A Fixpoint Theory of Label-Symmetric Classification

F. M. V. Feys
Independent Researcher, Antwerp, Belgium
FRANK@FMVFEYS.COM

Editor: To be assigned

Abstract

The structural reason a softmax classifier never outputs a class it has not been trained on is not gradient descent; it is a theorem. Every standard classifier (softmax, k-nearest-neighbors, decision trees with categorical splits) treats labels as bare names: relabeling the classes in the data relabels the predictions accordingly. We call this *label-symmetry* and ask what it forces, together with *universality* (definedness on every nonempty unambiguous dataset). The *constraint theorem* shows that on any dataset from which at least two labels are absent, a universal label-symmetric learner cannot predict a label outside the training data. The *rigidity theorem* shows that no proper subgroup of $\mathrm{Sym}(Y)$ achieves the same constraint: every weaker symmetry hypothesis admits a fabricating learner. The *characterization theorem* parametrizes the full class as a product over $\mathrm{Sym}(Y)$-orbits of stabilizer fixpoint sets, making the learner space explicitly computable, with a closed-form count via Bell and Stirling numbers. In the cross-entropy setting, a universal label-symmetric soft-label classifier may leak probability mass onto unseen labels, but only uniformly across them. Four impossibility theorems of axiomatic learning theory, namely locality, double-symmetry, ambiguity, and the compositional trichotomy, are recovered as specializations of a single fixpoint computation, the *master constraint*.

Keywords: label-symmetry, equivariant learning, fixpoint theory, impossibility theorems, axiomatic learning theory

1 Introduction

1.1 The Question

What does permutation-equivariance in the label space force upon a learner? Every standard classifier is implicitly permutation-equivariant in its label indices. Softmax cross-entropy, k-nearest-neighbors, decision trees with categorical splits, random forests: each treats the labels as bare names, so that relabeling the classes in the data relabels the classifier in step.[1]

A *learner* takes evidence and produces a predictor. Given a finite collection of labeled examples $D \subseteq X \times Y$, a learner L returns a function $L(D)\colon X \to Y$ used to predict the label of any new input. Nearest-neighbor methods, decision trees, neural networks, kernel methods: each is such a map $D \mapsto L(D)$ from data to predictors.

Equivariance under group actions has been a major theme in the structural theory of machine learning, beginning with the group-equivariant convolutional networks of Cohen and Welling (2016), the DeepSets formalism of Zaheer et al. (2017) for permutation-invariant set functions, the general framework of Kondor and Trivedi (2018) for equivariance under compact group actions, and the

1. The claim is about the idealized learning procedure (the architecture together with the update rule), abstracted from initialization noise and sampling randomness. Individual stochastic realizations of a label-symmetric procedure may not themselves be label-symmetric functions; we return to this distinction in Section 11.2. Architectures with label-specific priors or class-imbalance corrections (label smoothing with non-uniform target distributions, class-weighted loss, asymmetric regularization on individual logits) introduce label-specific structure into the learning procedure itself and therefore lie outside the framework; the framework applies to procedures that are symmetric in the labels at the level of the architecture and update rule.

probabilistic version developed by Bloem-Reddy and Teh (2020). These works study equivariance of predictors under groups acting on the *input* space: translations and rotations on images, node-permutations on graphs, set-permutations on point clouds. The aim is parameter-efficient architectures that exploit input symmetries.

We ask the dual question. *What does equivariance of the learner under the symmetric group acting on the* output *space force?* Concretely: if L is universal and label-symmetric, what does it look like? Universality says the learner is defined on every nonempty unambiguous dataset; it must handle whatever data we hand it. Label-symmetry says that relabeling the data relabels the prediction in step; it treats labels as bare names. Each demand is minimal on its own. Together they pin the learner down sharply, and the three theorems of this paper describe the resulting class completely.

1.2 A Picture in Three Labels

The central force of label-symmetry is best seen on a small example. The claim is this: *a universal label-symmetric learner cannot predict a label that is missing from the training data, provided at least two labels are missing.* The argument runs in three lines on three labels.

Let $Y = \{\text{red}, \text{green}, \text{blue}\}$ and $X = \{x_1, x_2\}$. A dataset D is formally a multiset of pairs in $X \times Y$ (Section 2), which on the unambiguous domain is equivalent to a partial function $X \rightharpoonup Y$, and we write $Y_D \subseteq Y$ for the set of labels actually appearing in D. Take $D = \{(x_1, \text{red})\}$, so $Y_D = \{\text{red}\}$. Two labels (green and blue) are missing, and we ask what the learner can return at the unseen input x_2.

It cannot return green. If it did, swap green and blue throughout: the dataset is unchanged (neither label appears in it), and by label-symmetry the prediction must swap too, from green to blue. But the original assumption requires the prediction to be green, while equivariance requires it to be blue; the same prediction cannot be both, contradiction. The only label stable under every permutation that fixes the dataset is the one already in the dataset, namely red. The learner is forced to predict red. This is the constraint theorem in three lines: a label outside the data is moved by some permutation that fixes the data, hence cannot be the prediction.

The hypothesis $|Y_D| \leq |Y| - 2$ (at least two labels missing) is essential. When only one label is missing, the argument breaks. With $Y = \{\text{red}, \text{green}\}$ and $D = \{(x_1, \text{red})\}$, the only nontrivial label permutation swaps red and green, which does *not* fix the dataset, so the rigidity of the previous argument vanishes; the learner is free to predict red or green at x_2. We call $|Y_D| \leq |Y| - 2$ the *generic case* and its complement $|Y_D| \geq |Y| - 1$ the *extremal case*. The constraint theorem says label-symmetry entails that the prediction lies in Y_D in the generic case, and the example just shown is the smallest nontrivial instance.

1.3 Three Theorems

Three theorems formalize and extend this picture.

The first records what we have just argued.

> *Constraint Theorem. If L is universal and label-symmetric, then on every generic dataset D (meaning $|Y_D| \leq |Y| - 2$) and every input x, $L(D)(x) \in Y_D$.*

This result says that label-symmetry, an axiom about permutations, forces no-fabrication, a property about prediction support. In fact, the two are not independent.

The second says that full label-symmetry is exactly what the argument requires. Nothing weaker suffices.

> *Rigidity Theorem. For $|Y| \geq 3$ and any proper subgroup $G \subsetneq \mathrm{Sym}(Y)$, there exists a universal G-equivariant learner L and a generic dataset D with $L(D)(x) \notin Y_D$.*

Replace $\mathrm{Sym}(Y)$ by, say, the cyclic group $\langle(\text{red green blue})\rangle$, and an explicit universal learner equivariant under this cyclic action can predict blue at the unseen input of $\{(x_1, \text{red})\}$. Under the cyclic group, the dataset's stabilizer is trivial (the cyclic generator sends red to green, so does not fix the dataset); the master constraint is therefore vacuous and any label, including blue, is permitted at the unseen input. Only the full symmetric group puts every label outside the data on equal footing.

The third theorem describes the full class.

> *Characterization Theorem. The space of universal label-symmetric learners on $\mathcal{D}_u \setminus \{\emptyset\}$ is in canonical bijection with the product*
>
> $$\prod_{[D]} \big(\mathrm{Fix}(D)\big)^{|X|},$$
>
> *where the product runs over $\mathrm{Sym}(Y)$-orbits of nonempty unambiguous datasets, D is a chosen representative of each orbit, and $\mathrm{Fix}(D)$ is the joint fixpoint set of the relabeling stabilizer of D (defined precisely in Definition 17 below).*

A universal label-symmetric learner is determined by its prediction on a single representative from each orbit, and that prediction must lie in a stabilizer fixpoint set. Indeed, the constraint theorem reappears here as the generic calculation of these fixpoint sets: when $|Y_D| \leq |Y| - 2$, the stabilizer fixes exactly Y_D, recovering $L(D)(x) \in Y_D$. In the extremal case, the stabilizer fixes all of Y, and the prediction is genuinely free. Counting orbits and counting fixpoint sets gives a closed-form expression for the total number of universal label-symmetric learners in terms of Bell and Stirling numbers (Section 5).

