-5/5-63

49- /

26084

12/

Symmetry as Bias: Rediscovering Special Relativity

Michael R. Lowry

AI Branch
M.S. 269-2
NASA Ames Research Center
Moffett Field, CA 94035
lowry@pluto.arc.nasa.gov

Abstract

This paper describes a rational reconstruction of Einstein's discovery of special relativity, validated through an implementation: the Erlanger program. Einstein's discovery of special relativity revolutionized both the content of physics and the research strategy used by theoretical physicists. This research strategy entails a mutual bootstrapping process between a hypothesis space for biases, defined through different postulated symmetries of the universe, and a hypothesis space for physical theories. The invariance principle mutually constrains these two spaces. The invariance principle enables detecting when an evolving physical theory becomes inconsistent with its bias, and also when the biases for theories describing different phenomena are inconsistent. Structural properties of the invariance principle facilitate generating a new bias when an inconsistency is detected. After a new bias is generated, this principle facilitates reformulating the old, inconsistent theory by treating the latter as a limiting approximation. The structural properties of the invariance principle can be suitably generalized to other types of biases to enable *primal-dual* learning.

Introduction[1]

Twentieth century physics has made spectacular progress toward a grand unified theory of the universe. This progress has been characterized by the development of unifying theories that are then subsumed under even more encompassing theories. Paradigm shifts are nearly routine, with the postulated ontology of the universe changing from the three dimensional absolute space of Newtonian physics, to the four dimensional space-time of relativistic physics, and through many other conceptual changes to current string theories embedded in ten dimensions. Theoretical physicists attribute much of the success of their discipline to the research strategy first invented by Einstein for discovering the theory of relativity [Zee 86].

At the heart of Einstein's strategy was the primacy of the principle of *invariance:* the laws of physics are the same in all frames of reference. This principle applies to reference frames in different orientations, displaced in time and space, and moreover to reference frames in relative motion. This principle also applies to many other aspects of physics, including symmetries in families of subatomic particles. The application of the invariance principle to "two systems of coordinates, in uniform motion of parallel translation relatively to each other" was Einstein's first postulate: the principle of special relativity [Einstein 1905].

Einstein's genius lay in his strategy for using the invariance principle as a means of unifying Newtonian mechanics and Maxwell's electrodynamics. This strategy of unifying different areas of physics through the invariance principle is responsible for many of the advances of theoretical physics. In the parlance of current machine learning theory, Einstein's strategy was to combine the principle of special relativity with his second postulate, the constancy of the speed of light in a vacuum, to derive a new bias. (This second postulate is a consequence of Maxwell's equations; [Einstein 1905] notes that experimental attempts to attribute it to a light medium were unsuccessful.) This new bias was designed and verified to be consistent with Maxwell's electrodynamics, but was inconsistent with Newton's mechanics. Einstein then reformulated Newton's mechanics to make them consistent with this new bias. He did this by treating Newton's mechanics as a limiting approximation, from which the relativistic laws were derived through generalization by the new bias.

Einstein's strategy is a model for scientific discovery that addresses a fundamental paradox of machine learning theory: to converge on a theory from experimental evidence in non-exponential time, it is necessary to incorporate a strong bias [Valiant 84], but the stronger the bias the more likely the 'correct' theory is excluded from consideration. Certainly any conventional analysis of what could be learned in polynomial time would exclude a grand unified theory of physics. The paradox can be avoided by

[1] This research was supported in part by a sabbatical while the author was a member of the Kestrel Institute, and in part by a subcontract through Recom Technologies to NASA Ames Research Center.

machine learning algorithms that have capabilities for reasoning about and changing their bias. Even if a strong bias is ultimately 'incorrect', it is still possible to do a great deal of useful theory formation before the inconsistencies between the bias and empirical facts becomes a limiting factor. The success of the Galilean/Newtonian framework is an obvious example. To avoid the paradox, a machine learning algorithm needs to detect when a bias is inconsistent with empirical facts, derive a better bias, and then reformulate the results of learning in the incorrect bias space into the new bias space [Dietterich 91]. The Erlanger program described in this paper is such an algorithm.

Einstein's strategy is essentially a mutual bootstrapping process between two interrelated hypothesis spaces: a space for biases, and a space for physical theories. The invariance principle defines the space of biases; each bias is a different postulated set of symmetries of the universe, formalized through a group of transformations. The invariance principle also defines a consistency relationship that mutually constrains the bias space and the space for physical theories. The hypothesis space for biases has a rich lattice structure that facilitates generating a new bias when a shift of bias is necessary. The hypothesis space for physical theories has an approximation relation between theories (limit homomorphisms) that, after a shift in bias, facilitates generating a new theory from an old (approximate) theory and the new bias. The entire process converges if learning in the bias space converges.

This paper builds upon the considerable body of literature on relativity and the role of symmetry in modern physics. Its contribution includes identifying and formalizing the structural relationships between the space of biases and the old and new theories that enabled Einstein's strategy to succeed, in other words, made it computationally tractable. The tactics for carrying out the components of this strategy have been implemented in the Erlanger program, written in Mathematica v.1.2.

The next section of this paper presents an overview of Einstein's strategy. The following section introduces the invariance principle, which determines the consistency relationship between a bias and a physical theory. It also describes the procedure for detecting inconsistency. The following section presents the tactic for computing a new bias using the invariance principle. It takes the reader through the Erlanger program's derivation of the Lorentz transformations. The section after defines *limit homomorphisms*, a formal semantics for approximation. The following section describes BEGAT: BiasEd Generalization of Approximate Theories, an algorithm that uses the invariance principle and the semantics of limit homomorphisms to generate components of the new theory. The paper concludes with a generalization of Einstein's strategy called *primal-dual* learning, which might be applied to other types of biases.

Overview of Einstein's Strategy

Einstein's strategy for deriving special relativity will first be explained through an analogy with symmetries and tangents of geometric figures. Then the structural components of the invariance principle interrelating the bias space and the space of physical theories will be outlined and the overall research strategy described with respect to these components. The next section will describe the mathematics of the invariance principle as it applies to theories of physics.

Symmetry and Group Theory

The symmetries of a geometric figure are invertible transformations that map the figure to itself. For example, a square is mapped to itself by various transformations about its center: horizontal reflections, vertical reflections, and ninety degree rotations. Because these transformations are invertible, they form a group.

A group is any set with a constant identity element, a binary operation defined on any two elements, and an inverse operation mapping any element to its inverse. A transformation group consists of elements which are transformations of some other set S; each transformation is a bijection from S to S. A transformation T defined on S is an *automorphism* of a subset $F \supseteq S$ iff $T(F) = F$. Hence if S is the two dimensional plane and F is a geometric figure such as a square, then the symmetries of F are those transformations T such that $T(F) = F$. Restrictions can be placed on the transformations considered; for example, transformations that preserve topological structure are called homeomorphisms while transformations that preserve distance are called isometries. The isometries of a square include horizontal reflections, vertical reflections, and multiples of ninety degree rotations about it center.

