Do First-Class Constraints Generate Gauge Transformations? A Geometric Resolution.

Clara Bradley

Department of Philosophy, University College London

clara.bradley@ucl.ac.uk

Abstract

In the Hamiltonian formalism, a *gauge transformation* is typically defined as a transformation generated by an arbitrary combination of first-class constraints. But gauge transformations are also understood as marking physical equivalence: they relate states or solutions that represent the same physical situation. Whether these two characterizations coincide has been a matter of debate. Pitts (2014*b*), for example, contends that first-class constraints can generate a "bad physical change". This paper defends the standard view, arguing that it correctly identifies both states and solutions that are equivalent from the perspective of the geometric structure of the Hamiltonian formalism. In doing so, it clarifies the relationship between mathematical and interpretational perspectives on gauge transformations.

1 Introduction

Gauge transformations are of philosophical interest because they are often characterized as symmetries that reveal "excess structure" or "redundancy" in a physical theory: they are transformations that relate mathematically distinct but physically equivalent situations.[1] Given their central role in the interpretation of a physical theory, it is crucial that the mathematical definition of gauge transformations aligns with their intended physical interpretation.

There is a longstanding tradition of using the "constrained Hamiltonian formalism" to identify

[1]For more on the notion of excess structure and its connection to symmetries of a theory, see, for example, Ismael & Van Fraassen (2003), Earman (2004), Baker (2010).

the gauge transformations of a theory. The standard account, due to Dirac (1964), identifies the gauge transformations as the transformations generated by an arbitrary combination of first-class constraints, which are the constraints on the dynamically allowed states that have vanishing Poisson bracket with all of the constraints. This definition has significant implications for the formulation of Hamiltonian gauge theories. On its basis, Dirac argued that the Hamiltonian governing the dynamics should be understood not as an equivalence class of Hamiltonians known as the *Extended Hamiltonian*. Furthermore, in contexts where gauge symmetry must be eliminated—most prominently in standard approaches to quantization—this definition justifies treating states related by first-class constraints as equivalent, and therefore for moving to a state space where the differences between such states are removed.

However, there have been several recent dissenters of Dirac's account of gauge transformations. For example, using the case of Electromagnetism, Pitts (2014*b*) argues that a first-class constraint can generate "a bad physical change". Similarly, Pons (2005) argues that Dirac's analysis of gauge transformations is "incomplete" since it does not provide an accurate account of the symmetries between solutions to the equations of motion; it only provides an account of the symmetries between individual states. Both authors conclude that Dirac was wrong about the definition of the gauge transformations in the Hamiltonian formalism and that formulating a theory in terms of the Extended Hamiltonian is therefore unmotivated.[2] If correct, these arguments could have implications for other issues in the foundations of the constrained Hamiltonian formalism. Notably, there is a puzzle called the "Problem of Time" that arises for theories where the Hamiltonian function is itself a first-class constraint: according to the standard definition, states along a single solution are equivalent to each other. If gauge transformations are not given by the standard definition, then this could be an avenue to avoiding the Problem of Time.[3]

In this paper, I defend the orthodoxy against the dissenters. I grant that they are correct in noting that Dirac's original reasoning is flawed, but I reject the claim that his conclusion therefore fails. Instead, I argue that the standard view—that arbitrary combinations of first-class constraints generate gauge transformations—can be independently motivated by examining the geometric structure of the constrained Hamiltonian formalism. In particular, I show that one can distinguish two precise notions of equivalence within this framework, and I demonstrate that the standard definition aligns with both,

[2]The alternative definition of a gauge transformation supported by Pitts and Pons can also be found elsewhere, including in earlier work by Anderson & Bergmann (1951) and Castellani (1982).

[3]See Pitts (2014*a*) for a response of this kind. For an introduction to the Problem of Time and its philosophical implications, see Thébault (2021).

thereby connecting the mathematical formalism to its interpretational significance.

There are several important consequences. First, the argument extends a recent response to Pitts (2014*b*) provided by Pooley & Wallace (2022). They demonstrate that Dirac's orthodoxy can be upheld in the case of Electromagnetism by showing that the standard definition is correct, provided one regards the Extended Hamiltonian as the appropriate equivalence class of Hamiltonians. However, they leave open the question of *why* one ought to formulate a theory in terms of the Extended Hamiltonian. The present argument answers this question: the equivalence class of solutions determined by the Extended Hamiltonian emerges naturally from the mathematical structure of the formalism.

Second, the analysis sharpens the distinction between viewing gauge transformations as capturing a notion of equivalence between *states* versus *solutions*. On Dirac's account, gauge transformations relate individual states along equivalent solutions, and so these two notions of equivalence are intertwined. Here, I show that these notions can be separated: one can identify a sense in which states are equivalent that is conceptually independent of the equivalence between solutions to the equations of motion. This may have implications for the Problem of Time, where a central question is whether gauge transformations should be interpreted differently when applied to states than when applied to solutions.

Finally, the argument brings into focus two questions that have not always been cleanly distinguished in the literature. The first is whether the standard definition of gauge transformations is justified from within the Hamiltonian formalism itself. The second is whether the Hamiltonian account yields the *same* account of gauge transformations as that arising in the Lagrangian formalism. This paper answers the first question affirmatively. The second, however, turns on the deeper issue of how different theories, and their associated symmetries, are to be compared.

The paper will go as follows. In Section 2, I present Dirac's version of the constrained Hamiltonian formalism and the original reasoning that led to the conclusion that arbitrary combinations of first-class constraints generate gauge transformations. In Section 3, I spell out the example that pitts uses as a counterexample to Dirac's account of gauge transformations, and I discuss where the tension between the two views lies. In Section 4, I consider the extent to which Pooley & Wallace (2022) provide a resolution, and in Sections 5 and 6, I extend this argument by showing that the geometric formulation of the constrained Hamiltonian formalism provides the theoretical ground for motivating the view that arbitrary combinations of first-class constraints generate gauge transformations. Finally, in Section 7, I consider and respond to two possible counterarguments.

2 Dirac's Theory

Dirac's version of the constrained Hamiltonian formalism is constructed by starting with the Lagrangian formalism. For the sake of simplicity, we will focus on the construction in the finite-dimensional case, but it can be extended naturally to the infinite-dimensional case.[4] The Lagrangian formalism has a state space composed of N degrees of freedom $q_n, n = 1, ..., N$, with corresponding velocities $\frac{dq_n}{dt} = \dot{q}_n$, where we assume an independent time variable t.[5] The dynamics are given by specifying a Lagrangian $L = L(q_n, \dot{q}_n)$ with corresponding action $I = \int L(q_n, \dot{q}_n)dt$, from which one derives the equations of motion called the Euler-Lagrange equations:

$$\frac{d}{dt}\frac{\partial L(q_n, \dot{q}_n)}{\partial \dot{q}_n} = \frac{\partial L(q_n, \dot{q}_n)}{\partial q_n}$$

To move to the Hamiltonian framework, one introduces 'canonical momentum variables' $p_n = \frac{\partial L}{\partial \dot{q}_n}$. When these momenta are not independent of each other, there are constraints of the form $\phi_m(q_n, p_n) \approx 0$ for $m = 1, ..., M$ where M is the number of constraints and the equality is *weak* equality, indicating that the constraints only hold on a subspace of phase space (the state space given by the collection of points (q_n, p_n)). Constraints of this kind are called the *primary constraints*.