A fourth result is concerned with the cross-entropy setting, where classifiers return soft labels (probability distributions over Y) rather than hard labels.

> *Soft Constraint Theorem. For a universal label-symmetric soft-label learner on $\mathcal{D}_u \setminus \{\emptyset\}$, every generic dataset D, and every input x, the prediction $L(D)(x) \in \Delta(Y)$ may assign mass to unseen labels in $Y \setminus Y_D$ only* uniformly *across them: no individual unseen label can be favored over another.*

This is the natural extension of the constraint theorem to soft outputs, with the hard-label conclusion $L(D)(x) \in Y_D$ recovered as the deterministic specialization. The soft constraint isolates the practical content of label-symmetry for cross-entropy classifiers: such a learner may leak probability mass onto unseen labels but cannot prefer one unseen label over another.

1.4 The Single Principle Behind All Three Theorems

The three theorems say one thing. *Label-symmetry is the right axiom, in a precise and complete sense: nothing weaker forces the learner to refrain from inventing labels, nothing stronger is needed, and the entire space of learners satisfying it is small and explicitly enumerable.*

A single principle drives every result. We call it the *master constraint* (Lemma 18, with generalizations Lemmas 29 and 56): if L is equivariant under a group H acting on X and Y, and if some $h \in H$ fixes D, then $L(D)$ is fixed by h; evaluated at a point x also fixed by h, the prediction $L(D)(x)$ lies in the joint fixpoint set of the (D, x)-stabilizer in H. The principle appears in three forms across the paper: the label-symmetric form (Lemma 18), the G-equivariant form (Lemma 29), and the joint H-equivariant form (Lemma 56); each is a specialization or generalization of the same fixpoint principle.

The constraint theorem is the specialization $H = \mathrm{Sym}(Y)$ acting by relabeling, with the stabilizer of an unambiguous D computed explicitly (it is the pointwise stabilizer of Y_D). The characterization theorem packages all the stabilizer fixpoint sets across orbits into a single product. The rigidity theorem says no proper subgroup $G \subsetneq \mathrm{Sym}(Y)$ admits the same strong constraint, and exhibits, for

each such G, an explicit witness learner that fabricates. The four impossibility theorems of locality, double-symmetry, ambiguity, and the compositional trichotomy each fall out by specializing the master constraint to a particular (H, D, x) and observing that the joint fixpoint set is empty. In fact, these four results are not independent; they are four instances of one computation.

The framework offers three concrete affordances for the working theorist. First, a *litmus test for label-fabrication*: given a classifier architecture, identify the symmetry group of its label space; the constraint theorem tells whether the architecture can fabricate on the generic case, and the rigidity theorem hands one an explicit witness when it can. Second, a *unification of impossibilities*: the four phenomena above, previously treated as separate results in axiomatic learning theory and classification aggregation (Maniquet and Mongin, 2016; Cailloux et al., 2024), are recovered here as specializations of a single fixpoint computation. Third, a *complete enumeration*: the characterization theorem describes the full space of universal label-symmetric learners as a product of stabilizer fixpoint sets, with a closed-form combinatorial count via Bell and Stirling numbers; at $|X| = 2$, $|Y| = 3$ there are exactly nine such learners (Section 9). The orbit-decomposition engine itself is folklore in equivariant combinatorics.

The framework is most informative precisely where label-symmetry is partial rather than full. Architectures with full $\mathrm{Sym}(Y)$-equivariance (softmax cross-entropy, k-nearest-neighbors, decision trees with categorical splits, kernel methods with label-symmetric kernels) inherit no-fabrication on the generic case for free, and the result here gives a structural reason for what was previously a per-architecture observation. The substantive regime is partial label-symmetry: hierarchical labels carrying the action of a Young subgroup, ordinal labels carrying a cyclic action or none, and embedded labels carrying the symmetry group of the embedding. In each case the rigidity theorem identifies an explicit fabricating witness via Corollary 36, locating the obstruction precisely. The framework therefore organizes a question the field has handled architecture-by-architecture into a single structural statement parametrized by the symmetry group.

1.5 Scope

Before proceeding, we note the scope of the present contribution. The framework is combinatorial and discrete. We work with finite X and Y, treat labels as bare set elements with no metric or order, and do not address probability, generalization, sample complexity, or optimization. While real classification problems are often effectively finite (images are finite pixel grids, text is bounded sequences over a fixed vocabulary, sensor data lives at finite precision), the finiteness assumptions in our framework are a tractability choice rather than a claim about realism: they enter only in the orbit-counting corollaries (Bell, Stirling). The infinite-input abstraction useful for asymptotic analysis (PAC bounds, VC dimension) is therefore not in tension with the present framework, simply orthogonal to its purposes. Two distinctions are worth noting. Finite X is not load-bearing in the spine theorems: the master constraint, the constraint theorem, and the rigidity theorem all go through verbatim for arbitrary X, and only the orbit-counting corollaries (Bell, Stirling) require finiteness. Finite Y, by contrast, is structural to the symmetry hypothesis itself: label-permutation symmetry requires labels that can be permuted, and continuous label spaces lack a natural finite symmetric group action and so fall outside the framework rather than carrying a weakened version of it. The results describe the *logical backbone* of empirical learning under symmetry, before any quantitative or distributional layer is added. The mathematics is elementary. The case $|Y| = 3$ recurs as a structural boundary throughout: many results split into $|Y| = 2$ and $|Y| \geq 4$ regimes with $|Y| = 3$ as the exception, governed by the core lemma of Section 7; we indicate each such occurrence in place.

In relation to the equivariant-ML literature, the closest precedent is the probabilistic equivariance theory of Bloem-Reddy and Teh (2020), which gives functional representations of probability distributions invariant under compact group actions. Their framework operates on probability distributions and applies to compact groups acting on input spaces; ours operates set-theoretically

and on the finite group $\mathrm{Sym}(Y)$ acting on the output. The two are complementary: the Bloem-Reddy-Teh representation describes *which functions* a probabilistic equivariant learner can express; the characterization theorem of this paper describes *which deterministic learners* are admissible at all under finite label-permutation symmetry, with a closed-form parameter count. The constraint and rigidity theorems in fact have no analogue in the probabilistic theory, since the generic obstruction is a discrete-finite phenomenon that requires the cardinality gap $|Y| - |Y_D| \geq 2$. The soft-label extension (Section 8.5) is the bridge between the two frameworks, recovering a probability-distribution version of the constraint theorem on $\Delta(Y)$.

1.6 Outline

Section 2 fixes the setup and motivates the axioms. Sections 3, 4, and 5 prove the three theorems in turn; together they form the spine. Section 6 derives the three impossibility theorems of locality, double-symmetry, and ambiguity as corollaries of the master constraint and revisits the no-fabrication axiom in light of constraint and rigidity; these are checkable consequences of the three main theorems with no new proofs. Section 7 proves the core combinatorial lemma on symmetric equivariant functions and uses it to derive the compositional trichotomy at $|Y| = 3$ as the headline extremal phenomenon. Section 8 extends the master constraint to general H-equivariance and to the soft-label setting, where the constraint theorem becomes a statement about probability distributions on Y. Section 9 maps the lattice of axiom-consistency and treats self-training closure and monotonicity in detail. Section 10 positions the contribution and Section 11 concludes.

2 Setup and Axioms

2.1 Datasets and Learners

We fix two finite nonempty sets X and Y. Elements of X are *inputs* (also called *features* in machine learning); elements of Y are *labels* (also called *classes* in machine learning, though we prefer *labels*, since the framework treats the elements of Y as bare names with no intrinsic content). The case $|Y| = 1$ is degenerate, since only one predictor exists and every axiom is satisfied trivially; we otherwise assume $|Y| \geq 2$, and most substantive results require $|Y| \geq 3$.