Symmetries can be represented through transformation equations; for example, the equations for a rotation of θ degrees about the origin in two dimensions define new primed coordinates for each point in terms of the original coordinates: $x' = x\cos\theta - y\sin\theta$, $y' = x\sin\theta + y\cos\theta$. If θ is a constant, then these equations represent a single transformation. If θ is a parameter, then these equations represent a set of transformations. Note that for any θ , a circle with its center at the origin is mapped to itself. Hence these equations denote a set of automorphisms of all origin-centered circles. One way to prove this algebraically is to solve for the equation of a circle, i.e., reduce $x^2 + y^2 = r^2$ to a set of functions for y in terms of x for different quadrants, plug the definitions of these functions into the transformation equations, and then show that the new points also satisfy these equations.

The method implemented in the Erlanger program is slightly different because it is based upon an equivalent but alternative approach to defining symmetries. (See [Friedman 83] for a thorough analysis of the relation between these two approaches, as applied to space-time theories.) Instead of viewing the transformations as mappings from points to points within a single reference frame, the transformations are viewed as mappings between reference frames. A figure is symmetric if it appears exactly the same in the new reference frame as it does in the old reference frame. In this alternative approach, the transformation equations are the same except that the sign of the parameter is inverted, because rotating the reference frame θ about the origin is equivalent to rotating the figure by $-\theta$ about the origin: $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$.

An Analogy to Einstein's Strategy

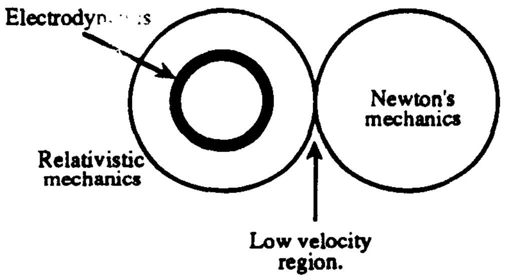

Figure 1.

Einstein's strategy for deriving special relativity is illustrated through the simple geometric analogy in Figure 1. Newton's mechanics is represented by the circle on the right, its set of symmetries are all rotations and reflections about its center. This set of symmetries is inconsistent with the invariance of the speed of light, a deductive consequence of Maxwell's electrodynamics that is represented by the small bold circle on the left.

Einstein derived the set of symmetries consistent with the constant speed of light by first generalizing from the particular circle representing Newton's mechanics to symmetries for all possible circles, i.e., rotations and reflections about all possible centers. He then specialized this set of all possible circular symmetries by solving for the center of the circle consistent with the constant speed of light. This new symmetry was verified to be consistent with Maxwell's electrodynamics.

Einstein then derived relativistic mechanics, represented by the larger left circle, through two constraints: that it be circularly symmetric around the same center as electromagnetic phenomena, and that it be tangent to Newton's mechanics as relative velocities approach 0. Note that Newton's mechanics had only been empirically verified at low velocities compared to light; the rest of the circle was assumed from the originally postulated symmetries dating back to Galileo. In this manner Einstein unified electromagnetism and mechanics under the same set of symmetries while still accounting for the wealth of experimental confirmations of Newton's theory at low velocities compared to the speed of light. Although simplistic, this geometric analogy captures the essential extensional relationships between Newton's mechanics, Maxwell's electromagnetism, and relativistic mechanics.

One of the crucial facts about symmetry as bias is that the groups corresponding to different figures form a lattice ordered by the subset relation. (More generally, the ordering is defined through group homomorphisms.) There is a contravariant relation between the complexity of an object and its set of symmetries. For example, a square is more complex than a circle, hence the group of transformations for a square is a subset of the group of transformations for a circle. As explained in the next section, this relation between geometric figures and their symmetries also holds between theories of physics and their symmetries. This contravariant relation is essential to the bootstrap learning of Einstein's strategy.

Structural Relations Exploited in Einstein's Strategy

Figure 2 illustrates the structural relations between the bias space and the space for physical theories that was exploited by Einstein, and indicates how these same structural relations might be exploited in other types of bias.

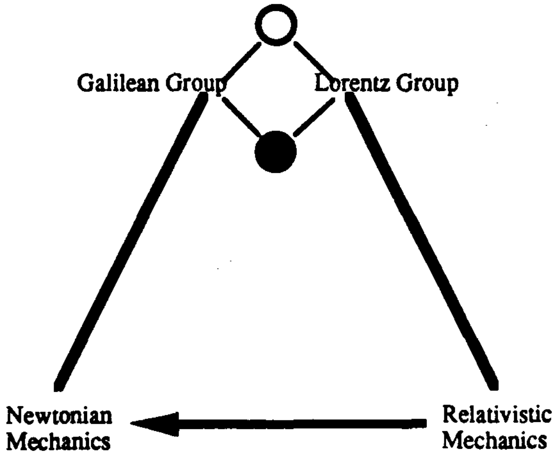

Figure 2

1. The diamond represents a space of biases for physical laws. The biases are different postulated symmetries of the universe. As modern physics has evolved, the bias has evolved. Each bias in this space is formalized as a transformation group.

2. The consistency relationship between a bias (a transformation group) and a physical theory is represented by a solid black line. The diagram illustrates that

Newtonian mechanics is consistent with the Galilean transformation group.

3. When an inconsistency is detected between an experimental fact and the current bias, then a new bias is computed. The new bias is computed by combining the new observation, an upper bound (represented by a hollow circle) and lower bounds (represented by a solid black circle). The upper bound is a superset of transformations that constrains the types of transformations that are considered. The transformations in this superset that are consistent with the new observation are selected for the new bias. This selection is done by symbolically solving for those transformations that are consistent with the new observation, rather than enumerating over all the transformations in the upper bound. The calculation is simplified through the use of lower bounds. Einstein derived the Lorentz transformations through this procedure.

4. Laws in the new hypothesis space are constrained to be consistent with the new bias and also to have, as a limiting approximation, the laws in the old hypothesis space. This limiting approximation is indicated by the arrow from relativistic mechanics to Newtonian mechanics. In fact, the new laws can often be derived from the old laws by using the new bias to reformulate the old laws. This was the method Einstein used to generate relativistic mechanics.

The power of Einstein's strategy is that his framework scales up from special relativity through the history of twentieth century physics, although the mathematics becomes considerably more complex. From the viewpoint of machine learning, the power of Einstein's strategy is his mutual bootstrapping between the bias space and the hypothesis spaces by exploiting the structural relationship between them: the invariance principle.

Symmetry as Bias: the Invariance Principle

Symmetry is a unifying aesthetic principle that has been a source of bias in physics since ancient times. In modern physics this principle is stated as: 'the laws of physics are invariant for all observers.' An invariance claim is a universally quantified statement of the form 'For all events/histories of type *F*, for all reference frames of type *R*, Physical Theory *P* holds'. An invariance claim implies that a group of transformations mapping measurements between different observers also maps physical theory *P* onto itself. Such a group of transformations defines the postulated symmetries of the universe, and is the type of bias used by theoretical physicists. The transformations are parameterized by the relation between two different observers, such as their relative orientation or velocity. For example, Galileo defined the following transformation equations relating measurements for observers in constant relative velocity v parallel to the x-axis: $\{x' = x - vt, t' = t\}$. These transformations are consistent with Newton's theory of mechanics.

The invariance principle defines a consistency relationship between physical theories and groups of transformations. The following definitions are standard and sufficient for our purpose of understanding and implementing Einstein's strategy for deriving special relativity. However, the reader should be aware that these definitions are a simple starting point for a deep, well developed mathematical theory that has had a profound impact on theoretical physics. (A good mathematical exposition focused on special relativity is [Aharoni 65], a more sophisticated philosophical and foundational treatment is [Friedman 83].)