The 'Hamiltonian' $H(q_n, p_n)$ can be defined via $FL^*(H) = E_L$ where E_L is the energy function associated with the Lagrangian, defined as $E_L(q_n, \dot{q}_n) = FL(\dot{q}_n)\dot{q}_n - L(q_n, \dot{q}_n)$ and FL is the Legendre transformation that takes the point $(q_n, \dot{q}_n)$ to $(q_n, \frac{\partial L}{\partial \dot{q}_n})$. This only unambiguously defines the Hamiltonian at the points (q_n, p_n) where the primary constraints hold; at all other points, the Hamiltonian is defined up to arbitrary combinations of the primary constraints. We call the equivalence class of Hamiltonians up to arbitrary combinations of the primary constraints the *Total Hamiltonian*, $H_T = H + u^m\phi_m$ where u^m are arbitrary functions of the canonical variables. From the variation in H_T, one can derive Hamilton's equations of motion with constraints:

$$\dot{q}_n = \frac{\partial H}{\partial p_n} + u^m\frac{\partial \phi_m}{\partial p_n}$$

$$\dot{p}_n = -\frac{\partial H}{\partial q_n} - u^m\frac{\partial \phi_m}{\partial q_n}$$

More generally, for any dynamical variable g, $\dot{g} \approx \{g, H\} + u^m\{g, \phi_m\} = \{g, H_T\}$ where $\{\}$ is the

[4] See Henneaux & Teitelboim (1994) for more details.

[5] In order to consider the Problem of Time, it is useful to drop this assumption and treat the time variable as an additional dynamical variable, but we keep this assumption for the purposes here.

Poisson bracket, defined by $\{f, g\} = \frac{df}{dq^n}\frac{dg}{dp_n} - \frac{df}{dp_n}\frac{dg}{dq^n}$.

In order for the solutions to the equations of motion to be consistent with the primary constraints, in the sense that the primary constraints hold at all times along a solution to the equations of motion, it ought to be the case that $\dot{\phi}_m \approx 0$. In other words, it ought to be the case that $\{\phi_m, H\} + u^p\{\phi_m, \phi_p\} \approx 0$ where $p = 0, ..., m$. For each m, this equation either is identically satisfied with the primary constraints, reduces to an equation independent of the u's of the form $\chi_k(q_n, p_n) \approx 0$, or it imposes conditions on the u's.

In the second case, we say that $\chi_k(q_n, p_n) \approx 0$ are *secondary constraints*, since they arise from applying the equations of motion to the primary constraints. If we have a secondary constraint, then we get new consistency conditions by requiring $\dot{\chi}_k \approx 0$, which is again one of the three kinds above. It may be that this procedure ends up showing that the system is dynamically inconsistent; when it doesn't, one can show that the process of determining (secondary) constraints will terminate and that one will be left with only the consistency conditions of the third kind. We can combine the primary and secondary constraints, writing them as $\phi_j \approx 0$ for $j = 1, ..., M + K$ where K is the number of secondary constraints.

For the remaining consistency conditions that do not reduce, we find solutions $u^m = U^m + v^a V_a^m$ where v^a is arbitrary and $V^m\{\phi_j, \phi_m\} \approx 0$. Substituting into the Total Hamiltonian, we get

$$H_T = H' + v^a\varphi_a$$

where $H' = H + U^m\phi_m$ and $\varphi_a = V_a^m\phi_m$. Notice that we have satisfied all the consistency conditions but still have coefficients v^a that are arbitrary functions of the canonical variables.

A dynamical variable $R(q_n, p_n)$ is said to be *first-class* if $\{R, \phi_j\} \approx 0$. In other words, a dynamical variable is first-class if the Poisson bracket with any constraint equals a linear function of the constraints. If it is not first-class, it is called *second-class*. Importantly, H' and φ_a are first-class. This means that H_T is an equivalence class of Hamiltonians given by a sum of a first-class Hamiltonian and an arbitrary combination of primary, first-class constraints.

Given some initial state $(q_n(t_0), p_n(t_0))$, the q's and p's at later times are underdetermined because of the arbitrariness in the coefficients v^a. There is therefore a form of indeterminism in the theory: there are multiple possible evolutions from an initial state. However, we might think that this indeterminism is an artifact of our mathematical description; it indicates that our theory contains redundant terms, rather than a mark of real indeterminism in the world. It is this reasoning that led Dirac to

propose the following definition of a gauge transformation:

> **State Gauge Transformation**: A gauge transformation relates any two states that are possible evolutions from an initial state under the dynamics generated by the Total Hamiltonian at some fixed (infinitesimal) interval δt.

In other words, Dirac proposes that physically equivalent states are precisely those that result from the arbitrariness in v^a in evolving the state of a system.

We can determine these transformations in the following way. For a given dynamical variable g with initial value g_0, its value after some infinitesimal δt under a specified choice of coefficients v^a is:

$$g(\delta t) = g_0 + \dot{g}\delta t = g_0 + \{g, H_T\}\delta t = g_0 + \delta t[\{g, H'\} + v^a\{g, \varphi_a\}] \tag{1}$$

However, one could have made different choices for v^a. Call another set of choices v'^a. The difference between the two values for g at δt under these two choices of coefficients is given by:

$$\Delta g(\delta t) = \delta t(v^a - v'^a)\{g, \varphi_a\} = \varepsilon^a\{g, \varphi_a\} \tag{2}$$

where ε^a is an arbitrary small number. This change will, according to the account above, describe the same physical state: it corresponds to a change from one state to another that arises merely from a different choice of arbitrary coefficient in the evolution from some initial state. Since φ_a are just the primary first-class constraints, Dirac concludes:

> All primary first-class constraints generate gauge transformations.

However, this isn't the end of the story. Take some value for $g(\delta t)$ and transform it by $\varepsilon^a\{g, \varphi_a\}$ *twice*. This new value for $g(\delta t)$ is related to the previous value by some amount generated by $\{\varphi_a, \varphi_{a'}\}$. The ϕ_a's are first-class constraints, and the Poisson bracket of two first-class quantities is first-class, so this generating function is a first-class constraint. However, it need not be a primary first-class constraint; it could be a *secondary* first-class constraint. Observing this, Dirac presents the following conjecture:

> **Dirac Conjecture**: All secondary first-class constraints generate gauge transformations.

If this conjecture is true, then one can conclude:

Arbitrary combinations of first-class constraints generate a State Gauge Transformation.

However, if we accept this conclusion, we have a puzzle. On the one hand, the dynamics are generated by the Total Hamiltonian, which includes the arbitrariness associated with the primary first-class constraints. On the other hand, there is arbitrariness associated with both the primary and secondary first-class constraints through the definition of the state gauge transformations. This mismatch between the dynamics and the state gauge transformations led Dirac to suggest that one should also add the first-class *secondary* constraints to the Total Hamiltonian, giving rise to the *Extended Hamiltonian*, $H_E = H_T + w^b\chi_b$ where χ_b are the first-class secondary constraints and w^b are arbitrary functions of the canonical variables. The equations of motion then read: $\dot{g} = \{g, H_E\}$.

The final picture of Dirac's theory is:

1. The symmetries that characterize physical equivalence are given by the "State Gauge Transformations", which are generated by arbitrary combinations of first-class constraints.

2. The dynamics are generated by an equivalence class of Hamiltonians represented by the Extended Hamiltonian.

Whether this picture is correct will be the subject of the rest of the paper.

3 The Case of Electromagnetism

Although Dirac's account of the gauge transformations in the constrained Hamiltonian formalism has been widely accepted as the standard framework, there are recent arguments that Dirac's account is flawed.[6] Here, I focus on the argument provided by Pitts (2014*b*) that contends that classical Electromagnetism is a counterexample to Dirac's account.