Definition 1 (Dataset)**.** *A* dataset *is a finite multiset D of pairs in $X \times Y$. We write $\mathcal{D}(X, Y)$ for the set of all such multisets, including the empty dataset $\emptyset$. A dataset D is* unambiguous *if each input $x \in X$ appears with at most one distinct label across D, allowing repetitions of the same pair; the collection of unambiguous datasets is $\mathcal{D}_u(X, Y) \subseteq \mathcal{D}(X, Y)$. The* input domain *of D is*

$$\mathrm{dom}(D) := \{x \in X \mid (x, y) \in D \text{ for some } y\},$$

and the label support *(also called the* label set of *D) is*

$$Y_D := \{y \in Y \mid (x, y) \in D \text{ for some } x\}.$$

A dataset is nonempty *if $\mathrm{dom}(D) \neq \emptyset$. We write $\mathcal{D}_u \setminus \{\emptyset\}$ for the collection of nonempty unambiguous datasets and work primarily on this domain.*

Definition 2 (Learner)**.** *A* learner *is a partial function $L \colon \mathcal{D}(X, Y) \rightharpoonup Y^X$, where Y^X denotes the set of total functions $X \to Y$, called* predictors*. We write $\mathrm{dom}(L) \subseteq \mathcal{D}(X, Y)$ for the set of datasets on which L is defined, $L(D) \colon X \to Y$ for the predictor returned on $D \in \mathrm{dom}(L)$, and $L(D)(x) \in Y$ for its value at $x \in X$.*

An unambiguous dataset is, up to multiplicities of identical pairs, equivalent to a partial function $X \rightharpoonup Y$: it specifies which label is attached to each input in its domain, together with how many copies of each (x, y) pair appear in D. The idempotence axiom (Definition 11 below) is the assumption

that learners do not see these multiplicities; under idempotence, the multiset description collapses to the partial-function description. We use both descriptions in the text, with the multiset structure primary and the partial-function summary invoked when idempotence is in force.

2.2 Generic and Extremal Datasets

We single out two classes of unambiguous datasets, distinguished by how much of the label set Y is covered by the label support Y_D.

Definition 3 (Generic and Extremal Datasets)**.** *An unambiguous dataset $D \in \mathcal{D}_u(X, Y)$ is* generic *if $|Y_D| \leq |Y| - 2$, that is, at least two labels of Y are missing from D. It is* extremal *if $|Y_D| \in \{|Y| - 1, |Y|\}$, that is, at most one label of Y is missing from D.*

Every unambiguous dataset is either generic or extremal; the two cases are mutually exclusive and exhaustive.

Example 1. *Let $|Y| = 4$ with $Y = \{a, b, c, d\}$. The dataset $D_1 = \{(x_1, a), (x_2, a)\}$ has $Y_{D_1} = \{a\}$, so $|Y_{D_1}| = 1$, and D_1 is generic (three labels missing). The dataset $D_2 = \{(x_1, a), (x_2, b)\}$ has $Y_{D_2} = \{a, b\}$, so $|Y_{D_2}| = 2 = |Y| - 2$, and D_2 is generic (exactly two labels missing). The dataset $D_3 = \{(x_1, a), (x_2, b), (x_3, c)\}$ has $Y_{D_3} = \{a, b, c\}$, so that $|Y_{D_3}| = 3 = |Y| - 1$, and D_3 is extremal (one label missing). The dataset $D_4 = \{(x_1, a), (x_2, b), (x_3, c), (x_4, d)\}$ has $Y_{D_4} = Y$, so $|Y_{D_4}| = 4$, and D_4 is extremal (no labels missing).*

Remark 4. *At $|Y| = 2$, the condition $|Y_D| \leq |Y| - 2 = 0$ implies $Y_D = \emptyset$, hence $D = \emptyset$. The only generic dataset at $|Y| = 2$ is therefore the empty one; every nonempty dataset is extremal. The generic/extremal distinction is correspondingly vacuous in the two-label setting and acquires substantive content only for $|Y| \geq 3$. On the domain $\mathcal{D}_u \setminus \{\emptyset\}$, the constraint theorem therefore has no content at $|Y| = 2$, since every nonempty unambiguous dataset is extremal and the generic case is empty.*

2.3 The Symmetry Actions

The symmetric group $\mathrm{Sym}(Y)$ acts on datasets by relabeling. For $\sigma \in \mathrm{Sym}(Y)$, we set

$$\sigma D := \{(x, \sigma(y)) \mid (x, y) \in D\},$$

extended to multisets in the natural way. The induced action on a predictor $f \colon X \to Y$ is $\sigma \cdot f := \sigma \circ f$, applied post-composition.

The symmetric group $\mathrm{Sym}(X)$ acts by input-relabeling. For $\tau \in \mathrm{Sym}(X)$, we set

$$\tau \cdot D := \{(\tau(x), y) \mid (x, y) \in D\},$$

and on predictors, $\tau \cdot f := f \circ \tau^{-1}$, so that the predictor still has type $X \to Y$.

2.4 Universality and Label-Symmetry

Each axiom in this section and the next is a property of a fixed learner $L \colon \mathcal{D}(X, Y) \rightharpoonup Y^X$ (Definition 2): a universal statement quantifying over *every* dataset, input, and permutation to which it refers, with L itself not quantified inside the axiom. Many axioms include a well-definedness qualifier on $\mathrm{dom}(L)$, for instance "for every $\sigma \in \mathrm{Sym}(Y)$ and every $D \in \mathrm{dom}(L)$ with $\sigma D \in \mathrm{dom}(L)$"; these qualifiers are essential, since without them the axiom would silently force closure of $\mathrm{dom}(L)$ under the relevant action.

We single out the two axioms that drive the main theorems and motivate each in detail.

Definition 5 (Universality)**.** *A learner L is* universal *(on $\mathcal{D}_u \setminus \{\emptyset\}$) if $\mathrm{dom}(L) = \mathcal{D}_u \setminus \{\emptyset\}$, that is, L is defined on every nonempty unambiguous dataset.*

Universality says the learner is defined on every reasonable training set: feed it data, receive a predictor back. The empty dataset is excluded throughout: requiring the learner to produce a prediction with no data at all forces a contradiction with label-symmetry when $|Y| \geq 2$, as Proposition 51 makes precise. The unambiguous restriction is also deliberate. Datasets with two distinct labels at the same input admit larger relabeling stabilizers, and the framework specializes accordingly (Theorem 47). The bulk of the structural picture lives on the unambiguous domain. Nearest-neighbor methods, bounded-depth decision trees, and neural-network classifiers trained by SGD are universal; learners defined only on total datasets $D\colon X \to Y$ are not.

Definition 6 (Label-Symmetry)**.** *A learner L is* label-symmetric *if for every $\sigma \in \mathrm{Sym}(Y)$ and $D \in \mathrm{dom}(L)$ with $\sigma D \in \mathrm{dom}(L)$ it holds that $L(\sigma D) = \sigma \circ L(D)$.*

Label-symmetry says the labels in Y play interchangeable roles: the learner does not privilege red over blue, and swapping red and blue in the data swaps the predictions accordingly. This is the natural symmetry of categorical classification with no external structure on the labels. Any classifier that refers to labels only through their identity, not through external properties, is label-symmetric: decision trees with category-valued splits, logistic regression without label-specific regularization, symmetric-softmax cross-entropy classifiers. What fails label-symmetry is any classifier with a fixed default label, a label ordering, a label-specific prior, or an ordinal structure. The label-symmetry assumption corresponds to taking labels as *bare names*, and the main theorems describe what this assumption entails. We emphasize that label-symmetry is a property of the *learning procedure* (the architecture together with the update rule and initialization distribution), not of any individual trained predictor obtained on a particular run; a single SGD trajectory of a label-symmetric procedure may yield a predictor that is not itself a label-symmetric function, but the procedure as a whole is.