Below, $\mathcal{G}$ is a transformation group. An invariant operation is a special case of a covariant operation. Laws are invariant if they define the same relation after they are transformed by the action of the transformation group. A sufficient condition for a theory to be invariant with respect to a transformation group $\mathcal{G}$ is if all the operations are covariant and all the laws are invariant.

Invariance of an operation or form:

$\text{Invariant}(op, \mathcal{G}) \Leftrightarrow \forall(g \in \mathcal{G}, x1 \ldots x_n)$

$op(x_1, x_2, \ldots, x_n) = op(g(x_1, x_2, \ldots, x_n))$

Covariance of an operation or form:

$\text{Covariant}(op, \mathcal{G}) \Leftrightarrow$

$\forall(g \in \mathcal{G}, x1 \ldots x_n)$

$op(g(x_1, x_2, \ldots, x_n)) = g(op(x_1, x_2, \ldots, x_n))$

$\Leftrightarrow \forall(g \in \mathcal{G}, x1 \ldots x_n)$

$op(x_1, x_2, \ldots, x_n) = g^{-1}(op(g(x_1, x_2, \ldots, x_n)))$

Invariance of a physcial law expressed as a universally quantified equation:

$\text{Invariant}(\forall(\ldots)\ \text{t1}(\ldots) = \text{t2}(\ldots), \mathcal{G}) \Leftrightarrow$

$\forall(g \in \mathcal{G}, x1 \ldots x_n)$

$\text{t1}(x_1, x_2, \ldots, x_n) = \text{t2}(x_1, x_2, \ldots, x_n)$

$\equiv \text{t1}(g(x_1, x_2, \ldots, x_n)) = \text{t2}(g(x_1, x_2, \ldots, x_n))$

More generally, a theory is invariant with respect to a transformation group $\mathcal{G}$ iff all the transformations in the group are automorphisms of the models of the theory. This is equivalent to proving that the theory and the transformation equations together imply the same theory in other frames of reference (though see [Friedman 83] for qualifications).

Invariance of a theory $\mathcal{T}$:

$\text{Invariant}(\mathcal{T}, \mathcal{G}) \Leftrightarrow$

$\forall(g \in \mathcal{G})\,\text{Models}(\mathcal{T}) \equiv g(\text{Models}(\mathcal{T}))$

$\equiv \forall(g \in \mathcal{G})\,\mathcal{T} \models \mathcal{T}/g$ and $\mathcal{T}/g \models \mathcal{T}$

where $\mathcal{T}/g$ denotes substituting variables with the terms defined by the transformation equations.
Because of the inverse property of groups,
the two conjunctions imply each other:

$\forall(g \in \mathcal{G})\,\mathcal{T} \models \mathcal{T}/g^{-1}$

implies $\forall(g \in \mathcal{G})\,\mathcal{T}/g \models \left(\mathcal{T}/g^{-1}\right)/g$

implies $\forall(g \in \mathcal{G})\,\mathcal{T}/g \models \mathcal{T}$

To check an invariant predicate, the Erlanger program back-substitutes transformation equations into a form or law and then compares the result to the original form or law. If the function or relation are the same, then the invariant predicate is true. In essence the Erlanger program assumes the law holds good in the original reference frame and then transforms the law into measurements that would be observed in a new frame of reference. (This can be done independent of whether the law is invariant.) If these measurements agree with the law stated in the new frame of reference, then the law is invariant. The steps of the algorithm are described below and illustrated with the example of determining whether the Galilean transformations are consistent with the constant speed of light, Einstein's second postulate. The input is the definition of the law for the constant speed of light, and the transformation equations relating variables in the original frame of reference to the variables in the new (primed) frame of reference:

Invariant($x^2 = c^2t^2$, $\{x = x' + vt', t = t'\}$)

1. Solve the law in the new frame of reference to derive expressions for dependent variables (This turns a relation between variables into a disjunction of sets of substitutions.): $\{\{x' = ct'\}, \{x' = -ct'\}\}$

2. Use the parameterized transformation equations to substitute expressions in the new frame of reference for variables in the old frame of reference; this yields a new law relating measurements in the new frame of reference: $(x' + vt')^2 = c^2t'^2$

3. The substitutions derived in step 1 are applied to the new law derived in step 2:

$$\{(ct' + vt')^2 = c^2t'^2, (-ct' + vt')^2 = c^2t'^2\}$$

4. If the law(s) derived in step 3 is a valid equality(ies), then the law(s) is invariant. For this example they are not, so the Erlanger program determines that Einstein's second postulate is inconsistent with the Galilean transformations.

Deriving a New Bias

The invariance principle can be used not only to verify that a physical law is consistent with a particular bias, but also to generate a new bias when a physical law is inconsistent with the current bias, as when the constant speed of light is inconsistent with the Galilean transformations. There are important structural aspects of the invariance principle that enabled this aspect of Einstein's strategy to succeed. In particular, the consistency relationship is contravariant: a weaker physical theory is consistent with a larger set of transformations. (For the purposes of this paper, 'weaker' can be thought of as 'fewer deductive consequences', though this is not entirely correct. This only holds if each law transforms into itself.) Thus when an inconsistency is detected between a bias represented by a set of transformations and an evolving physical theory, the physical theory can be relaxed, leading to an enlarged set of transformations. This enlarged set is then filtered to compute the new bias.

Assume that a physical theory T (e.g. Newton's mechanics) is consistent with a transformation group G (e.g. the Galilean group). Further assume that G is the largest transformation group consistent with T. Then a new empirical fact e is observed (e.g. the constant speed of light), such that e is not consistent with G. Then T is relaxed to T' (e.g. Newton's first law), thereby enlarging G to G' (e.g. the set of all linear transformations). The new bias is the subset of G', i.e. G''(e.g. the Lorentz group), such that T' with e is consistent with G''. Then the laws in $(T - T')$ are transformed so that they are consistent with G'' and have as limiting approximations the original laws. This section describes an implemented algorithm for deriving G'', while the next sections describe transforming the laws in $(T - T')$. These same algorithms can also be used when trying to unify theories with different biases, such as Newton's mechanics and Maxwell's electromagnetism.

The Lorentz group is a set of transformations that relate the measurements of observers in constant relative motion. The Lorentz group is a sibling to the Galilean group in the space of biases. Einstein's derivation of the Lorentz transformations implicitly relied upon structural properties of the lattice of transformation groups. In particular, Einstein constrained the form of the transformations with an upper bound, derived from Newton's first law: a body in constant motion stays in constant motion in the absence of any force. This is his assumption of inertial reference frames, an assumption he relaxed in his theory of general relativity. The largest set of transformations consistent with Newton's first law are the four dimensional linear transformations. Of these, the spatial rotations and spatial/temporal displacements can be factored out of the derivation, because they are already consistent with Einstein's second postulate. (The Erlanger program does not currently have procedures implemented to factor out subgroups of transformations - these are under development.) This leaves an upper bound for a subgroup with three unknown parameters (a,d,f) whose independent parameter is the relative velocity (v):

$$x = a(x' + vt') \qquad t = dx' + ft'$$

This upper bound includes both the Galilean transformations and the Lorentz transformations. The DeriveNewBias algorithm takes the definition of an upper bound, such as the one above, including lists of the unknown and independent parameters, a list of invariants, a list of background assumptions, and information on the group properties of the upper bound. When this algorithm is applied to Einstein's second postulate of the constant speed of light, the derivation of the Lorentz transformations proceeds along roughly the same lines as that in Appendix 1 of [Einstein 1916]. This derivation and others are essentially a gradual accumulation of constraints on the unknown parameters of the transformations in the upper bound, until they can be solved exactly in terms of the independent parameter which defines the relation between two reference frames. The algorithm is described below, illustrated with the example of deriving the Lorentz transformations.