The Lagrangian for classical Electromagnetism can be written relative to a given Lorentz frame as

$$\mathcal{L}(\vec{A}, V; \dot{\vec{A}}, \dot{V}) = \int \frac{1}{2}(\dot{\vec{A}} - \nabla V)^2 - \frac{1}{2}(\nabla \times \vec{A})^2 - (V\rho + \vec{A} \cdot \vec{J})$$

where $\vec{A}$ and V are time-dependent functions on $\mathbb{R}^3$ and the integral is over $\mathbb{R}^3$. The conjugate momenta are $p_{\vec{A}} = \frac{\delta L}{\delta \dot{\vec{A}}} = \dot{\vec{A}} - \nabla V$ and $p_V = \frac{\delta L}{\delta \dot{V}} = 0$. This means that there is one primary constraint, $\phi_0 = p_V$. The Total Hamiltonian is:

[6] See in particular Pitts (2014*a*,*b*) and Pons (2005) but also Pons et al. (1997) and Barbour & Foster (2008).

$$H_T = \int \frac{1}{2}(p_{\vec{A}}^2 + \vec{B}^2) + \lambda p_V + p_{\vec{A}} \cdot \nabla V + (V\rho + \vec{A} \cdot \vec{J}) \tag{3}$$

where the integral is over $\mathbb{R}^3$ and λ is an arbitrary function of the canonical coordinates. Integrating by parts with appropriate boundary conditions, we can rewrite the Total Hamiltonian as:

$$H_T = \int \frac{1}{2}(p_{\vec{A}}^2 + \vec{B}^2) + \vec{A} \cdot \vec{J} + \lambda p_V - V(\nabla \cdot p_{\vec{A}} - \rho) \tag{4}$$

We can then find the evolution of the primary constraint:

$$\{p_V, H_T\} = \frac{\delta H}{\delta V} = \nabla \cdot p_{\vec{A}} - \rho. \tag{5}$$

So there is a secondary constraint given by $\phi_1 = \nabla \cdot p_{\vec{A}} - \rho$. The evolution of the secondary constraint is zero, so there are two constraints in total, and both constraints are first-class.

The equations of motion for $\vec{A}$ and V are given by:[7]

$$\begin{aligned} \frac{\partial \vec{A}}{\partial t} &= \{\vec{A}, H_T\} = \frac{\partial H_T}{\partial p_{\vec{A}}} = p_{\vec{A}} + \nabla V \\ \frac{\partial V}{\partial t} &= \{V, H_T\} = \frac{\partial H_T}{\partial p_V} = \lambda \end{aligned} \tag{6}$$

The question that Pitts (2014*b*) asks is whether the arbitrary combinations of the primary and secondary constraint generate gauge transformations for these equations. In other words, we want to know whether, if $(\vec{A}(t), V(t); p_{\vec{A}}(t), p_V(t))$ satisfies these equations of motion, then transforming this solution by an arbitrary combination of the first-class constraints, $\int \alpha\phi_0 + \beta\phi_1$, also satisfies the equations of motion, where α and β are arbitrary functions of the canonical coordinates and time.

We have that:

$$\begin{aligned} \{\vec{A}, \int \alpha\phi_0 + \beta\phi_1\} &= \{\vec{A}, \int \alpha p_V + \beta(\nabla \cdot p_{\vec{A}} - \rho)\} \\ &= \{\vec{A}, \int \alpha p_V\} + \{\vec{A}, \int \beta(\nabla \cdot p_{\vec{A}} - \rho)\} \end{aligned} \tag{7}$$

The first term vanishes. Since $\int \beta \nabla \cdot p_{\vec{A}} = -\int p_{\vec{A}} \cdot \nabla\beta$ by integration by parts (with appropriate

[7] We leave out the equations of motion for $p_{\vec{A}}$ and p_V for convenience, since they aren't important for the argument.

boundary conditions), the second term is equal to $\{\vec{A}, -\int p_{\vec{A}} \cdot \nabla\beta + \beta\rho)\} = \nabla\beta$. Therefore, the transformed quantity is given by $A' = A + \nabla\beta$.

Similarly:

$$\{V, \int \alpha\phi_0 + \beta\phi_1\} = \{V, \int \alpha p_V\} + \{V, \int \beta(\nabla \cdot p_{\vec{A}} - \rho)\} \tag{8}$$

The second term here vanishes, and the first term is equal to α. Thus, the transformed potential is given by $V' = V + \alpha$.

We also have that $\{p_{\vec{A}}, \int \alpha p_V + \beta(\nabla \cdot p_{\vec{A}} - \rho)\} = \{p_V, \int \alpha p_V + \beta(\nabla \cdot p_{\vec{A}} - \rho)\} = 0$ and so the conjugate momenta do not change under the transformation generated by an arbitrary combination of the constraints. We can therefore write the transformed equations of motion for $\vec{A}$ and V as:

$$\begin{aligned} \frac{\partial \vec{A}'}{\partial t} &= \frac{\partial \vec{A}}{\partial t} + \frac{\partial \nabla\beta}{\partial t} = p_{\vec{A}} + \nabla(V + \alpha) \\ \frac{\partial V'}{\partial t} &= \frac{\partial V}{\partial t} + \frac{\partial \alpha}{\partial t} = \lambda \end{aligned} \tag{9}$$

Since we assumed that $\frac{\partial \vec{A}}{\partial t} = p_{\vec{A}} + \nabla V$, the first equation is satisfied only when $\frac{\partial \nabla\beta}{\partial t} - \nabla\alpha = 0$. In particular, in the case where either α or β is zero (where one considers the transformation generated by only one of the primary or secondary constraints), the first equation is not satisfied.

On the basis of this argument, Pitts (2014*b*) concludes that arbitrary combinations of first-class constraints do not generate gauge transformations; only particular combinations of first-class constraints generate gauge transformations. And since the form of the gauge transformations was the basis for introducing the Extended Hamiltonian, one ought to also conclude that the Extended Hamiltonian is not motivated. Indeed, notice that since $\nabla\alpha = \frac{\partial \nabla\beta}{\partial t}$, we only need one arbitrary function (and its time derivative) to specify the gauge transformations; not, as the Extended Hamiltonian implies, as many arbitrary functions as there are first-class constraints. Therefore, this argument suggests that the Extended Hamiltonian is not the right equivalence class of Hamiltonians from the perspective of capturing the arbitrariness in the dynamics of a gauge theory.

3.1 Where The Issue Lies

There is an immediate sense in which the above argument fails on its own to show that Dirac was wrong. As we discussed in Section 2, Dirac provides an account of the "State Gauge Transformations":

transformations relating two states that are possible evolutions from some initial state. However, the argument we just ran, following Pitts (2014*b*), doesn't consider whether two *states* are equivalent; it considers whether two *solutions* are equivalent. That is, it considers whether arbitrary combinations of first-class constraints generate a transformation that takes one from a solution to the equations of motion to another solution. We might alternatively call this notion of a gauge transformation a "Solution Gauge Transformation":

> **Solution Gauge Transformation**: A gauge transformation relates any two *curves* that are possible evolutions from an initial state under the dynamics generated by the Total Hamiltonian.

What Pitts' argument demonstrates is that the Solution Gauge Transformations are not generated by arbitrary combinations of first-class constraints in the context of classical Electromagnetism. Indeed, arbitrary combinations of first-class constraints *do* generate State Gauge Transformations in classical Electromagnetism. To see this, recall that we can write the Solution Gauge Transformations as $\int \dot{\epsilon}\phi_0 + \epsilon\phi_1$. At a fixed time, ϵ and $\dot{\epsilon}$ become independent of each other. And so, we can write the State Gauge Transformations as $\int \alpha\phi_0 + \beta\phi_1$, as would be the case if arbitrary combinations of first-class constraints generate gauge transformations. So what Pitts (2014*b*) shows is that Solution Gauge Transformations do not always match the State Gauge Transformations.