Remark 7 (On the Unobjectionability of Universality and Label-Symmetry)**.** *We record the precise sense in which each of universality and label-symmetry is unobjectionable, since these senses differ.*

Label-symmetry is forced by the modeling choice. *The framework studies classification in which Y carries no external structure (no metric, no order, no semantic content beyond bare names). Under that choice, the labels are interchangeable, and a learner that gives different predictions when we rename the labels is responding to the renaming rather than to the data. That is what label-symmetry rules out. If Y does carry external structure (ordinal ranking, semantic content, a default class), label-symmetry is no longer the right axiom and the framework simply does not apply. Architectures that learn label-specific representations (prototypical networks, label-embedding methods) similarly sit outside the bare-name modeling choice and require a separate analysis. Within the bare-name setting, however, label-symmetry follows automatically.*

Universality is what a learner is for. *A learner is a procedure that takes data and returns a predictor; refusing some legitimate input means the procedure does not finish its job. Every working classifier (nearest-neighbor methods, decision trees, neural-network classifiers) is universal in this sense. The honest caveat is that universality is also a uniformity demand, and several of the impossibility theorems are partly impossibilities about uniformity: dropping it (by restricting to total datasets) reconciles axioms that are otherwise incompatible, which is exactly the role of the total-data identity learner L_{tot} as a witness in Theorem 68. The reader should therefore read universality as the demand that the learner works on every nonempty unambiguous input, accepting that this demand is doing genuine load-bearing work in the theorems.*

2.5 The Remaining Axioms

The remaining axioms enter the corollaries and refinements of the main theorems. We define each here and motivate them briefly, with a more careful treatment deferred to Sections 6 and 9.

Definition 8 (Consistency)**.** *A learner L is* consistent *if for every $D \in \mathrm{dom}(L)$ and every $x \in \mathrm{dom}(D)$: if every example at x in D has the same label y, then $L(D)(x) = y$.*

On the unambiguous domain, consistency simplifies to $L(D)(x) = D(x)$ for $x \in \mathrm{dom}(D)$. The learner agrees with the data where the data is unambiguous. Every empirical learning method in standard use satisfies consistency on the unambiguous domain.

Definition 9 (Input-Symmetry)**.** *A learner L is* input-symmetric *if for every $\tau \in \mathrm{Sym}(X)$ and every $D \in \mathrm{dom}(L)$ with $\tau \cdot D \in \mathrm{dom}(L)$,*

$$L(\tau \cdot D) = L(D) \circ \tau^{-1}.$$

Input-symmetry is the analogue of label-symmetry for inputs: the learner does not privilege any particular input. A nearest-neighbor method that breaks ties by a fixed input ordering fails input-symmetry. A neural network with no spatial structure on X (no positional encoding) satisfies input-symmetry.

Definition 10 (Strict Locality)**.** *A learner L is* strictly local *if for every $D \in \mathrm{dom}(L)$ and every $x \in X \setminus \mathrm{dom}(D)$, $\emptyset \in \mathrm{dom}(L)$ and $L(D)(x) = L(\emptyset)(x)$.*

Strict locality is the strongest form of locality. The prediction at x depends only on data at x. Most empirical learning methods are nonlocal: nearest-neighbor methods use data at other inputs, and neural networks use the whole data. Strict locality enters the framework as a limiting case, useful chiefly because the locality impossibility (Theorem 45) shows it cannot coexist with universality and label-symmetry.

Definition 11 (Idempotence)**.** *A learner L is* idempotent *if $L(D \uplus \{(x, y)\}) = L(D)$ whenever $(x, y) \in D$ and $D \uplus \{(x, y)\} \in \mathrm{dom}(L)$.*

Idempotence says multiplicities of identical examples do not matter. If we treat datasets as sets rather than multisets, idempotence is automatic; we state it explicitly to keep the multiset structure available.

Definition 12 (Monotonicity)**.** *A learner L is* monotone *if, for every $D \in \mathrm{dom}(L)$, every $(x', y') \notin D$ with $D \cup \{(x', y')\} \in \mathrm{dom}(L)$, and every $x \in X$, if $L(D)(x) = y'$ then $L(D \cup \{(x', y')\})(x) = y'$.*

Monotonicity says: adding more evidence of label y' never removes y' from a prediction.

Definition 13 (No-Fabrication)**.** *A learner L is* nonfabricating *if $L(D)(x) \in Y_D$ for every $D \in \mathrm{dom}(L)$ and every $x \in X$.*

No-fabrication says the learner predicts only labels that appear in the data. Every k-nearest-neighbor classifier satisfies it by construction; a softmax neural network classifier trained with cross-entropy on a dataset D may, in principle, output any label in Y after argmax, and so satisfies no-fabrication only by virtue of the structural results of this paper (the constraint theorem, applied to the learning procedure as a whole) rather than by construction; the cross-entropy case is made precise in the soft-label extension of Section 8.5 (Theorem 61). The constraint theorem (Section 3) shows that this axiom is a *consequence* of label-symmetry on the generic case, not an independent demand. A complete analysis of its regime structure under label-symmetry is given in Proposition 50 (No-Fabrication Revisited).

Self-training closure says: retraining the learner on its own predictions yields the same predictor. This captures the fixpoint structure of empirical self-training as a learning technique.

Definition 14 (Self-Training Closure)**.** *For nonempty unambiguous $D \in \mathrm{dom}(L)$ and $S \subseteq X$ satisfying $\mathrm{dom}(D) \subseteq S$ (so that the self-training closure extends, not replaces, the training data), define*

$$D_S^L := \{(x, L(D)(x)) \mid x \in S\}.$$

We say L is closed under self-training *if $L(D_S^L) = L(D)$ for every such D and S with $D_S^L \in \mathrm{dom}(L)$. Under universality, the condition $D_S^L \in \mathrm{dom}(L)$ is automatic: D_S^L has input domain $S \supseteq \mathrm{dom}(D)$ which is nonempty, and D_S^L is unambiguous because $L(D)$ is a function.*

Definition 15 (Coordinate-Wise Compositionality)**.** *A learner L is* coordinate-wise compositional *if there exists a function $\Phi\colon Y\times Y\to Y$, called the* aggregator*, such that for every $D_1, D_2 \in \mathcal{D}_u \cap \mathrm{dom}(L)$ with disjoint domains ($\mathrm{dom}(D_1)\cap\mathrm{dom}(D_2)=\emptyset$) and $D_1\cup D_2\in\mathrm{dom}(L)$, and every $x\in X$,*

$$L(D_1\cup D_2)(x)=\begin{cases} L(D_1)(x) & \text{if } x\in\mathrm{dom}(D_1),\\ L(D_2)(x) & \text{if } x\in\mathrm{dom}(D_2),\\ \Phi(L(D_1)(x),L(D_2)(x)) & \text{otherwise.}\end{cases}$$

Coordinate-wise compositionality is a strong locality property in dataset space: the prediction on a disjoint union depends only on the predictions on the components, via a fixed aggregator that combines pairs of labels. It is rarely satisfied in practice; it enters the framework because the compositional trichotomy (Section 7) shows that even this strong locality is compatible with universality and label-symmetry only at $|Y|=3$.

Definition 16 (Joint H-Equivariance)**.** *Let H be a group acting on both X and Y. The diagonal action on a dataset is $h\cdot D:=\{(h\cdot x, h\cdot y)\mid (x,y)\in D\}$. The action on a predictor $f\colon X\to Y$ is $(h\cdot f)(x):=h\cdot f(h^{-1}\cdot x)$. We say L is H-*equivariant *if $L(h\cdot D)=h\cdot L(D)$ for every $h\in H$ and every $D\in\mathrm{dom}(L)$ with $h\cdot D\in\mathrm{dom}(L)$.*

Joint equivariance subsumes both label-symmetry ($H=\mathrm{Sym}(Y)$ acting trivially on X) and input-symmetry ($H=\mathrm{Sym}(X)$ acting trivially on Y). We use it in Section 8 to state the general H-equivariant master constraint.