The input in this example to the DeriveNewBias algorithm is the upper bound given above, two invariants for a pulse of light - one going forward in the x direction and one going backwards in the x direction $\{x = ct, x = -ct\}$, the assumptions that the speed of light is not zero and that the relative velocity between reference frames is less than the speed of light, and information for computing the inverse of a transformation. The steps of the DeriveNewBias algorithm:

1. Constraints on the unknown parameters for the transformation group are derived separately from each individual invariant. This step is similar to the procedure which checks whether a law is invariant under a transformation group. However, instead of steps 3 and 4 of that procedure, the system of equations from steps 1 and 2 are jointly solved for constraints on the unknown parameters. For the two invariants for a pulse of light, the derived constraints are:

$$a = (-c^2 d + cf)/(c - v), \quad a = (c^2 d + cf)/(c + v)$$

2. The constraints from the separate invariants are combined through Mathematica's SOLVE function. In the example of the Lorentz derivation, this reduces the unknown parameters to a single unknown (*f*):

$$a = f, \quad d = (fv)/c^2$$

3. In the last step, the group properties are used to further constrain the unknown parameters. Currently the implemented algorithm only uses the inverse property of a group, but the compositional property is another source of constraints that could be exploited. First, the constraints on the unknown parameters are substituted into the upper bound transformation definition, yielding a more constrained set of transformations. For the Lorentz example this yields:

$$x = f(x' + vt') \qquad t = ft' + fvx'/c^2$$

Second, the inverse transformations are computed. The information given to the algorithm on the group properties of the upper bound define how the independent parameter for the transformation is changed for the inverse transformation. For relative velocity, this relation is simply to negate the relative velocity vector . This then yields the inverse transformations:

$$x' = f(x - vt) \qquad t' = ft - (fvx)/c^2$$

The inverse transformations are then applied to the right hand side of the uninverted transformations, thereby deriving expressions for the identity transformation:

$$x = f\left(f(x - vt) + v\left(ft - \frac{fvx}{c^2}\right)\right)$$

$$t = f\left(ft - \frac{fvx}{c^2}\right) + \frac{fvf(x - vt)}{c^2}$$

These expressions are then solved for the remaining unknown parameters of the transformation (e.g. *f*), whose solution is substituted back into the transformations:

$$x = (x' + vt')\frac{\sqrt{c}}{\sqrt{2}}\sqrt{\frac{1}{c - v} + \frac{1}{c + v}}$$

$$t = \left(t' + \frac{v}{c^2}x'\right)\frac{\sqrt{c}}{\sqrt{2}}\sqrt{\frac{1}{c - v} + \frac{1}{c + v}}$$

The result is the new bias, which in this example is equivalent to the standard definition of the Lorentz transformations (the definitions above are in Mathematica's preferred normal form).

Limit Homomorphisms: Approximations between Theories.

Once a new bias is derived, a learning algorithm needs to transfer the results of learning in the old bias space into the new bias space. Unless the relationship between the old bias and the new bias can be exploited, in the worst case this means running the learning algorithm with the new bias over all the examples used to derive the old theory. The shift in bias from the Galilean transformation group to the Lorentz transformation group required a global reformulation of all the theories of physics, from kinematics to fluid dynamics, and later quantum mechanics. Yet in all these reformulations, the relativistic theory was derived from its non-relativistic counterpart without exhaustively considering the experimental evidence justifying the non-relativistic theory. This was done by treating the non-relativistic theory as an approximation to the new, unknown relativistic theory; and combining this constraint with the Lorentz transformations to derive a corresponding relativistic theory.

A theory such as Newton's mechanics that has a high degree of experimental validation over a range of phenomena (e.g. particles interacting at low velocities compared to the speed of light), represents a summary of many experimental facts. If a new theory is to account for these same experimental facts, it must agree with the observable predictions of the old theory over the same range of phenomena. Hence the old theory must approximate, to within experimental error, the new theory over this range of phenomena (and vice versa). By showing that an old theory is a limiting approximation to a new theory, it is unnecessary to exhaustively reconsider all the experimental evidence justifying the old theory. This approximation criteria for partially validating a new theory is well accepted, both within scientific communities and within the philosophy of science. However, the development of relativity theory went beyond a *post-hoc* verification of this approximation criteria: the approximation criteria was used to *derive* the new theory.

Various notions of "approximation" have been developed in AI to support reasoning between approximate theories, and even generating approximate theories from detailed theories [Ellman 90,92]. The problem of generating a new theory from an approximate theory and a new bias requires a precise definition of approximation with a well defined semantics. This section describes *limit homomorphisms*, which are homomorphisms that only hold in the limiting value of some parameter. Limit homomorphisms can be viewed as an extension of fitting parameter approximations [Weld 92] with additional algebraic structure that adds the constraints needed to derive the new theory, and not just model the approximation relation.

A limit homomorphism is a map *m* from one domain to another such that for corresponding functions f_1 and f_2 the following equality converges as the limit expression goes to the limiting value:

$$\lim_{\text{expr}\rightarrow\text{value}} m(f_1(x_1 \ldots x_n)) = f_2(m(x_1)\ldots m(x_n))$$

Another motivation for this definition of approximation is to resolve a fundamental disagreement between Kuhn's view of the paradigm shift from Newtonian to relativistic physics, and the view of most physicists. Most physicists agree with the logical positivists: Newtonian physics is a limiting approximation of Einstein's physics. Kuhn argues that this is spurious [Kuhn 62, pg. 102], because the corresponding concepts in the relativistic and Newtonian mechanics are different. A limit homomorphism combines a map between corresponding concepts and a limiting approximation, thus achieving a limiting approximation between two different conceptual domains.

A well known type of limit homomorphism within computer science is Ω-order computational complexity. For example, the Ω-order computational complexity of a sequence of program statements is the maximum of the Ω-order computational complexity of the individual statements:

$$\lim_{\mathit{input}\rightarrow\infty} \Omega(S_1;S_2;\ldots S_n) = \mathrm{Max}(\Omega(S_1),\Omega(S_2),\ldots\Omega(S_n))$$

To determine the Ω-order computational complexity of a program, this limit homomorphism is recursively applied to the definition of a program. Similarly, to determine the non-relativistic quantity corresponding to a relativistic quantity, the appropriate limit homomorphism is recursively applied to the definition of the relativistic quantity.

Within physics, limit homomorphisms define the relationship between new, unified theories and the older theories they subsume. If the mapping function *m* is invertible, then a limit homomorphism can be defined in the reverse direction. The limit homomorphisms between Newton's mechanics and different formulations of relativistic mechanics are invertible. Thus from an *a priori*, mathematical viewpoint neither Newtonian mechanics nor relativistic mechanics is intrinsically more general than the other - the mathematical relationship is symmetric; each is a limit homomorphism of the other. These theories agree on their predictions when velocities are low, but diverge as velocities approach the speed of light. Relativistic mechanics is *a posteriori* more general because its predictions agree with experimental facts for high velocities, hence the theory is more generally applicable. Relativistic mechanics is also extrinsically more general in the sense that its bias is consistent with electrodynamics, and hence relativistic mechanics and electrodynamics can be unified.

