Mathematics and Its Applications. Cambridge University Press, Cambridge.

Hodges, W. (1997). *A Shorter Model Theory*. Cambridge University Press, Cambridge.

Hudetz, L. (2019). Definable Categorical Equivalence. *Philosophy of Science*, 86(1):47–75.

Ismael, J. and van Fraassen, B. C. (2003). Symmetry as a guide to superfluous theoretical structure. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*, pages 371–392. Cambridge University Press, Cambridge.

Ketland, J. (2004). Empirical Adequacy and Ramsification. *The British Journal for the Philosophy of Science*, 55(2):287–300.

Kosso, P. (2000). The empirical status of symmetries in physics. *The British Journal for the Philosophy of Science*, 51(1):81–98.

Leeds, S. (1999). Gauges: Aharonov, Bohm, Yang, Healey. *Philosophy of Science*, 66(4):606–627.

Leeds, S. (2006). Discussion: Malament on Time Reversal. *Philosophy of Science*, 73(4):448–458.

Lutz, S. (2015). What Was the Syntax-Semantics Debate in the Philosophy of Science About? *Philosophy and Phenomenological Research*, 91(3).

Malament, D. B. (2004). On the time reversal invariance of classical electromagnetic theory. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 35(2):295–315.

Malament, D. B. (2008). Norton's slippery slope. *Philosophy of Science*, 75(5):799–816.

Malament, D. B. (2009). Notes on Geometry and Spacetime.

Malament, D. B. (2012). *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory*. University of Chicago Press, Chicago, IL.

Manzano, M. (1996). *Extensions of First Order Logic*. Cambridge University Press, Cambridge.

Marquis, J.-P. (2020). Category Theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition.

Maudlin, T. (1998). Healey on the Aharonov-Bohm Effect. *Philosophy of Science*, 65(2):361–368.

Maudlin, T. (2012). *Philosophy of Physics: Space and Time*. Princeton University Press, Princeton, NJ.

Maudlin, T. (2018). Ontological Clarity via Canonical Presentation: Electromagnetism and the Aharonov–Bohm Effect. *Entropy*, 20(6):465.

Møller-Nielsen, T. (2017). Invariance, Interpretation, and Motivation. *Philosophy of Science*, 84(5):1253–1264.

Newman, M. H. A. (1928). Mr. Russell's "Causal Theory of Perception". *Mind*, 37(146):137–148.

Norton, J. D. (2008). The Dome: An Unexpectedly Simple Failure of Determinism. *Philosophy of Science*, 75(5):786–798.

Nounou, A. M. (2003). A fourth way to the Aharanov-Bohm effect. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*, pages 174–199. Cambridge University Press, Cambridge.

Olver, P. J. (1986). *Applications of Lie Groups to Differential Equations*. Springer-Verlag, New York, NY.

Pooley, O. (2006). Points, particles, and structural realism. In Rickles, D., French, S., and Saatsi, J., editors, *The Structural Foundations of Quantum Gravity*, pages 83–120. Oxford University Press, Oxford, UK.

Psillos, S. (2000). Carnap, the Ramsey-Sentence and Realistic Empiricism. *Erkenntnis*, 52(2):253–279.

Ramsey, F. P. (1931). Theories (1929). In Braithwaite, R. B., editor, *The Foundations of Mathematics and Other Logical Essays*. Routledge and Kegan Paul, London.

Read, J. and Møller-Nielsen, T. (2018). Motivating dualities. *Synthese*.

Roberts, J. T. (2008). A Puzzle about Laws, Symmetries and Measurability. *The British Journal for the Philosophy of Science*, 59(2):143–168.

Russell, B. (1927). *The Analysis of Matter*. Kegan Paul, London.

Saunders, S. (2003). Physics and Leibniz's principles. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*, pages 289–308. Cambridge University Press, Cambridge, UK.

Saunders, S. (2013). Rethinking Newton's Principia. *Philosophy of Science*, 80(1):22–48.

Shapiro, S. (1991). *Foundations without Foundationalism: A Case for Second-Order Logic*. Oxford University Press, Oxford.

Sklar, L. (1982). Saving the noumena. *Philosophical Topics*.

Stein, H. (1967). Newtonian space-time. *Texas Quarterly*, 10:174–200.

Teh, N. J. (2016). Galileo's Gauge: Understanding the Empirical Significance of Gauge Symmetry. *Philosophy of Science*, 83(1):93–118.

Wallace, D. (2003). Time-dependent symmetries: The link between gauge symmetries and indeterminism. In Brading, K. and Castellani, E., editors, *Symmetries in Physics: Philosophical Reflections*, pages 163–173. Cambridge University Press, Cambridge.

Wallace, D. (2014). Deflating the Aharonov-Bohm Effect. *arXiv:1407.5073 [quant-ph]*.

Wallace, D. (n.d.a). Isolated Systems and their Symmetries, Part I: General Framework and Particle-Mechanics Examples. http://philsci-archive.pitt.edu/16623/.

Wallace, D. (n.d.b). Isolated systems and their symmetries, part II: Local and global symmetries of field theories. http://philsci-archive.pitt.edu/16624/.

Wallace, D. (n.d.c). Observability, redundancy and modality for dynamical symmetry transformations.

Weatherall, J. O. (2016). Are Newtonian Gravitation and Geometrized Newtonian Gravitation Theoretically Equivalent? *Erkenntnis*, 81(5):1073–1091.

Wheeler, J. T. (2007). Gauging Newton's law. *Canadian Journal of Physics*, 85(4):307–344.

Wilson, M. (2009). Determinism and the Mystery of the Missing Physics. *The British Journal for the Philosophy of Science*, 60(1):173–193.

Winnie, J. A. (1986). Invariants and objectivity: A theory with applications to relativity and geometry. In Colodny, R. G., editor, *From Quarks to Quasars, University of Pittsburgh Press, Pittsburgh*, pages 71–180. University of Pittsburgh Press, Pittsburgh.