2.6 Summary of the Axioms

For ease of reference, Table 1 summarizes all axioms used in the paper.

Symbol	Name	One-line content
U	Universality	L is defined on every nonempty unambiguous dataset
LS	Label-Symmetry	$L(\sigma D)=\sigma\circ L(D)$ for $\sigma\in\mathrm{Sym}(Y)$
C	Consistency	$L(D)(x)=D(x)$ for $x\in\mathrm{dom}(D)$, on $\mathcal{D}_u$
NF	No-Fabrication	$L(D)(x)\in Y_D$ for every $x\in X$
Idem	Idempotence	Multiplicities of identical examples are ignored
IS	Input-Symmetry	$L(\tau D)=L(D)\circ\tau^{-1}$ for $\tau\in\mathrm{Sym}(X)$
Loc	Strict Locality	$L(D)(x)=L(\emptyset)(x)$ for $x\notin\mathrm{dom}(D)$, with $\emptyset\in\mathrm{dom}(L)$
Mon	Monotonicity	$L(D)(x)=y'\Rightarrow L(D\cup\{(x',y')\})(x)=y'$
CL	Self-Training Closure	$L(D_S^L)=L(D)$ for self-training closures D_S^L
Comp	Coordinate-Wise Compositionality	$L(D_1\cup D_2)(x)=\Phi(L(D_1)(x),L(D_2)(x))$ on disjoint D_1, D_2

Table 1: The ten axioms of the catalogue $\mathcal{A}$. Universality and label-symmetry appear in the hypotheses of the constraint, characterization, and rigidity theorems; the remaining eight axioms are the subject of the consistency lattice of Section 9. Formal definitions are in Section 2.

3 The Constraint Theorem

We state and prove the first of our three theorems. The constraint theorem says that universality and label-symmetry together force predictions to lie in the data, on the generic case $|Y_D|\le|Y|-2$. The proof is a paragraph, resting on two short lemmas: the master constraint (a fixpoint principle from equivariance), and an explicit computation of the stabilizer of an unambiguous dataset under the relabeling action.

3.1 The Master Constraint

The fundamental observation is that equivariance, applied at a fixed point of the group action, becomes *invariance under the stabilizer.*

Definition 17 (Relabeling Stabilizer and Fixpoint Set)**.** *For a dataset $D \in \mathcal{D}_u(X, Y)$, the* relabeling stabilizer *of D is*

$$\mathrm{Stab}_{\mathrm{Sym}(Y)}(D) := \{\sigma \in \mathrm{Sym}(Y) \mid \sigma D = D\}.$$

The joint fixpoint set *of D is*

$$\mathrm{Fix}(D) := Y^{\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)} = \{y \in Y \mid \sigma(y) = y \text{ for every } \sigma \in \mathrm{Stab}_{\mathrm{Sym}(Y)}(D)\}.$$

Intuitively, $\mathrm{Fix}(D)$ is the largest set of candidate predictions that label-symmetry alone permits at any input, given training data D. We use the notation $\mathrm{Fix}(D)$ as a lightweight shorthand for the joint fixpoint set throughout.

Lemma 18 (Master Constraint, Label-Symmetric Form)**.** *Let L be a label-symmetric learner. For every $D \in \mathrm{dom}(L)$ and every $\sigma \in \mathrm{Sym}(Y)$ with $\sigma D = D$ and $\sigma D \in \mathrm{dom}(L)$, we have*

$$L(D) = \sigma \circ L(D).$$

Consequently, for every $x \in X$,

$$L(D)(x) \in \mathrm{Fix}(D).$$

Proof By label-symmetry, $L(\sigma D) = \sigma \circ L(D)$. The hypothesis $\sigma D = D$ gives $L(D) = \sigma \circ L(D)$. Evaluating at any $x \in X$ yields $L(D)(x) = \sigma(L(D)(x))$, so $L(D)(x)$ is fixed by σ. Since this holds for every σ in the stabilizer, $L(D)(x)$ lies in the joint fixpoint set in Y. ■

The lemma reduces the description of $L(D)(x)$ to computing the fixpoint set $\mathrm{Fix}(D)$. We now compute this set explicitly for unambiguous D.

3.2 Stabilizers on the Unambiguous Domain

The next lemma identifies the stabilizer of an unambiguous dataset as the pointwise stabilizer of its label support.

Lemma 19 (Stabilizer on Unambiguous Data)**.** *Let $D \in \mathcal{D}_u(X, Y)$ be unambiguous. Then*

$$\mathrm{Stab}_{\mathrm{Sym}(Y)}(D) = \{\sigma \in \mathrm{Sym}(Y) \mid \sigma(y) = y \text{ for every } y \in Y_D\},$$

the pointwise stabilizer of Y_D in $\mathrm{Sym}(Y)$. This subgroup is canonically isomorphic to $\mathrm{Sym}(Y \setminus Y_D)$ via restriction to the unused labels.

Proof Since D is unambiguous, each label $y \in Y_D$ is attached to a unique nonempty set of inputs $\mathrm{dom}(D, y) := \{x \in X \mid (x, y) \in D\}$. Distinct labels in Y_D have disjoint $\mathrm{dom}(D, y)$.

Suppose $\sigma \in \mathrm{Sym}(Y)$ satisfies $\sigma D = D$. The action sends each $(x, y) \in D$ to $(x, \sigma(y))$ without changing the input. First, σ preserves Y_D setwise: if $y \in Y_D$, then $(x, y) \in D$ for some x, so $(x, \sigma(y)) \in \sigma D = D$, hence $\sigma(y) \in Y_D$. The condition $\sigma D = D$ therefore requires $\mathrm{dom}(D, \sigma(y)) = \mathrm{dom}(D, y)$ for every $y \in Y_D$, where both sets sit inside the pairwise disjoint family $\{\mathrm{dom}(D, y') \mid y' \in Y_D\}$.Two distinct labels in Y_D have disjoint nonempty dom-blocks by unambiguity of D; hence the identity $\mathrm{dom}(D, \sigma(y)) = \mathrm{dom}(D, y)$, which matches the block at y with the block at $\sigma(y)$, implies that $\sigma(y) = y$ for every $y \in Y_D$. Since $\sigma|_{Y_D}$ is a bijection of Y_D to itself, σ fixes Y_D pointwise. Conversely, any σ fixing Y_D pointwise sends each $(x, y) \in D$ to itself, so $\sigma D = D$.

The pointwise stabilizer of Y_D in $\mathrm{Sym}(Y)$ permutes only the elements of $Y \setminus Y_D$, and is therefore canonically isomorphic to $\mathrm{Sym}(Y \setminus Y_D)$. ■

The key fact behind this lemma is that unambiguous data *pins down which label appears at which input, leaving the unused labels unconstrained.* A label permutation that stabilizes such data must therefore fix every label that occurs, leaving free only the unused labels.

3.3 The Fixpoint Set of the Stabilizer

We can now compute the fixpoint set in Y of the stabilizer.

Lemma 20 (Fixpoint Set)**.** *Let D be unambiguous with $|Y_D| = m$. Then*

$$\mathrm{Fix}(D) = \begin{cases} Y_D & \text{if } m \leq |Y| - 2, \\ Y & \text{if } m \geq |Y| - 1. \end{cases}$$

Proof By Lemma 19, $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)$ is the pointwise stabilizer of Y_D in $\mathrm{Sym}(Y)$. Restriction to the unused labels gives a canonical isomorphism $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D) \cong \mathrm{Sym}(Y \setminus Y_D)$; this is the fact we use. Every element of Y_D is fixed by every element of $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)$, since the action on Y_D is trivial by construction. An element $y \in Y \setminus Y_D$ is fixed by every element of $\mathrm{Sym}(Y \setminus Y_D)$ if and only if $|Y \setminus Y_D| \leq 1$, equivalently $m \geq |Y| - 1$. The extremal regime $m \geq |Y| - 1$ splits further into two subcases with the same conclusion: at $m = |Y| - 1$, $\mathrm{Sym}(Y \setminus Y_D)$ is trivial as the group on a singleton, and every element of the singleton $Y \setminus Y_D$ is fixed; at $m = |Y|$, $Y \setminus Y_D$ is empty and the question is vacuous, with the fixpoint set equal to $Y_D = Y$. The two cases of the lemma split on this condition. ■

3.4 The Constraint Theorem

The constraint theorem is now a direct corollary.