BEGAT: (BiasEd Generalization of Approximate Theories)

While the intrinsic mathematical relationship between Newtonian and relativistic physics is not one of generalization [Friedman 83], the *process* of generating relativistic mechanics from Newtonian mechanics is one of generalization. This section describes the mathematics justifying this process, and an implemented algorithm based on these mathematics that derives relativistic kinematics. Extensions currently undergoing implementation are described that will enable it to derive different formulations of relativistic dynamics.

It is clear from a reading of [Einstein 1905] that Einstein derived relativistic mechanics from Newtonian mechanics, by treating the latter as a limiting approximation that was valid in low velocity reference frames and applying the Lorentz transformations in order to generalize to the relativistic laws. For example, in section 10, paragraph 2 of [Einstein 1905]: "If the electron is at rest at a given epoch, the motion of the electron ensues in the next instant of time according to the equations [Newton's equations of motion] ... as long as its motion is slow." Einstein then generalized to the relativistic equation of motion by applying the Lorentz transformations to Newton's equations of motion. Einstein even constrained the laws of relativistic dynamics to have the same form as Newtonian dynamics.

This point needs to be made because [Kuhn 62], which many in AI take as a definitive source on scientific revolutions, argues otherwise with respect to the genetic relationship between Newtonian and relativistic mechanics [Kuhn 62, pg. 103]: "Though an out-of-date theory can always be viewed as a special case of its up-to-date successor, it must be transformed for the purpose. And the transformation is one that can be undertaken only with the advantages of hindsight, the explicit guidance of the more recent theory. .... but it [the old theory] could not suffice for the guidance of research." The first sentence is true, but the remaining part of the paragraph is demonstrably false as applied to Einstein's derivation of relativistic mechanics. As is clear from the selection of Einstein's paper in the preceding paragraph, Einstein not only used Newton's theory to guide his search for the proper relativistic laws, he transformed, with foresight, the old (Newtonian) laws to obtain the new (relativistic) laws. Few physicists or philosophers/historians of science currently subscribe to Kuhn's interpretation.

When both the old theory and the new theory comply with the invariance principle, then the difference in the biases will determine the limit point, i.e. the range of phenomena over which they must agree. The following mathematical sketch explains what this limit point must be, when the theories postulate the same number of dimensions. The two biases will share some subgroups in common (e.g. the spatial rotations) and differ in other subgroups (e.g. the subgroup for relative velocity). For the subgroups that differ, the identity transformations will be the same. Hence the value of the parameter (e.g. relative velocity) that yields the identity transformation must be the limit point (e.g. 0 relative velocity). Furthermore, assuming that the transformations in the differing subgroups are a continuous and smooth function of their parameter(s), and that the functions in the respective theories are smooth and continuous, then the bounding epsilon-delta requirements for a limit are satisfied.

Thus, given a new bias, the new theory must be derived so that it satisfies two constraints: the theory is invariant under the new bias, and the old theory is a limit homomorphism of the new theory. The limit homomorphisms between Newtonian physics and relativistic physics can be defined through the composition of tupling (or projections) that are invertible, with Lorentz transformations applied to the various entities of the theory. Because the Lorentz transformations are also invertible, the composition is invertible. In other words, the limit homomorphism is defined through a standard homomorphism at the limit point, which will be denoted h, and Lorentz transformations denoted g.

The two constraints on the new theory, that it be invariant under the new bias and that it have as a limiting approximation the old theory, can be solved to generate the new theory when the limit homomorphism is invertible. The new theory and the limit homomorphism are derived in tandem. In essence, the transformations in the new bias are used to 'rotate away' from the limit point, as Einstein 'rotated' a description of Newton's equations for an electron initially at rest to reference frames in which it was not at rest. (Here 'rotate' means applying the transformations in the subgroups of the new bias not contained in the old bias, e.g. the Lorentz transformations.)

For the operations of the new theory, these two constraints can often be directly combined as follows:

1. New, unknown operation is covariant wrt new bias:

$$\mathrm{op}(g(x_1,x_2,\dots,x_n)) = g(\mathrm{op}(x_1,x_2,\dots,x_n))$$

Equivalently: $\mathrm{op}(x_1,x_2,\dots,x_n) = g^{-1}(\mathrm{op}(g(x_1,x_2,\dots,x_n)))$

2. New, unknown operation has limit homomorphism to old operation op′:

$$\lim_{\substack{x_1,\dots,x_n \to \\ \text{limit point}}} h(\mathrm{op}(x_1,x_2,\dots,x_n)) = \mathrm{op}'(h(x_1),h(x_2),\dots,h(x_n))$$

Thus: $\mathrm{op}(x_1,x_2,\dots,x_n) =$

$$g^{-1}(h^{-1}(\mathrm{op}'(h(g(x_1)),h(g(x_2)),\dots,h(g(x_n)))))$$

where $g(x_1,x_2,\dots,x_n) = \text{limit point}$

In words, the new operation is obtained by :

1. Finding a transformation g that takes its arguments to a reference frame where the old operation is valid.

2. Applying the inverse transformation to define the value of the new operation in the original reference frame.

Applying BEGAT to derive the laws of the new theory is a similar two step process: first, a transformation is determined that takes the variables to a reference frame in which the old laws are valid, and then the inverse transformations are symbolically applied to the equations for the old laws.

The algorithm is underconstrained, because of the interaction of the definition of the new (unknown) operation and the definition of the (unknown) homomorphism h. In parts of [Einstein 1905], Einstein assumes that h is the identity, for example in his derivation of the relativistic composition of velocities (described below), and then derives an expression for the new operation. In other parts of [Einstein 1905], he assumes that the old operation and the new operation are identical, for example in his derivation of the relativistic equation of motion. In that derivation he kept the same form as the Newtonian equation (i.e. force = mass * acceleration) and then solved for a relativistic definition of inertial mass, and hence h. To his credit, Einstein recognized that he was making arbitrary choices [Einstein 1905 section 10, after definition of transverse mass]: "With a different definition of force and acceleration we should naturally obtain other values for the masses."

The following illustrates how the BEGAT algorithm works for a simple operation when h is the identity.

Note that when h is the identity: $\mathrm{op}(x_1,x_2,\dots,x_n) = g^{-1}(\mathrm{op}'(g(x_1),g(x_2),\dots,g(x_n))$
where $g(x_1,x_2,\dots,x_n) = \text{limit point}$

BEGAT takes as input the definition of the old operation, the list of transformations for the new bias, and a definition of the limit point. For the composition of velocities, the old operation is simply the addition of velocities:

Newton - Compose$(v1, v2) = v1 + v2$ where:
$v1$ is the velocity of reference frame R_1 w.r.t. R_0
$v2$ is the velocity of object A w.r.t. reference frame R_1
and the output is defined in reference frame R_0

The transformations are the Lorentz transformations derived earlier. The limit point is when R_1 is the same as R_0, i.e. $v1 = 0$. The first part of the reasoning for the BEGAT algorithm is at the meta-level, so it is necessary to understand some aspects of the notation used in the Erlanger program. Variables are represented by an uninterpreted function of the form:

var[event, component, reference-frame]. This form facilitates pattern matching. Transformations have representations both as lists of substitutions and as a meta-level predicate of the form:

Transform[start-frame, end-frame, independent-parameter] The independent parameter for relative velocity has the form: var[end-frame,relvelocity,start-frame]. Thus $v1$ is represented as var[R_1,relvelocity,R_0] and $v2$ as var[A,velocity,R_1].