In light of this, one might think that what this shows is that we really have two distinct notions of a gauge transformation, 'State Gauge Transformation' and 'Solution Gauge Transformation', and it turns out that these notions do not coincide. This would suggest that there is not really a debate here at all; different parties in the debate are just focusing on different concepts, and we can accept that both are right.

Although this would be unproblematic if gauge transformations were purely a formal notion, there remains an issue if gauge transformations are taken to mark physical equivalence. The reason is that accepting that both notions of a gauge transformation are adequate means accepting that the individual states along two curves can be physically equivalent without it being the case that the curves that they make up are physically equivalent, since the transformations that generate Solution Gauge Transformations are more restrictive than those that generate State Gauge Transformations. Conceptually, this is not coherent: solutions just consist of a series of states, and so if all of these states are physically equivalent to some other series of states, then the solutions ought to also be physically equivalent.

Therefore, if one wants to accept that "Solution Gauge Transformation" is the right account of equivalence between solutions, then one must also accept that "State Gauge Transformation" fails to independently capture the notion of equivalence between states. Rather, a state gauge transformation would have to be understood as derivative to the solution gauge transformations: a state gauge transformation is simply the action of a solution gauge transformation at a single point along the solution. If this is correct, then one cannot conclude that any two states related by a transformation generated by an arbitrary combination of first-class constraints are physically equivalent; one would only be able to conclude that they correspond to two states along equivalent solutions. This would be a significant position to hold, given that the standard Hamiltonian picture of gauge theories—including the process of quantization—relies on treating individual states related by a gauge transformation as being equivalent.

This sets the stage for the rest of the paper. I will argue that one can maintain independent notions of state and solution gauge transformations as notions of equivalence, but it means that one has to deny that "State Gauge Transformation" and "Solution Gauge Transformation" provide the correct characterizations of gauge transformations on states and solutions, respectively. In particular, a common part of the definition of a "State Gauge Transformation" and "Solution Gauge Transformation" is the assumption that state or solution notions of equivalence are derived from the equivalence between Hamiltonian functions, which in turn is provided by the *Total Hamiltonian*. I will argue that this assumption is misplaced: the Total Hamiltonian draws distinctions that a Hamiltonian gauge theory, properly understood, is unable to draw. Rather, the distinctions that a Hamiltonian gauge theory can draw match those distinctions provided by the *Extended Hamiltonian*, which in turn is aligned with the standard definition of the gauge transformations. The last part of this argument parallels the recent response to Pitts (2014*b*) developed by Pooley & Wallace (2022). To clarify how my argument goes beyond that of Pooley & Wallace (2022), I will first outline their position and then indicate where it falls short of resolving the issue.

4 Extended Hamiltonian Electromagnetism

Pooley & Wallace (2022) show that if one starts with the Extended Hamiltonian, arbitrary combinations of first-class constraints generate gauge transformations between solutions for classical Electromagnetism. Their argument can be summarised as follows. Consider the Extended Hamiltonian

for classical Electromagnetism, where we add to the Total Hamiltonian the secondary constraint multiplied by an arbitrary function μ:

$$H_E = \int \frac{1}{2}(p_{\vec{A}}^2 + \vec{B}^2) + \vec{A}\cdot\vec{J} + \lambda p_V - (V+\mu)(\nabla \cdot p_{\vec{A}} - \rho) \tag{10}$$

With this Hamiltonian, the equations of motion become:

$$\begin{aligned} \frac{\partial \vec{A}}{\partial t} &= \frac{\partial H_E}{\partial p_{\vec{A}}} = p_{\vec{A}} + \nabla(V+\mu) \\ \frac{\partial V}{\partial t} &= \frac{\partial H_E}{\partial p_V} = \lambda \end{aligned} \tag{11}$$

When we now consider the transformation generated by an arbitrary combination of primary and secondary constraints, $\int \alpha\phi_0 + \beta\phi_1$, we find:

$$\begin{aligned} \frac{\partial \vec{A}'}{\partial t} &= \frac{\partial \vec{A}}{\partial t} + \frac{\partial \nabla\beta}{\partial t} = p_{\vec{A}} + \nabla(V+\mu+\alpha) \\ \frac{\partial V'}{\partial t} &= \frac{\partial V}{\partial t} + \frac{\partial \alpha}{\partial t} = \lambda \end{aligned} \tag{12}$$

We can rewrite the first equation as $\frac{\partial \vec{A}'}{\partial t} = \frac{\partial \vec{A}}{\partial t} = p_{\vec{A}} + \nabla(V+\mu+\alpha-\dot{\beta})$. Notice that μ, α and $\dot{\beta}$ are all arbitrary functions, so we can write this equation as

$$\frac{\partial \vec{A}'}{\partial t} = \frac{\partial \vec{A}}{\partial t} = p_{\vec{A}} + \nabla(V+\mu')$$

where μ' is arbitrary. This is just the untransformed equation of motion, with μ' in place of μ. In other words, if $(\vec{A}(t), V(t); p_{\vec{A}}(t), p_V(t))$ is a solution to $\frac{\partial \vec{A}}{\partial t} = p_{\vec{A}} + \nabla(V+\mu)$, then $(\vec{A}(t)+\nabla\beta, V(t)+\alpha; p_{\vec{A}}(t), p_V(t))$ is also a solution. Therefore, arbitrary combinations of first-class constraints generate gauge transformations on solutions, for the dynamics generated by the Extended Hamiltonian.

Although this argument shows that when we start with the Extended Hamiltonian, the gauge transformations are generated by arbitrary combinations of first-class constraints, it leaves open the question of what the justification is for starting with the Extended Hamiltonian. Indeed, it seems that the proponents of "Solution Gauge Transformation" will deny that this is the right starting point; they would say that it is the Total Hamiltonian that one should use to determine the gauge transformations.

Pooley & Wallace (2022) do provide one kind of response: the dynamics generated by the Extended Hamiltonian is *empirically equivalent* to the dynamics generated by the Total Hamiltonian, in the sense that they give rise to the same predictions for those quantities that both agree are gauge-invariant. In particular, what they notice is that the difference between the solutions of the Total and Extended Hamiltonian lies only in the quantity that plays the role of the electric field: when the Total Hamiltonian is used to generate the dynamics, it is $\dot{\vec{A}} - \nabla V$ that plays the role of the electric field, but when the Extended Hamiltonian is used, it is $p_{\vec{A}}$. And so, given that it is the electric field that one measures through its interaction with charges (and not the quantities $\dot{\vec{A}}$ and V), there is no empirical difference between these choices of Hamiltonian.

Although I consider this response to be both convincing and informative, I will argue that we can go further: the Extended Hamiltonian does not just capture the same empirical content as that of the Total Hamiltonian, it is the correct equivalence class of Hamiltonians from the perspective of the mathematical structure of the theory. That is, once we formalize the mathematical structure of a Hamiltonian gauge theory, the fact that the Extended Hamiltonian generates the equivalence class of solutions to the equations of motion can be motivated directly.

To develop this argument, we will employ the standard geometric formulation of the constrained Hamiltonian formalism, as it offers a formal framework for clarifying the issues at hand. In particular, the geometric framework highlights the role of first-class constraints within the structure of the formalism. This, in turn, makes it possible to see more clearly the theoretical motivations underlying specific definitions of state and solution gauge transformations.

5 Geometric Reformulation

The constrained Hamiltonian formalism can be expressed naturally in a geometric way using the theory of symplectic manifolds.[8] A symplectic manifold consists of a pair (M, ω) where M is a smooth manifold and ω is a *symplectic form*: it is a two-form (a smooth, anti-symmetric tensor field of rank (0,2)), that satisfies the following conditions:

1. ω is *non-degenerate*, i.e. if $\omega(X_i, X_j) = 0$ for all $X_j \in TM$ and some $X_i \in TM$, then $X_i = 0$.