Theorem 21 (Constraint Theorem)**.** *Let L be a universal label-symmetric learner on $\mathcal{D}_u \setminus \{\emptyset\}$. For every $D \in \mathcal{D}_u \setminus \{\emptyset\}$ with $|Y_D| \leq |Y| - 2$ and every $x \in X$,*

$$L(D)(x) \in Y_D.$$

Proof By Lemma 18, $L(D)(x) \in \mathrm{Fix}(D)$. By Lemma 19, the stabilizer of an unambiguous D is $\mathrm{Sym}(Y \setminus Y_D)$. By Lemma 20, the fixpoint set in Y of $\mathrm{Sym}(Y \setminus Y_D)$ is Y_D when $|Y_D| \leq |Y| - 2$. Therefore, $L(D)(x) \in Y_D$. ■

The proof is short; the work is in the three lemmas: equivariance forces stabilizer-invariance, the stabilizer of unambiguous data is the pointwise stabilizer of the label support, and the fixpoint set of this stabilizer is Y_D exactly on the generic case. Geometrically: a label-symmetric learner cannot tell the unused labels apart, since they are interchangeable under $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)$, so it cannot predict any of them, except in the degenerate case where only one unused label remains.

3.5 Sharpness and the Extremal Case

The hypothesis $|Y_D| \leq |Y| - 2$ is sharp.

Sharpness in $|Y_D|$. For $|Y_D| = |Y| - 1$ the stabilizer $\mathrm{Sym}(Y \setminus Y_D)$ is trivial, the fixpoint set is all of Y, and the constraint disappears. At $|Y_D| = |Y|$ the stabilizer is again trivial. These two cases together form the *extremal case*; both classical exceptional phenomena (the $|Y| = 3$ compositional trichotomy, the ambiguity escape) live there. Section 7 characterizes the combinatorics of the extremal case via a core lemma on $\mathrm{Sym}(Y)$-equivariant functions. Concretely, at $|Y| = 3$ and $D = \{(x_1, y_1), (x_2, y_2)\}$, the missing-label learner that predicts y_3 at every unseen input is universal

and label-symmetric yet fabricates $y_3 \notin Y_D$. Universality on $\mathcal{D}_u \setminus \{\emptyset\}$ extends: on datasets with $|Y_D| = |Y| - 1$, predict the unique missing label off-domain; on datasets with $|Y_D| \in \{1, |Y|\}$, predict according to any label-symmetric default consistent with the master constraint (the existence of such an extension is established by Theorem 38 in Section 5).

Sharpness in universality. Without universality the constraint is vacuous: simply restrict $\mathrm{dom}(L)$ to datasets with $|Y_D| \geq |Y| - 1$ (or to any domain that excludes all generic datasets) and the generic case is empty.

Sharpness in label-symmetry. This is the converse direction. The rigidity theorem in Section 4 shows that any strictly weaker symmetry hypothesis admits a universal equivariant learner violating the constraint.

Remark 22 (The Extremal Boundary $|Y| = 3$)**.** *The cardinality $|Y| = 3$ is the smallest at which the extremal case $|Y_D| = |Y| - 1$ admits a nontrivial dataset, and it recurs as a structural boundary throughout the paper: the double-symmetry impossibility (Theorem 46), the ambiguity impossibility (Theorem 47), and the compositional trichotomy (Theorem 53) each admit an exception at $|Y| = 3$ (all three are proved in Sections 6 and 7). The common engine is the core lemma of Section 7 on symmetric $\mathrm{Sym}(Y)$-equivariant functions, which produces the missing-label function as the unique nonzero solution exactly when $|Y| = k + 1$ for the relevant arity k. The $|Y| = 3$ exceptions are accordingly not independent anomalies; they are the same combinatorial phenomenon viewed through different axioms.*

3.6 Immediate Consequences

Three corollaries follow immediately.

Corollary 23 (Label-Symmetry Implies No-Fabrication on Generic Datasets)**.** *Every universal label-symmetric learner on $\mathcal{D}_u \setminus \{\emptyset\}$ is nonfabricating on every dataset with $|Y_D| \leq |Y| - 2$.*

Proof Direct from Theorem 21: the conclusion $L(D)(x) \in Y_D$ holds for every $x \in X$, which is exactly the no-fabrication condition. ■

Remark 24 (No-Fabrication Does Not Imply Label-Symmetry)**.** *Corollary 23 gives the implication $\mathsf{U} + \mathsf{LS} \Rightarrow \mathsf{NF}$ on the generic case; the converse however fails. To witness the failure of the converse, fix a total order $\prec$ on Y and define*

$$L_{\prec}(D)(x) := \begin{cases} D(x) & \text{if } x \in \mathrm{dom}(D), \\ \min_{\prec} Y_D & \text{otherwise.} \end{cases}$$

This learner is universal, consistent, and nonfabricating, since $\min_{\prec} Y_D \in Y_D$. It is not label-symmetric: relabeling by a $\sigma \in \mathrm{Sym}(Y)$ that swaps the $\prec$-minimum of Y_D with some other element of Y_D changes which label is selected at unseen inputs, breaking $L_{\prec}(\sigma D) = \sigma \cdot L_{\prec}(D)$.

Corollary 25 (Singleton Triviality)**.** *For $|Y| \geq 3$, every universal label-symmetric learner L satisfies $L(\{(x_0, y_0)\})(x) = y_0$ for every $x \in X$ and every singleton dataset.*

Proof A singleton has $|Y_D| = 1 \leq |Y| - 2$ when $|Y| \geq 3$. The constraint theorem yields $L(D)(x) \in \{y_0\}$. ■

Corollary 26 (Generic Determination)**.** *For $|Y_D| \leq |Y| - 2$, the constraint theorem reduces $L(D)(x)$ to a function $X \to Y_D$, independent of which specific labels of Y are in Y_D up to relabeling.*

Proof By Theorem 21, $L(D) \in Y_D^X$. Label-symmetry makes the dependence equivariant in Y_D rather than in specific labels. ■

Remark 27 (Image and Boundary Observations)**.** *Two immediate extensions of the constraint theorem are worth noting.*

Image surjectivity. *Under universality, label-symmetry, and consistency, the constraint theorem yields $L(D)(X) \subseteq Y_D$ on the generic case; for the reverse inclusion, each $y \in Y_D$ appears at some $x \in \mathrm{dom}(D)$, and consistency gives $L(D)(x) = y$, so $y \in L(D)(\mathrm{dom}(D)) \subseteq L(D)(X)$ since $\mathrm{dom}(D) \subseteq X$. Thus, $L(D)$ is a surjection $X \twoheadrightarrow Y_D$, and the predictor factors through the inclusion $Y_D \hookrightarrow Y$.*

Boundary regimes. *At $|Y| = 2$ the generic case $|Y_D| \leq 0$ is vacuous on $\mathcal{D}_u \setminus \{\emptyset\}$: every nonempty dataset has $|Y_D| \geq 1$ and is extremal, so the constraint theorem has no content. In this case Theorem 38 still applies: every orbit is in the extremal case and contributes a factor $|Y|^{|X|} = 2^{|X|}$ per orbit, giving $N(|X|, 2) = 2^{|X| \cdot |\mathcal{O}|}$ (writing $\mathcal{O}$ for the set of $\mathrm{Sym}(Y)$-orbits in $\mathcal{D}_u \setminus \{\emptyset\}$; formally introduced in Section 5), the maximally unconstrained regime, in which every predictor at every orbit representative is free. This is why $|Y| \geq 3$ is typically assumed: at $|Y| = 2$ label-symmetry imposes no constraint at all on the generic case.*

4 The Rigidity Theorem

The constraint theorem showed label-symmetry *sufficient* for no-fabrication on the generic case. The rigidity theorem of this section identifies it as *necessary*: any weaker symmetry hypothesis admits a universal equivariant learner that fabricates on the generic case.