1. BEGAT first solves for g, the transformation which takes the arguments to the limit point. This transformation maps the reference frame for the output to the reference frame for the limit point. The result is obtained by applying a set of rewrite rules at the meta-level:

Transform[R_0,R_1,var[R_0,relvelocity,R_1]]

This transformation maps reference frame R_0 to reference frame R_1.

2. BEGAT next solves for the value of the variables which are given to the old operation, i.e. $g(v1)$, $g(v2)$. For $g(v1)$ it symbolically solves at the meta-level for:

Apply[Transform[R_0,R_1,var[R_0,relvelocity,R_1]],
var[R_1,relvelocity, R_0]],

obtaining var[R_1,relvelocity,R_1], i.e. $g(v1)=0$

For $g(v2)$ it symbolically solves at the meta-level for:

Apply[Transform[R_0,R_1,var[R_0,relvelocity,R_1]],
var[A,velocity,
R_1]],

obtaining var[A,velocity,R_1], i.e. $g(v2)=v2$ since $v2$ is measured in R_1.

This meta-level reasoning about the application of transformations is necessary when the input variables and the output variables are defined in different reference frames.

3. BEGAT next symbolically applies the old operation to the transformed variables:

$$\text{Newton - compose}(g(v1), g(v2)) = 0 + v2 = v2$$

4. BEGAT finally applies the inverse transformation to this result to obtain the definition for the relativistic operation: Relativistic-compose($v1$,$v2$) =

Apply[Transform[R_1,R_0,var[R_1,relvelocity,R_0]],
var[A,velocity, R_1]]

The transformation derived previously for velocities is now applied to var[A,velocity, R_1], yielding the definition of the operator for relativistic composition of velocities: so BEGAT calls DeriveCompositeTransformation with the definition for velocity (i.e.$v = {}^{\Delta x}\!/\!_{\Delta t}$), and the Lorentz Transformations for the components of the definition of velocity - namely the transformations for the x co-ordinate and the time co-ordinate derived earlier. DeriveCompositeTransformation then symbolically applies these transformations to the components of the definition, and then calls Mathematica's SOLVE operation to eliminate the Δx, Δt components from the resulting expression. The result is the same definition as Einstein obtained in section 5 of [Einstein 1905]:

$$\text{Relativistic} - \text{compose}(v1, v2) = (v1 + v2) / (1 + (v1v2) / c^2)$$

Deriving Relativistic Dynamics

This subsection describes how the invariance principle can be used to derive other components of the new theory and the limit homomorphism, illustrated with one derivation of relativistic dynamics. Different background assumptions lead to different limit homomorphisms m and different formulations of the equations for relativistic dynamics. In his original paper, Einstein reformulated the Newtonian equation by measuring the force in the reference frame of the moving object and the inertial mass and acceleration in the reference frame of the observer. (In essence, Einstein did not complete step 2, for reasons too complex to explain here.) This leads to a projection of the Newtonian mass into separate transverse and longitudinal relativistic masses.

A subsequent formulation of relativistic dynamics consistently measures masses, accelerations, momentum and energy in the reference frame of the observer, resulting in a single relativistic mass that varies with the speed of the object. In this formulation the mass of a system is the sum of the masses of its components, and is conserved in elastic collisions. The modern formulation of relativistic dynamics, based on Minkowski's space-time and Einstein's tensor calculus, requires that components that transform into each other be tupled together. Thus because time coordinates transform into spatial coordinates, time and space are tupled into a single 4-vector. Consequently energy and momentum are also tupled together. In this case m maps Newtonian inertial mass to rest mass, and maps Newtonian acceleration and forces to their 4-vector counterparts.

In all three cases the derivation strategy is based directly on the invariance principle and the principle that the non-relativistic theory be a limiting approximation to the relativistic theory. The strategy is to assume that the laws of dynamics are invariant under the Lorentz transformations, and then to solve for the limit homomorphism that makes them invariant. (If it is not possible to consistently solve for the limit homomorphism, then the theory cannot be invariant.) These limit homomorphisms are composed of two maps: first a tupling or projection map from the components of the original theory to components of the new theory ($\mathcal{H}$), and second of Lorentz transformations for components of the new theory($\mathcal{G}$). These two maps are generated by the derivations.

Derivations based on the tensor calculus are the most elegant because the tensor calculus is essentially a syntactic encoding of the invariance principle, as applied to biases defined by groups of linear homogenous transformations. However, an explanation of the group-theoretic basis of the tensor calculus is beyond the scope of this paper. Instead we will describe the justification and strategy that applies to the first two derivations of relativistic dynamics, and then illustrate it with part of Einstein's original derivation. This derivation has been partially simulated in interactive mode with Mathematica 1.2. The justification and steps of this derivation are also the same as that for relativistic electrodynamics; more specifically, the derivation of the Lorentz transformations for electric and magnetic fields.

Recall the definition of the invariance of a theory under a transformation group $\mathcal{G}$, where $\mathcal{NT}$ is the new theory:

$$\text{Invariant}(\mathcal{NT}, \mathcal{G}) \Leftrightarrow \forall (g \in \mathcal{G}) \mathcal{NT} / g \models \mathcal{NT}$$

This is combined with the constraint that the old theory is a limit homomorphism of the new theory, where $\mathcal{LH}$ is the definition of the components of the old theory in terms of the components of the new theory:

$$\mathcal{NT} \cup \mathcal{LH} \models \mathcal{OT}$$

When the limit homomorphism is invertible, we also have:

$\mathcal{OT} \cup \mathcal{ILH} \models \mathcal{NT}$

Because this inverse limit homomorphism can be factored into a tupling/projection map $\mathcal{H}$ and the new bias $\mathcal{G}$, this last constraint can be combined directly with the invariance principle to yield a single constraint between the old theory and the new theory. :

$\forall(g \in \mathcal{G})(\mathcal{OT} \cup \mathcal{H}) / g \models \mathcal{NT}$

By the definition of a limit homomorphism, the old theory is defined with respect to the reference frame for which the limiting value holds (e.g. zero relative velocity). The transformations in $\mathcal{G}$ take the result of applying the tupling/projection map $\mathcal{H}$ to this reference frame and transform it to all other reference frames. The constraint is satisfied when the new theory, defined with respect to any reference frame $\mathcal{R}$, is a consequence of the old theory, the tupling/projection map $\mathcal{H}$, and the transformation g from the reference frame for the old theory to the reference frame $\mathcal{R}$. We will now show how this constraint can be used to derive the new theory, illustrated with Einstein's derivation of relativistic dynamics.

In all derivations of relativistic dynamics, it is assumed that the new equation has the same form as the Newtonian equation, but that the definition of the components might be different; according to $\mathcal{H}$ and $\mathcal{G}$. Thus if $\mathcal{H}$ and $\mathcal{G}$ are partially known, say $\mathcal{H}'$ and $\mathcal{G}'$ are defined for some of the components, then the remaining parts of $\mathcal{H}$ and $\mathcal{G}$ are derived by setting up the following unified constraint and solving for the remaining parts of the limit homomorphism:

$\forall(g \in \mathcal{G})(\mathcal{OT} \cup \mathcal{H}') / g' \models \mathcal{OT}'$

where $\mathcal{OT}'$ has the same form as the Newtonian theory but with new variables which are functions of corresponding variables in $\mathcal{OT}$ and the parameters of the transformation group $\mathcal{G}$.