2. ω is *closed*, i.e., $\mathbf{d}\omega = \mathbf{0}$, where $\mathbf{d}$ is the exterior derivative operator, which is such that $\mathbf{d}f = df$, the differential of a function f, $\mathbf{d}(\mathbf{d}\alpha) = 0$ where α is a k-form, and $\mathbf{d}(f\alpha) = df \wedge \alpha + f\mathbf{d}\alpha$.

[8] This formalism is widely used to express the constrained Hamiltonian formalism. For further details of this formalism, see Henneaux & Teitelboim (1994), Butterfield (2006).

There is a sense in which every symplectic manifold comes equipped with the Poisson bracket that we defined (in coordinate-dependent form) in Section 2: Let (M, ω) be a symplectic manifold and $C^\infty(M)$ the space of smooth maps on M. In addition, let ω' be the inverse of ω (a smooth, anti-symmetric tensor field of rank $(2,0)$).[9] Then the map $\{\cdot,\cdot\} : C^\infty(M) \times C^\infty(M) \to C^\infty(M)$ defined by $f, g \mapsto \{f, g\} = \omega'(df)(dg)$ is the Poisson bracket on M.

A constrained Hamiltonian theory can be defined as a symplectic manifold in the following way. The manifold is the cotangent bundle of configuration space (otherwise known as phase space), T^*Q, whose points can be written as $\{(q_n, p_n), n = 1, ..., N\}$. T^*Q comes equipped with a one-form, the *Poincaré one-form*, given by $\theta = p_i dq^i$. The corresponding two-form is given by $\omega = \mathbf{d}\theta = dp_b \wedge dq^b$, which is symplectic.

Given a function f, one can uniquely define a smooth tangent vector field X_f through:

$$\omega(X_f, \cdot) = \mathbf{d}f \tag{13}$$

where $\{\cdot\}$ represents any vector field tangent to T^*Q. In particular, one can uniquely define a vector field corresponding to the Hamiltonian $H = p^i q_i - L$ through $\omega(X_H, \cdot) = dH$. This provides an alternative way to write Hamilton's equations. In particular, $\{f, H\} = \omega(X_f, X_H) = df(X_H) = \mathcal{L}_{X_H}(f)$. If we define the flow parameter of X_H to be time, then this says that $\{f, H\} = \frac{df}{dt}$, which is Hamilton's equation.

We can understand the primary constraints $\phi_m(q_n, p_n) = 0$ for $j = 1, .., M$ where M is the total number of constraints as giving rise to a smooth, embedded sub-manifold of phase space of dimension $N - M$, which we call the *primary constraint surface*, given by $\Sigma_p = \{(q_n, p_n) \in \Gamma | \forall_m : \phi_m(q_n, p_n) = 0\}$. The first-class primary constraints are those constraints whose associated vector field is tangent to Σ_p, while the second-class primary constraints are those constraints whose associated vector field is not tangent to Σ_p. For the purposes here, we will restrict ourselves to the case where we just have first-class constraints, since these are the relevant ones for defining the gauge transformations.

We can define an induced two-form on the primary constraint surface $\tilde{\omega}_p$ as the pullback along the embedding $i : \Sigma_p \to \Gamma$ of ω. This induced two-form is in general *degenerate* i.e. it is not invertible. In particular, it possesses M linearly independent null vector fields that form the null space of $\tilde{\omega}_p$. These are the vector fields that satisfy $\tilde{\omega}(X_m, \cdot) = 0$ where $\{\cdot\}$ is any vector field tangent to Σ_p. But these are precisely the vector fields that off the constraint surface satisfy $\omega(X_m, \cdot) = d\phi_m$ where ϕ_m are the

[9] This is well-defined because ω is non-degenerate.

primary first-class constraints, since $d\phi_m|_{\Sigma_p} = 0$. Thus, we will write X_{ϕ_m} for these null vector fields to indicate that they are the tangent vector fields associated with the primary first-class constraints. The degeneracy of $\tilde{\omega}_p$ means that one cannot associate a unique vector field with any smooth function on the constraint surface through the equation $\tilde{\omega}_p(X_f, \cdot) = \mathbf{d}f$, since *if* X_f satisfies this equation (if it is tangent to the primary constraint surface), then so does $X_f + X_{\phi_m}$ since the two-form acts linearly.

We can write the equations of motion on the primary constraint surface as $\tilde{\omega}_p(X_H, \cdot) = dH|_{\Sigma_p}$. However, this equation of motion may not have solutions everywhere, since X_H may not be tangent to the primary constraint surface. In order for the solutions to be tangent to the primary constraint surface, it must be that $\tilde{\omega}_p(X_H, X_{\phi_m}) = dH(X_{\phi_m}) = 0$. This is geometrically what gives rise to the *secondary constraints*, and we can think of these secondary constraints as leading to the specification of a further submanifold.

Continuing this process of requiring the solutions to be tangent to the constraint surface terminates in some *final constraint surface* Σ_f, defined by the satisfaction of a collection of constraints $\varphi_j(q_n, p_n) = 0$ for $j = 1, ..., J$ where J is the total number of constraints. We can also define an induced two-form on Σ_f, $\tilde{\omega}_f$, whose null vector fields are the vector fields associated with all of the first-class constraints, which we will write as X_{φ_j} (since we are just considering the case where all the constraints are first-class, although it is easy to extend to the case where there are second-class constraints). The equations of motion are $\tilde{\omega}_f(X_H, \cdot) = dH|_{\Sigma_f}$, which has (non-unique) solutions everywhere on Σ_f.

The integral curves of the null vector fields on the final constraint surface are known as *gauge orbits*. Equivalently, a gauge orbit is the set of points connected by a curve whose tangent vectors are null. The gauge orbits on the final constraint surface coincide with the notion of a gauge transformation in the Dirac formalism in the following sense: the null vector fields generating the orbits are precisely the vector fields X_{φ_j} associated with the first-class constraints. Arbitrary combinations of first-class constraints thus generate transformations that move points along a given gauge orbit. In this way, the geometric formulation inherits the standard picture of gauge transformations. As emphasized in the canonical textbook on the constrained Hamiltonian formalism:

> "The identification of the gauge orbits with the null surfaces of the induced two-form relies strongly on the postulate made throughout the book that all first-class constraints generate gauge transformations." (Henneaux & Teitelboim (1994, p.54))

Here, however, our aim is to derive the gauge transformations directly from the geometric formal-

ism. We will therefore not assume that the null surfaces of the induced two-form automatically identify physically equivalent states. Accordingly, by "gauge orbits" we will mean only the integral curves of the null vector fields, without presupposing that these curves correspond to gauge equivalence.

6 A Geometric Resolution

The above presentation of the geometric formulation of the constrained Hamiltonian approach shows that it is natural to formulate the theory on the final constraint surface: it captures the dynamically accessible points of phase space, such that solutions to the equations of motion are well-defined at every point. We now consider whether, by formulating the theory on the final constraint surface, one can resolve the issue raised earlier, namely, how to reconcile the gauge transformations on states with the gauge transformations on solutions. Recall that both Dirac (1964) and Pitts (2014*b*) take gauge transformations to be determined through the dynamics generated by the Total Hamiltonian, but this leads to different mathematical transformations being counted as gauge transformations for states and for solutions, and consequently different opinions about whether one should extend the equivalence class of Hamiltonians. We can summarize the reasoning common to Dirac (1964) and Pitts (2014*b*) as follows:

1. First, one determines the gauge transformations via the solutions to the Total Hamiltonian.

2. Then, one uses the gauge transformations to say whether one should extend the equivalence class of Hamiltonians or not.