4.1 G-Equivariance and the Subgroup Constraint

To make "weaker symmetry hypothesis" precise, we generalize label-symmetry to equivariance under an arbitrary subgroup of $\mathrm{Sym}(Y)$.

Definition 28 (G-Equivariance)**.** *Let $G \leq \mathrm{Sym}(Y)$. A learner L is* G-equivariant *if for every $\sigma \in G$ and every $D \in \mathrm{dom}(L)$ with $\sigma D \in \mathrm{dom}(L)$,*

$$L(\sigma D) = \sigma \circ L(D).$$

Label-symmetry (Definition 6) is the special case $G = \mathrm{Sym}(Y)$. The trivial group $G = \{e\}$ imposes no equivariance condition and admits every learner. Subgroups in between impose correspondingly partial equivariance. A G-equivariant learner respects only those relabelings that lie in G.

Both the master constraint and the stabilizer computation generalize to G unchanged. The following is the G-equivariant form; Lemma 18 is the special case $G = \mathrm{Sym}(Y)$, and Lemma 56 in Section 8 is the further generalization to joint H-actions on X and Y.

Lemma 29 (Master Constraint, G-Equivariant Form)**.** *Let L be a G-equivariant learner and $D \in \mathrm{dom}(L)$. For every $x \in X$,*

$$L(D)(x) \in Y^{\mathrm{Stab}_G(D)},$$

where $\mathrm{Stab}_G(D) := \{\sigma \in G \mid \sigma D = D\}$.

Proof Identical to Lemma 18 with $\mathrm{Sym}(Y)$ replaced by G. ■

We write $\mathrm{Fix}_G(D) := Y^{\mathrm{Stab}_G(D)}$ for the joint fixpoint set of $\mathrm{Stab}_G(D)$ in Y, in exact analogy with $\mathrm{Fix}(D) = Y^{\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)}$ of Definition 17; the two notations coincide when $G = \mathrm{Sym}(Y)$. We use $\mathrm{Fix}_G(D)$ freely below where it improves readability, with $Y^{\mathrm{Stab}_G(D)}$ as the unabbreviated form.

Lemma 30 (*G*-Stabilizer on Unambiguous Data)**.** *For unambiguous D and any $G \leq \mathrm{Sym}(Y)$,*

$$\mathrm{Stab}_G(D) = G \cap \mathrm{Sym}(Y \setminus Y_D),$$

the intersection of G with the pointwise stabilizer of Y_D in $\mathrm{Sym}(Y)$.

Proof By Lemma 19, $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D) = \mathrm{Sym}(Y \setminus Y_D)$, embedded in $\mathrm{Sym}(Y)$ as permutations fixing Y_D pointwise. The stabilizer in G is the intersection of this with G. ■

The joint fixpoint set $Y^{\mathrm{Stab}_G(D)}$ always contains Y_D, since the stabilizer acts trivially on Y_D by definition. The question is whether it contains anything beyond Y_D, which would allow a G-equivariant learner to predict an unused label. For $G = \mathrm{Sym}(Y)$, the answer is no on the generic case (Lemma 20). For other G, the answer depends on whether G contains enough permutations to fix unused labels.

4.2 Strongly Constraining Subgroups

We name the property we want to investigate.

Definition 31 (Strongly Constraining Subgroup)**.** *A subgroup $G \leq \mathrm{Sym}(Y)$ is* strongly constraining *if every universal G-equivariant learner L on $\mathcal{D}_u \setminus \{\emptyset\}$ satisfies*

$$L(D)(x) \in Y_D$$

for every $D \in \mathcal{D}_u \setminus \{\emptyset\}$ with $|Y_D| \leq |Y| - 2$ and every $x \in X$.

By Theorem 21, $\mathrm{Sym}(Y)$ is strongly constraining. The rigidity theorem will say no other subgroup is.

4.3 Group-Theoretic Reformulation

The bridge from the learner-level statement to a group-theoretic condition on G is the following lemma. It is the technical heart of the rigidity theorem.

Lemma 32 (Group-Theoretic Characterization)**.** *Let $G \leq \mathrm{Sym}(Y)$ and $|Y| \geq 3$. Then are equivalent:*

(i) *G is strongly constraining.*

(ii) *For every subset $T \subseteq Y$ with $1 \leq |T| \leq |Y| - 2$, the intersection $G \cap \mathrm{Sym}(Y \setminus T)$ has no global fixed point in $Y \setminus T$: no element of $Y \setminus T$ is fixed by every $\sigma \in G \cap \mathrm{Sym}(Y \setminus T)$.*

The condition (ii) says: for every label support T of intermediate size, the permutations of $Y \setminus T$ lying in G have no joint fixed point in $Y \setminus T$; this does not mean the action is free.

Proof *(ii) ⇒ (i).* Suppose (ii) holds. Let L be a universal G-equivariant learner and let $D \in \mathcal{D}_u \setminus \{\emptyset\}$ with $|Y_D| \leq |Y| - 2$. Set $T := Y_D$. Then it holds that $1 \leq |T| \leq |Y| - 2$ since D is nonempty. By Lemma 29, $L(D)(x) \in Y^{\mathrm{Stab}_G(D)}$, and by Lemma 30, $\mathrm{Stab}_G(D) = G \cap \mathrm{Sym}(Y \setminus T)$.

The joint fixpoint set $Y^{\mathrm{Stab}_G(D)}$ trivially contains $T = Y_D$. By (ii), $G \cap \mathrm{Sym}(Y \setminus T)$ has no global fixed point in $Y \setminus T$, so no element of $Y \setminus T$ lies in $Y^{\mathrm{Stab}_G(D)}$. Hence, $Y^{\mathrm{Stab}_G(D)} = T = Y_D$, and $L(D)(x) \in Y_D$.

(i) ⇒ (ii). Suppose (ii) fails: there exist $T \subseteq Y$ with $1 \leq |T| \leq |Y| - 2$ and $y_0 \in Y \setminus T$ fixed by every $\sigma \in G \cap \mathrm{Sym}(Y \setminus T)$. Pick a nonempty unambiguous dataset D with $Y_D = T$ (possible since X is nonempty and $|T| \geq 1$), so $|Y_D| \leq |Y| - 2$. Define $L_0(D)(x) := y_0$ for every x; by hypothesis $y_0 \in Y^{\mathrm{Stab}_G(D)}$, so this is consistent with the master constraint (Lemma 29). Extend L_0 to the rest of the G-orbit of D by G-equivariance, and to other G-orbits arbitrarily; the equivariant extension is always possible by Theorem 33 below. Then L_0 is universal and G-equivariant, but $L_0(D)(x) = y_0 \notin T = Y_D$, contradicting strong constraint. ■

The (ii) $\Rightarrow$ (i) direction is the master constraint applied to G. The (i) $\Rightarrow$ (ii) direction is constructive: when (ii) fails, the proof exhibits an explicit witness learner. We collect this construction as Corollary 36 after the rigidity theorem itself. The construction relies on the characterization theorem for G, which we record next. We write $[D]_G$ for the G-orbit of D in $\mathcal{D}_u \setminus \{\emptyset\}$.

Theorem 33 (Characterization Theorem, G-Equivariant Version)**.** *The space of universal G-equivariant learners on $\mathcal{D}_u \setminus \{\emptyset\}$ is in canonical bijection with*

$$\prod_{[D]_G} \left(Y^{\mathrm{Stab}_G(D)}\right)^{|X|},$$

where the product runs over G-orbits and D denotes the chosen representative of each orbit.

Proof We construct maps in both directions and verify they are inverse.