Einstein's derivation of relativistic dynamics proceeded as follows. First, the old theory ($\mathcal{OT}$) was Newton's dynamics relating a particle's inertial mass, acceleration, and the force exerted upon the particle (Einstein considered the case where the force was exerted by an electric field with a particle of charge ε). This law is valid in the reference frame of the particle:

$$\mathcal{OT} \equiv m\frac{d^2x}{dt^2} = \varepsilon E_x \qquad m\frac{d^2y}{dt^2} = \varepsilon E_y \qquad m\frac{d^2z}{dt^2} = \varepsilon E_z$$

Through previous derivations, $\mathcal{H}'$ and $\mathcal{G}'$ were known for space, time, and electromagnetic fields; though Einstein did not use the transformations for the electromagnetic field. The map $\mathcal{H}'$ for space and time was the identity, while $\mathcal{G}'$ was the Lorentz transformation equations for space and time generated by DeriveNewBias. (A different background assumption where $\mathcal{H}'$ tuples space and time into a single 4-vector would yield the tensor formulation of relativistic dynamics). Thus Einstein needed to solve for the relativistic definition of inertial mass as a function of the non-relativistic mass and the parameter of the Lorentz transformation group; namely, the relative velocity between reference frames. Because the relative velocity is a vector quantity with x,y,z components; the definition of the inertial mass is also set up with x,y,z components. These components of the inertial mass might later be identified. In the following, v is the relative velocity between the reference frame of the particle and an observer moving in the positive x direction, and β is a term defined with respect to the magnitude of v: $\beta = {}^{1}\!/\!\sqrt{1-{}^{v^2}\!/_{c^2}}$. The unprimed variables are in the reference frame of the particle, while the primed variables are in the reference frame of the observer. The constraint relating Newton's dynamics, the Lorentz transformations, and relativistic dynamics is instantiated from the unified constraint above:

$$\forall v\ \mathcal{OT} \cup \left\{\begin{array}{c} x = \beta(x' + vt') \\ y = y' \\ z = z' \\ t = \beta\left(t' + x'\,{}^{v^2}\!/_{c^2}\right) \end{array}\right\} \vdash \left\{\begin{array}{l} m_x'(m,v)\dfrac{d^2x'}{dt'^2} = \varepsilon E_x \\ m_y'(m,v)\dfrac{d^2y'}{dt'^2} = \varepsilon E_y \\ m_z'(m,v)\dfrac{d^2z'}{dt'^2} = \varepsilon E_z \end{array}\right.$$

Note that Einstein defines the force in the reference frame of the particle, even on the right hand side. The equations for Newton's dynamics are then partially transformed into the reference frame of the observer by applying the Lorentz transformations, yielding a simplified constraint:

$$\left.\begin{array}{l} m\beta^3\dfrac{d^2x'}{dt'^2} = \varepsilon E_x \\ m\beta^2\dfrac{d^2y'}{dt'^2} = \varepsilon E_y \\ m\beta^2\dfrac{d^2z'}{dt'^2} = \varepsilon E_z \end{array}\right\} \vdash \left\{\begin{array}{l} m_x'(m,v)\dfrac{d^2x'}{dt'^2} = \varepsilon E_x \\ m_y'(m,v)\dfrac{d^2y'}{dt'^2} = \varepsilon E_y \\ m_z'(m,v)\dfrac{d^2z'}{dt'^2} = \varepsilon E_z \end{array}\right.$$

This constraint is then solved for definitions of the relativistic inertial mass in terms of the Newtonian inertial mass and the parameter between the reference frame of the particle and the observer. Solving this constraint is a simple directed inference problem [Smith 91]; reasoning backwards from the right hand side a match is derived between the variables for the relativistic inertial mass and terms on the left hand side:

$m_x'(m,v) = m\beta^3$

$m_y'(m,v) = m\beta^2$

$m_z'(m,v) = m\beta^2$

The definitions for the y and z components of the inertial mass are identical, so they can be combined into a single `transverse' inertial mass. In alternative derivations of relativistic dynamics, all the components of the inertial mass are identical.

While the particular derivation tactics currently implemented or undergoing implementation in the BEGAT algorithm might not be directly applicable to other types of biases, it is likely that analogues can be found. Research toward generalizing BEGAT is described after a review of related work.

Related Research

Within AI, this research is related to scientific discovery and theory formation [Shrager and Langley 90], qualitative physics [Weld and de Kleer 90], change of bias in machine learning [Benjamin 90a], and use of group theory [Benjamin 90b]. The research in this paper appears to be the first addressing the automated rediscovery of scientific revolutions of twentieth century theoretical physics. Most of the work in scientific theory formation has been on incremental theory revision (normal science). Previous research on scientific revolutions includes conceptual and qualitative accounts of the geological revolution in plate tectonics [Thagard and Nowak 90] and the chemical revolution of the oxygen theory [O'Rork, Morris, and Schulenburg 90]. Recently, [Thagard 92] has addressed automating the comparison of competing theories, and applied it to comparing Einstein's relativity theories with competing theories.

The notions of approximation within qualitative physics are closely related to limit homomorphisms. The well known calculii for qualitative physics reasoning usually include some sort of homomorphism from the reals [Forbus 84] [Kuipers 86]. The use of limits (fitting parameters) to define approximation relations between models is described in [Weld 89]. Within machine learning, research on declarative representations and reasoning about bias is most important, see the collection of papers in [Benjamin 90a]. The research described in this paper is one approach to addressing an open problem presented in [Dietterich 91]: analytically comparing biases. The declarative bias used in theoretical physics is group theory. A good collection of papers, many of which focus on the use of group theory in AI reasoning and problem solving, is in the workshop proceedings [Benjamin 90b].

The mathematical model and the research strategy presented in this paper are consistent with the physics literature. References accessible to the layman include [Zee 86] and [Davies and Brown 88]. With respect to that literature the chief innovations of this paper are the result of focusing on the structure of derivations with the aim of formalizing them. This focus is peculiar to AI; to the best of my knowledge it has not been addressed before. The closest previous works may be various pedagogical explanations found in textbooks such as [Skinner 82], [Taylor and Wheeler 66], and [French 68].

Conclusion: Toward Primal-Dual Learning

A hypothesis of this research is that Einstein's strategy for mutually bootstrapping between a space of biases and a space of theories has wider applicability than theoretical physics. Below we generalize the structural relationships of the invariance principle which enabled the computational steps of Einstein's derivation to succeed. We conjecture that there is a class of *primal-dual* learning algorithms based on this structure that have similar computational properties to primal-dual optimization algorithms that incrementally converge on an optimal value by alternating updates between a primal space and a dual space.