I will show that this reasoning is flawed in three parts. First, I show that the Extended Hamiltonian—or more precisely, the generator of the solutions to the equations of motion—is motivated independently from consideration of the gauge transformations, and so (2) is wrong: the gauge transformations do not determine the equivalence class of Hamiltonians. Second, I show that the gauge transformations on states can be defined as structure-preserving maps on the final constraint surface, rather than as transformations between points along solutions, and so (1) is wrong: the gauge transformations on states are not simply a special case of the gauge transformations on solutions. Finally, I show that the gauge transformations on solutions are generated by arbitrary combinations of first-class constraints, and I discuss how they are related to the gauge transformations on states.

6.1 Motivating the Extended Hamiltonian

First, let us start with why the Extended Hamiltonian is motivated if one takes the theory to be formulated on the final constraint surface. There is an immediate sense in which the Extended Hamiltonian is the right equivalence class of Hamiltonians from the perspective of the final constraint surface: distinct Hamiltonians on T^*Q that are related by an arbitrary combination of first-class constraints are identified when one restricts to the points on the final constraint surface. However, this requires us to refer back to the theory on the full phase space; what we also want is an intrinsic characterization of the equivalence class of solutions. This is captured by the fact that on the final constraint surface, the vector fields corresponding to solutions to the equations of motion for some Hamiltonian are defined up to arbitrary combinations of vector fields associated with the first-class constraints. In other words, what we have is not an equivalence class of Hamiltonian functions but rather an equivalence class of vector fields associated with the Hamiltonian.

In more detail, take a (first-class) Hamiltonian vector field X_H and transform it to $X_H + a^j X_{\varphi_j}$ where X_{φ_j} are the null vector fields associated with the first-class constraints φ_j and a^j are arbitrary functions. We have that

$$\tilde{\omega}_f(X_H + a^j X_{\varphi_j}, \cdot) = \tilde{\omega}_f(X_H, \cdot) = dH|_{\Sigma_f}$$

since X_{φ_j} are null vector fields. But this means that transforming X_H by an arbitrary combination of the vector fields associated with the first-class constraints preserves the dynamical equations on the final constraint surface. In other words, the structure of the final constraint surface is such that the evolution generated by X_H and that generated by $X_H + a^j X_{\varphi_j}$ is not distinguished: if f satisfies $\tilde{\omega}_f(X_f, X_H) = \frac{df}{dt}|_{\Sigma_f}$, then it satisfies $\tilde{\omega}_f(X_f, X_H + a^j X_{\varphi_j}) = \frac{df}{dt}|_{\Sigma_f}$. Therefore, we can think of the vector fields $X_H + a^j X_{\varphi_j}$ on the final constraint surface as characterizing the equivalence class of vector fields that generate solutions to the equations of motion; they cannot be distinguished by the structure of the theory. Let us call this equivalence class of vector fields the "Extended Hamiltonian vector fields".

Notice that in such reasoning, we have not made any assumptions about the X_{φ_j} being associated with primary or secondary first-class constraints, nor about what the gauge-transformations are; each first-class constraint constitutes a null direction on the final constraint surface, and it is this property that is important in determining which Hamiltonian vector fields are equivalent. This provides one

argument for why restricting to the Total Hamiltonian is unnatural in the geometric framework: it is to privilege a class of null vector fields (those that correspond to primary first-class constraints) as generating equivalent solutions, despite the fact that the equations of motion on the final constraint surface provide no way to distinguish this class of null vector fields from those corresponding to the secondary first-class constraints.

6.2 State Gauge Transformations

Second, let us consider the notion of a gauge transformation on states. We want this notion to identify those points on the final constraint surface that are physically equivalent to each other. In Dirac's account, this is understood in terms of transformations connecting individual points that lie along equivalent solutions. If we follow the same reasoning, but took the solutions to be generated by the Extended Hamiltonian vector fields, we would end up with the same conclusion as Dirac: arbitrary combinations of first-class constraints generate gauge transformations. However, this would still have the issue of being derivative, conceptually, of the transformations linking equivalent solutions. What we want is an independent way of capturing whether two points on the state space represent the same (instantaneous) physical situation.

One way to approach this task is by asking which states are *structurally* equivalent, much as we considered which Hamiltonian vector fields are equivalent relative to the structure of the final constraint surface. More precisely, we can ask which points are equivalent in the sense that the theoretical structure is unaffected by how the value of a given function is distributed across them. In other words, if there is some vector field such that arbitrarily changing the value of any function along that vector field leaves the relevant structure intact, then the points along the integral curves of such vector field ought to be regarded as equivalent from the perspective of the theory: the theory doesn't depend on which point is taken to represent the "actual" state. If the theory captures all genuine physical distinctions, then this same perspective also tells us which states should count as physically indistinguishable.

We can represent arbitrary changes in how one distributes the value of some property across some set of points in terms of moving the points themselves around while keeping the properties 'fixed'. In other words, we can think about which states are equivalent by thinking about which points on the final constraint surface are such that arbitrarily transforming one point to another preserves the structure of the final constraint surface.

It turns out that such points are given by the points along the integral curves of the null vector fields i.e. the points along the gauge orbits. To see this precisely, consider the (smooth) diffeomorphism h that takes one along the gauge orbits by an arbitrary amount at each point on Σ_f. Then the following is true:

> **Proposition:** h is an automorphism of the structure $(\Sigma_f, \tilde{\omega}_f)$ i.e. h is a diffeomorphism $h : \Sigma_f \to \Sigma_f$ such that $h^*(\tilde{\omega}_f) = \tilde{\omega}_f$.
>
> *Proof*: Since h takes each point on Σ_f to another arbitrary point along the gauge orbit associated with the first-class constraints φ_j at that point, we can represent h as the flow of the vector field associated with $\alpha^j d\varphi_j$ where α^j are arbitrary functions. Since $d\varphi_j = 0$ on Σ_f, $\alpha^j d\varphi_j = 0$. This means that $\alpha^j d\varphi_j$ is closed i.e. $d(\alpha^j d\varphi_j) = 0$. But this means that one can (locally) associate a vector field Y with $\alpha^j d\varphi_j$ via $\tilde{\omega}_f(Y, \cdot) = \alpha^j d\varphi_j$. It follows that the flow of Y on Σ_f consists of maps that preserve $\tilde{\omega}_f$.[10] So h is a diffeomorphism that takes $\tilde{\omega}_f$ to itself.

This proposition shows that arbitrary transformations along the gauge orbits are structure-preserving maps of the presymplectic structure final constraint surface. Importantly, the proof of the proposition relies centrally on the fact that the arbitrary transformations are along the *null* vector fields—if this were not true, then one may not be able to associate a vector field with the transformation whose maps preserve the induced two-form. This highlights that what makes gauge orbits special—as compared to other (integral curves of) vector fields—is precisely that the structure of the state space doesn't discriminate between different ways of distributing the value of a function along the gauge orbits.[11]

This proposition therefore provides a way of associating transformations along the null vector fields with equivalence in terms of the structure of the state space. However, the 'theory' under consideration is not just a theory of the state space; it is also a theory of the dynamics of a system, which is specified in terms of the Hamiltonian function. One therefore might worry that the proposition doesn't capture the points that are *physically* equivalent to one another, if the Hamiltonian is able to distinguish these points. In particular, if the value of the Hamiltonian is different at points along the gauge orbits, then

[10] This follows from Abraham & Marsden (1987) Proposition 3.3.6. (when we extend the proposition to presymplectic manifolds).

[11] Indeed, the flow along any vector field X_f associated with a function f via $\tilde{\omega}_f(X_f, \cdot) = df$ consists of maps that are automorphisms of $(\Sigma_f, \tilde{\omega}_f)$. This implies that one can change the value of f and preserve the structure of the state space. However, this is to change f by the *same* amount at each point and not, as in the proposition above, by an arbitrary amount along X_f at each point of Σ_f.

even if the structure of the state space cannot distinguish these values, one should still treat them as physically distinct.