Let $\mathcal{B}$ be a set of biases with ordering relation $\lhd$ that forms a lattice. Let $\mathcal{T}$ be a set of theories with ordering relation $\prec$ that forms a lattice. Let $\mathcal{C}$ be a consistency relation on $\mathcal{B}\times\mathcal{T}$ such that:

$$\mathcal{C}(b,t) \text{ and } b' \lhd b \Rightarrow \mathcal{C}(b',t)$$

$$\mathcal{C}(b,t) \text{ and } t' \prec t \Rightarrow \mathcal{C}(b,t')$$

This definition is the essential property for a well-structured bias space: As a bias is strengthened, the set of theories it is consistent with decreases; as a theory is strengthened, the biases it is consistent with decreases. Hence $\mathcal{C}$ defines a contravariant relation between the ordering on biases and the ordering on theories.

Let $\mathcal{U}$ be the max bias function from $\mathcal{T} \rightarrow \mathcal{B}$ such that $\mathcal{C}(\mathcal{U}(t),t)$ and $\forall b\, \mathcal{C}(b,t) \Rightarrow b \lhd \mathcal{U}(t)$. Let $\mathcal{D}$ be a function from $\mathcal{B} \times \mathcal{T} \rightarrow \mathcal{B}$ such that $\mathcal{D}(b,t) = b \wedge \mathcal{U}(t)$, where $\wedge$ is the lattice meet operation.

$\mathcal{D}$ is the DeriveNewBias function, which takes an upper bound on a bias and filters it with a (new) theory or observation to obtain a weaker bias. (For some applications of primal-dual learning, $\mathcal{D}$ should take a lower bound on a bias and filter it with a new theory or observation to obtain a stronger bias.) $\mathcal{D}$ is well-defined whenever $\mathcal{B}$, $\mathcal{T}$, and $\mathcal{C}$ have the properties described above. However, depending on the type of bias, it might or might not be computable. If it is computable, then it defines the bootstrapping from the theory space to the bias space when an inconsistency is detected.

The bootstrapping of BEGAT from a new bias to a new theory that has a limiting approximation to the old theory requires two capabilities. First, given the old bias and the new sibling bias, the restriction of the old theory to those instances compatible with the new bias must be defined and computable. Second, given this restriction, its generalization by the new bias must also be defined and computable.

As an example of BEGAT with a different type of bias, consider the problem of learning to predict a person's native language from attributes available in a data base. We will assume that one's native language is the same as the language spoken by one's mother, but that the mother's language is not in the data base. A declarative representation for biases that includes functional dependencies was presented in [Davies and Russell 87] and subsequent work. Let the original bias be that the native language is a function of the birth place. This bias would likely be consistent with data from Europe, but might be inconsistent with the data from the U.S. because of its large immigrant population. Assume that a function $\mathcal{D}$ derives a new bias where the native language is a function of the mother's place of origin. The following limit homomorphism formalizes the intersection of the original bias and the new bias:

$$\lim_{\#immigrants \to 0} \text{mother's} - \text{origin}(x) = \text{birth} - \text{place}(x)$$

The restriction of the original theory to concepts derived from the limiting value (e.g. non-immigrant data) is compatible with this new bias. Furthermore, the concepts learned from this restricted set can be transferred directly to the new theory by substituting the value of the birth place attribute into the value for the mother's place of origin.

Future research will explore the theory and application of primal-dual learning to theoretical physics and other domains. Given the spectacular progress of twentieth century physics, based on the legacy of Einstein's research strategy, the computational advantages of machine learning algorithms using this strategy might be considerable.

Acknowledgments

Thanks to Robert Holte, Laura Jones, Hugh Lowry, Thomas Pressburger, and Jeffrey Van Baalen for their many helpful comments on improving this paper.

References

Aharoni, J. 1965. *The Special Theory of Relativity.* New York: Dover.

Benjamin, P. editor. 1990a. *Change of Representation and Inductive Bias.* Boston: Kluwer.

Benjamin, P. editor. 1990b. *Workshop Proceedings for Algebraic Approaches to Problem Solving and Perception.* June 1990.

Davies, P.C.W. and J. Brown 1988. *Superstrings: A Theory of Everything?* Cambridge University Press.

Davies, T. R. and Russell, S.J. 1987. A Logical Approach to Reasoning by Analogy. In IJCAI-87.

Dietterich, T. 1991. Invited Talk on Machine Learning at AAAI91, Los Angeles, CA.

Einstein, A. 1905. On the Electrodynamics of Moving Bodies. In *The Principle of Relativity, A Collection of Original Memoirs on the Special and General Theory of Relativity,* contributors H.A. Lorentz, A. Einstein, H. Minkowski, and H. Weyl. NewYork: Dover (1952).

Einstein, A. 1916. *Relativity: The Special and General Theory. A clear explanation that anyone can understand.* New York: Crown.

Ellman, T. editor. 1990. *AAAI90 Workshop proceedings on Automatic Generation of Abstractions and Approximations.* Boston, MA.

Ellman, T. editor. 1992. *AAAI902Workshop proceedings on Automatic Generation of Abstractions and Approximations.* San Jose, CA.

Forbus K.D. 1984. Qualitative Process Theory. *Artificial Intelligence*(24):85-168.

French, A. P. 1968. *Special Relativity.* New York: Norton.

Friedman, M. 1983. *Foundations of Space-Time Theories: Relativistic Physics and Philosophy of Science.* New Jersey: Princeton University Press.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Kuipers, B. 1986. Qualitative Simulation. *Artificial Intelligence(29): 289-388.*

Minkowski, H. 1908. Space and Time. In *The Principle of Relativity, A Collection of Original Memoirs on the Special and General Theory of Relativity,* contributors H.A. Lorentz, A. Einstein, H. Minkowski, and H. Weyl. NewYork: Dover (1952).

O'Rorke, P., Morris, S. and D. Schulenburg 1990. Theory Formation by Abduction: A Case Study Based on the Chemical Revolution. In *Computational Models of Scientific Discovery and Theory Formation,* editors J. Shrager and P. Langley.

Shrager, J. and Langley, P. eds. 1990. *Computational Models of Scientific Discovery and Theory Formation,* editors. San Mateo: Morgan Kaufmann.

Skinner, R. 1982. *Relativity for Scientists and Engineers.* New York: Dover.

Smith, D.R. 1982. Derived Preconditions and Their Use in Program Synthesis. In *Sixth Conference on Automated Deduction,* ed. D.W. Loveland, 172-193. Lecture Notes in Computer Science, volume 138. Berlin: Springer-Verlag.

Taylor, E.F. and Wheeler, J.A. 1966. *Spacetime Physics.* San Francisco: Freeman.

Thagard, P. and G. Nowak 1990. The Conceptual Structure of the Geological Revolution. In *Computational Models of Scientific Discovery and Theory Formation,* editors J. Shrager and P. Langley.

Thagard, P. 1992. *Conceptual Refvolutions.* Princeton, New Jersey: Princeton University Press.

Valiant, L.G. 1984. A Theory of the Learnable. In *CACM* (27):1134-1142.

Weld,D.S. 1989. *Automated Model Swithching: Discrepency Driven Selection of Approximation Reformulations.* University of Washington Computer Science Department Technical Report 89-08-01.

Weld, D.S. and de Kleer, J. eds. 1990. *Readings in Qualitative Reasoning about Physical Systems.* editors. San Mateo, CA: Morgan Kaufmann.

Weld, D.S. 1992. *Reasoning about Model Accuracy.* University of Washington Computer Science Department Technical Report To appear in Artificial Intelligence.

Zee, A. 1986. Fearful Symmetry: The Search for Beauty in Modern Physics. New York: Macmillan.