However, recall that the final constraint surface is defined in terms of the fact that the dynamics has solutions everywhere. What this means is just that a well-defined vector field can be associated with the Hamiltonian via $\tilde{\omega}_f(X_H, \cdot) = dH|_{\Sigma_f}$ and is tangent to the final constraint surface. But this in turn implies that the Hamiltonian is invariant along the gauge orbits, since its change along the gauge orbits is given by $\tilde{\omega}_f(X_H, X_{\varphi_j}) = 0$. In terms of the proposition above, this is the same as saying that $h^*(H|_{\Sigma_f}) = H|_{\Sigma_f}$. Therefore, the proposition can be extended to show that h is an automorphism of the structure $(\Sigma_f, \tilde{\omega}_f, H|_{\Sigma_f})$, which includes not just the structure of the state space, but also the structure that defines the dynamics.

This demonstrates that the gauge orbits are not simply a way of geometrically encoding the gauge transformations on states: the null surfaces correspond to the gauge-equivalent points *because* such points cannot be distinguished by the theory, where the theory is determined by the structure of the constraint surface along with the Hamiltonian on this space. This suggests a revision to the definition of the state gauge transformations:

> **State Gauge Transformation, Geometrically:** A state gauge transformation is a transformation along the gauge orbits on the final constraint surface that preserves the induced two-form and the Hamiltonian.

On this definition of the state gauge transformations, arbitrary combinations of first-class constraints generate gauge transformations, in line with Dirac's conclusion. However, the reasoning is importantly different to Dirac's. In particular, one doesn't need to restrict to the points of the state space where the equations of motion are satisfied in order to determine the gauge transformations on states. One therefore cannot object to this definition on the basis that it makes the state gauge transformations a special case of the solution gauge transformations. Rather, it provides an independent account of what the state gauge transformations are: they relate points on the state space that are equivalent from the perspective of the structure of the theory.

6.3 Solution Gauge Transformations

Finally, let us turn to gauge transformations acting on solutions—that is, transformations relating solutions that are physically equivalent. In the geometric framework, solutions correspond to integral

curves of the vector field(s) generated by the Hamiltonian. What we seek, then, is a notion of equivalence between these integral curves.

We have already determined the equivalence class of vector fields that generate solutions; they are the non-unique vector fields associated with the Hamiltonian, which we called the *Extended Hamiltonian vector fields*. So solutions that differ just in terms of which vector field in the Extended Hamiltonian vector fields generate it are equivalent from the perspective of the structure of the theory.

This motivates the following definition of the gauge transformations on solutions as a way to capture those solutions that are physically equivalent to each other, according to the theory:

> **Solution Gauge Transformation, Geometrically:** A solution gauge transformation relates any two integral curves of the Extended Hamiltonian vector fields.

Since the integral curves of the Extended Hamiltonian vector fields differ only with regard to where on the gauge orbit they lie at each point in time, transforming a solution by an arbitrary amount along the gauge orbit at each point gives rise to another solution generated by a Hamiltonian vector field with a different combination of null vector fields. Therefore, solutions that differ just in terms of where each point lies along the gauge orbit are related by a solution gauge transformation in the revised sense. In other words, arbitrary combinations of first-class constraints generates solution gauge transformations in this sense, just as Pooley & Wallace (2022) show in the case of Electromagnetism.

The solution gauge transformations are defined as acting on individual curves, while the state gauge transformations are defined as acting on the entire state space. However, they can be related in the following way. First, the state gauge transformations by definition preserve the induced two-form, and they preserve the Hamiltonian. They therefore also preserve the equations of motion. This implies that if one fixes some solution to the equations of motion, a state gauge transformation preserves the fact that it is a solution. In other words, the action of the state gauge transformation on a solution will be a solution gauge transformation. Similarly, a solution gauge transformation can be extended to a structure-preserving map on the final constraint surface. Therefore, we can say that both state and solution gauge transformations are generated by arbitrary combinations of first-class constraints, while keeping the notions conceptually distinct.

This provides another argument for why it is problematic to take solution gauge transformations to relate integral curves of the vector fields associated with the *Total* Hamiltonian: it would be to say that there is a physical difference between certain states along a gauge orbit, even though these

states cannot be distinguished by the structure of the state space on which the dynamics is defined, nor by the Hamiltonian function on this state space. It would therefore require one to commit to there being some further structure that distinguishes the points along the integral curves of the vector fields associated with the secondary constraints. Inasmuch as the Hamiltonian is the generator of the dynamics—and therefore determines the predictions of the theory—it is not clear what structure could play this role.

We can thus conclude that the standard view—that gauge transformations are generated by arbitrary combinations of first-class constraints—emerges naturally from the geometric formulation of the constrained Hamiltonian formalism. However, we have also seen that this view conflates two ways of understanding gauge transformations: as relating equivalent ways of representing the state space of the theory, and as relating equivalent solutions to the equations of motion. Both are motivated by the mathematical structure of the theory, but distinguishing them clarifies the relationship between, and significance of, taking gauge transformations to act on *states* as opposed to on *solutions*.

7 Possible Counterarguments

There is one clear avenue for responding to the argument in the previous section: one can reject the claim that the theory defined on the final constraint surface captures the physical content of the theory, and therefore provides the basis for determining the gauge transformations. Let us consider two versions of this counterargument. First, that one shouldn't restrict to the points of the final constraint surface. Second, that we shouldn't think that the geometric formulation of the constrained Hamiltonian formalism is adequate more generally.

Starting with the first objection, one might maintain that points off the final constraint surface should not be discarded, since they continue to play a role in the theory. In particular, although the *dynamics* are restricted to the points on the final constraint surface, we might take the points off the final constraint surface to be *kinematically* possible states. The reason is that the secondary constraints are determined by the consistency of the primary constraints with the dynamics, and so the points of the final constraint surface are those that are dynamically accessible rather than kinematically accessible. If one adopts this view, then because the vector fields associated with the secondary first-class constraints are not null vectors of the two-form on the full phase space nor the primary constraint surface, we cannot use the fact of them being null to argue that they generate gauge

transformations. Indeed, on the primary constraint surface, the equivalence class of Hamiltonians is the Total Hamiltonian, and the null surfaces are just the points along the integral curves of the vector fields associated with the *primary* first-class constraints.

One natural response is that the points off of the final constraint surface are 'excess structure': although there is nothing inconsistent about including them, the content of the theory is given by the final constraint surface. Indeed, why should one formulate a theory in terms of points that the dynamics could never reach, even in principle?

Another response is to point out that the idea that we start out with the primary constraints and then generate the secondary constraints through the dynamics is somewhat an accident of the way that the constrained Hamiltonian formalism is usually set up. As I presented Dirac's version of the theory, one starts with a Lagrangian function, from which one derives the primary constraints. Only once we have the primary constraints and the Hamiltonian in hand do we determine the secondary constraints. But we could have set up the Hamiltonian formalism in a different way: we could say that our theory is given by specifying a Hamiltonian function, a symplectic two-form, and a collection of constraints. In this way of setting up the formalism, although there is a functional relationship between the primary and secondary constraints, there is no clear difference in the role that they play. In particular, the only relevant difference seems to be which constraints are first-class; these are the ones that generate transformations that keep one along the constraint surface and correspond to null vector fields of the induced two-form on the constraint surface.

To push back on this response, one would have to argue that there is something wrong with treating the primary and secondary constraints on the same footing. This leads to the second objection, namely that the geometric formulation of the constrained Hamiltonian formalism is not adequate. There is a clear sense in which this formulation is well-motivated from within the Hamiltonian framework—it is a natural extension of the widely accepted formulation of Hamiltonian mechanics using symplectic manifolds. But one might want to argue that it is inadequate in a different way: it is inadequate because it fails to capture the same content as the Lagrangian formalism. The argument might go as follows. The Hamiltonian formalism for gauge theories is derived from the Lagrangian one, and so its adequacy depends on whether it captures the same content as the Lagrangian theory. From this perspective, the primary constraints are necessary; they have to be imposed in order for the map from the Lagrangian to Hamiltonian state spaces (the Legendre transformation) to be invertible. However, the secondary constraints are not: they are 'extra' constraints on the Hamiltonian side that are not

motivated from the Lagrangian perspective.

Therefore, this argument goes, restricting to the final constraint surface—and consequently having the view that arbitrary combinations of first-class constraints generate gauge transformations—leads to a theory that is inequivalent to the Lagrangian theory, and so is not the right theory to consider. Indeed, one can show that the Total Hamiltonian formalism, understood as relying on the primary constraint surface, gives rise to solutions that are equivalent to the solutions to the Euler-Lagrange equations (Batlle et al. (1986)). Therefore, even if one can show that the solutions to the Extended Hamiltonian are empirically equivalent to those of the Total Hamiltonian, restricting to the final constraint surface gives rise to a theory whose solutions are not equivalent to the Lagrangian solutions. And so, if one takes the view that the Lagrangian formalism is the "fundamental" one, then one might conclude that the definition of a gauge transformation in the Hamiltonian formalism should be inherited from the Lagrangian formalism, and consequently not the definition motivated by the geometry of the constraint surface.

Whether this argument holds up depends on the details of what makes one theory more fundamental than another, and how best to characterize the equivalence of theories—questions that this paper has not addressed directly.[12] However, the discussion here indicates that if one wants to advocate for "Solution Gauge Transformation" as defined in Section 3.1 as capturing the right notion of equivalence, these are questions that one is forced to face. In particular, one must explain what distinguishes primary from secondary constraints, a task that seems to require taking a definite stance on the relationship between the Hamiltonian and Lagrangian formalisms.

8 Conclusion

To summarize, I have argued that the debate about the correct characterization of the gauge transformations in the constrained Hamiltonian formalism can be resolved in favor of the orthodox position. However, the resolution relied on deriving the gauge transformations in a novel way—through consideration of the mathematical structure of the geometric formulation of the theory. In doing so, I showed that there are (at least) two notions of equivalence that the gauge transformations might be thought to capture that I think have not been clearly distinguished previously.

There are some further issues that we have not had space to discuss in any detail. An important issue is the "Problem of Time". Recall that the problem arises from the fact that for theories that are

[12] These issues are addressed in Bradley (2025).

time-reparameterization invariant, the standard account of gauge transformations implies that time evolution is itself a gauge transformation since the Hamiltonian is a first-class constraint. At first glance, endorsing the orthodox view might seem to carry these difficulties along with it. However, the distinctions developed here point to what is especially puzzling about the Hamiltonian constraint: contrary to the cases we have discussed here, the separation between gauge transformations on states and on solutions cannot be clearly drawn when the Hamiltonian is a first-class constraint, since the gauge orbits coincide with the solutions to the equations of motion. As a result, states along a gauge orbit cannot be understood independently of the dynamics, and the possibility of drawing two distinct notions of equivalence is cast into doubt.[13]

A related issue concerns the relationship between equivalence of mathematical structure and physical equivalence. We have argued that the first ought to be used to bridge the two ways of thinking about gauge transformations—as a transformation linking physically equivalent situations and as a mathematical transformation. However, there is more to be said about the notion of physical equivalence that is implied by the definition of the gauge transformations that we laid out in Section 6. For example, if two states are equivalent from the perspective of the mathematical structure of a theory, does that mean that they *cannot* represent physically distinct states? Answering this question in the positive is part of the puzzle leading to the Problem of Time. But this answer is not necessarily implied by our analysis here; it depends on the interpretational significance of two individual states being related by an isomorphism of the structure of the entire state space. The arguments presented here may therefore provide a basis for further exploration of this question.

Acknowledgements

I am especially grateful to Jim Weatherall for valuable comments on the drafts of the paper and for many conversations about the topic. I also thank Karim Thébault, Sean Gryb, and David Wallace for discussions related to several of the ideas developed here. Finally, I am grateful to the audiences at the Foundations of Physics series at Harvard, the 2023 Foundations of Physics conference in Bristol, and the Quantum Spacetime in the Cosmos conference at Perimeter Institute for their helpful questions and suggestions.

[13] For further discussion, see Gryb & Thébault (2023).

References

Abraham, R. & Marsden, J. E. (1987), *Foundations of Mechanics, Second Edition*, Addison-Wesley Publishing Company, Inc.

Anderson, J. L. & Bergmann, P. G. (1951), 'Constraints in covariant field theories', *Physical Review* **83**(5), 1018.

Baker, D. J. (2010), 'Symmetry and the metaphysics of physics', *Philosophy Compass* **5**(12), 1157–1166.

Barbour, J. & Foster, B. Z. (2008), 'Constraints and gauge transformations: Dirac's theorem is not always valid', *arXiv preprint arXiv:0808.1223* .

Batlle, C., Gomis, J., Pons, J. M. & Roman-Roy, N. (1986), 'Equivalence between the Lagrangian and Hamiltonian formalism for constrained systems', *Journal of Mathematical Physics* **27**(12), 2953–2962.

Bradley, C. (2025), 'The relationship between lagrangian and hamiltonian mechanics: The irregular case', *Philosophy of Physics* **3**(1), 1–23.

Butterfield, J. (2006), On symplectic reduction in classical mechanics, *in* J. Butterfield & J. Earman, eds, 'The Handbook of Philosophy of Physics', North Holland, pp. 1–131.

Castellani, L. (1982), 'Symmetries in constrained hamiltonian systems', *Annals of Physics* **143**(2), 357–371.

Dirac, P. A. M. (1964), *Lectures on Quantum Mechanics*, Dover.

Earman, J. (2004), 'Laws, symmetry, and symmetry breaking: Invariance, conservation principles, and objectivity', *Philosophy of Science* **71**(5), 1227–1241.

Gryb, S. & Thébault, K. (2023), *Time Regained: Volume 1: Symmetry and Evolution in Classical Mechanics*, Oxford University Press.

Henneaux, M. & Teitelboim, C. (1994), *Quantization of Gauge Systems*, Princeton University Press.

Ismael, J. & Van Fraassen, B. C. (2003), Symmetry as a guide to superfluous theoretical structure, *in* K. Brading & E. Castellani, eds, 'Symmetries in physics: Philosophical reflections', Cambridge University Press, pp. 371–392.

Pitts, J. B. (2014*a*), 'Change in hamiltonian general relativity from the lack of a time-like killing vector field', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **47**, 68–89.

Pitts, J. B. (2014*b*), 'A first class constraint generates not a gauge transformation, but a bad physical change: The case of electromagnetism', *Annals of Physics* **351**, 382–406.

Pons, J. M. (2005), 'On dirac's incomplete analysis of gauge transformations', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* **36**(3), 491–518.

Pons, J. M., Salisbury, D. C. & Shepley, L. C. (1997), 'Gauge transformations in the lagrangian and hamiltonian formalisms of generally covariant theories', *Phys. Rev. D* **55**, 658–668.

Pooley, O. & Wallace, D. (2022), 'First-class constraints generate gauge transformations in electromagnetism (reply to pitts)', *arXiv preprint arXiv:2210.09063* .

Thébault, K. P. Y. (2021), The problem of time, *in* E. Knox & A. Wilson, eds, 'The Routledge Companion to Philosophy of Physics', Routledge.