($\Rightarrow$) Given a universal G-equivariant learner L, define the tuple $(L(D))_{[D]_G}$ of predictors at chosen G-orbit representatives. By Lemma 29, $L(D)(x) \in Y^{\mathrm{Stab}_G(D)}$ for every $x \in X$, so $L(D) \in (Y^{\mathrm{Stab}_G(D)})^{|X|}$ and the tuple lies in the indicated product.

($\Leftarrow$) Given a tuple $(f_{[D]_G})$ with each $f_{[D]_G} \in (Y^{\mathrm{Stab}_G(D)})^{|X|}$, define L on any $D' \in \mathcal{D}_u \setminus \{\emptyset\}$ as follows. The group G acts on $\mathcal{D}_u \setminus \{\emptyset\}$ via the relabeling action of $\mathrm{Sym}(Y)$ restricted to G, partitioning $\mathcal{D}_u \setminus \{\emptyset\}$ into G-orbits; every D' therefore lies in a unique orbit $[D']_G$ with a chosen representative D and at least one $\sigma \in G$ satisfying $\sigma D = D'$. Set

$$L(D')(x) := \sigma \cdot f_{[D]_G}(x) \quad \text{for every } x \in X.$$

Well-definedness: if $\sigma D = \sigma' D = D'$, then $\sigma'^{-1}\sigma \in \mathrm{Stab}_G(D)$, and since $f_{[D]_G}(x) \in Y^{\mathrm{Stab}_G(D)}$ we have $\sigma'^{-1}\sigma \cdot f_{[D]_G}(x) = f_{[D]_G}(x)$, hence $\sigma \cdot f_{[D]_G}(x) = \sigma' \cdot f_{[D]_G}(x)$. Universality holds because the construction defines L on every $D' \in \mathcal{D}_u \setminus \{\emptyset\}$: the orbit representative D exists for every D' by the orbit partition, and the value $L(D')(x) := \sigma \cdot f_{[D]_G}(x)$ is defined for every $x \in X$. Equivariance under G holds because for any $\rho \in G$ and any D' with $\sigma D = D'$, we have $\rho\sigma D = \rho D'$, and

$$L(\rho D')(x) = (\rho\sigma) \cdot f_{[D]_G}(x) = \rho \cdot (\sigma \cdot f_{[D]_G}(x)) = \rho \cdot L(D')(x).$$

It remains to show that the two maps are mutually inverse. The forward map followed by the backward map recovers L: for D' with $\sigma D = D'$, the backward map gives $\sigma \cdot L(D) = L(\sigma D) = L(D')$ by G-equivariance of L. The backward map followed by the forward map recovers $(f_{[D]_G})$: at each representative D, the backward map constructs L with $L(D) = \mathrm{id} \cdot f_{[D]_G} = f_{[D]_G}$, and the forward map applied to L returns $f_{[D]_G}$. ■

Remark 34 (Non-Circularity of the Dependency Chain)**.** *The well-definedness verification above uses only the master constraint (Lemma 29) applied to the chosen orbit representative, together with the orbit structure of G acting on $\mathcal{D}_u \setminus \{\emptyset\}$; it does not invoke Lemma 32. The dependency chain is therefore: Theorem 33 rests on Lemma 29 and the orbit structure of G; Lemma 32 then invokes Theorem 33 in its (i)$\Rightarrow$(ii) direction; and Theorem 35 (the rigidity theorem) invokes Lemma 32. The use of Theorem 33 inside the proof of Lemma 32 is consequently non-circular.*

Theorem 38 in Section 5 is the special case $G = \mathrm{Sym}(Y)$; its proof is structurally identical to the one just given. We prove the general version here because it is needed for the rigidity witness construction, and state the special case in its natural position, namely after the spine theorems.

4.4 The Rigidity Theorem

The group-theoretic characterization (Lemma 32) reduces strong constraint to a fixpoint condition on subgroups; we now extract from it the rigidity statement itself.

Theorem 35 (Rigidity of Label-Symmetry)**.** *Let $|Y| \geq 3$. A subgroup $G \leq \mathrm{Sym}(Y)$ is strongly constraining if and only if $G = \mathrm{Sym}(Y)$.*

Proof The forward direction, $G = \mathrm{Sym}(Y) \Rightarrow G$ strongly constraining, is Theorem 21.

For the reverse, suppose G is strongly constraining. By Lemma 32, condition (ii) holds: for every $T \subseteq Y$ with $1 \leq |T| \leq |Y| - 2$, the intersection $G \cap \mathrm{Sym}(Y \setminus T)$ has no global fixed point in $Y \setminus T$.

It suffices to consider the case $|T| = |Y| - 2$: a group containing every transposition of $\mathrm{Sym}(Y)$ already satisfies condition (ii) for every smaller T, since transpositions generate $\mathrm{Sym}(Y)$ and the full symmetric group on $Y \setminus T$ has no global fixed point whenever $|Y \setminus T| \geq 2$.

We specialize to $|T| = |Y| - 2$, so $|Y \setminus T| = 2$. Write $Y \setminus T = \{a, b\}$ with $a \neq b$. The symmetric group $\mathrm{Sym}(\{a, b\}) \cong \mathbb{Z}/2$ has exactly two elements: the identity, which fixes both a and b, and the transposition $(a\ b)$, which fixes neither. For $G \cap \mathrm{Sym}(\{a, b\})$ to have no global fixed point in $\{a, b\}$, it must contain the transposition $(a\ b)$.

This holds for every pair $\{a, b\} \subseteq Y$ (since every such pair arises as $Y \setminus T$ for some T of size $|Y| - 2$). Hence, G contains every transposition of $\mathrm{Sym}(Y)$. The transpositions generate $\mathrm{Sym}(Y)$, so $G \supseteq \mathrm{Sym}(Y)$. Combined with $G \leq \mathrm{Sym}(Y)$, this gives $G = \mathrm{Sym}(Y)$. ■

4.5 Witness Learners for Weaker Symmetries

The proof of Lemma 32 (i) $\Rightarrow$ (ii) is constructive: when G is a proper subgroup of $\mathrm{Sym}(Y)$, the construction produces an explicit fabricating learner. We record the construction concretely.

Corollary 36 (Fabricating Witness for Proper Subgroups)**.** *Let $|Y| \geq 3$ and $G \subsetneq \mathrm{Sym}(Y)$. Pick a transposition $(a\ b) \in \mathrm{Sym}(Y) \setminus G$, set $T := Y \setminus \{a, b\}$, and let D be any nonempty unambiguous dataset with $Y_D = T$. The learner L_0 with $L_0(D)(x) := a$ for every x, extended by G-equivariance to $[D]_G$ and arbitrarily to other G-orbits, is universal and G-equivariant; since $a \notin Y_D$, it fabricates on the generic dataset D.*

Proof A transposition $(a\ b) \notin G$ exists since the transpositions generate $\mathrm{Sym}(Y)$ and G is proper. Then $G \cap \mathrm{Sym}(\{a, b\}) = \{e\}$ (since $(a\ b) \notin G$ by choice of $(a\ b)$, and $\mathrm{Sym}(\{a, b\}) = \{e, (a\ b)\}$ has no other non-identity element), so a is fixed by every element of $G \cap \mathrm{Sym}(Y \setminus T)$ and thus lies in $Y^{\mathrm{Stab}_G(D)}$. By Lemma 29, $L_0(D)(x) = a$ is valid, and G-equivariance extends L_0 to a universal learner, since by Theorem 33 the master constraint is satisfiable on every orbit. ■

4.6 Examples and Verification

Three concrete subgroups at $|Y| = 4$, $Y = \{y_1, y_2, y_3, y_4\}$, illustrate the dichotomy. The full symmetric group $\mathrm{Sym}(Y)$ is strongly constraining and admits no fabricating witness. The cyclic subgroup $\langle (y_1\ y_2\ y_3\ y_4) \rangle \cong \mathbb{Z}/4$ is not: $(y_1\ y_2) \notin \mathbb{Z}/4$, so taking $T := \{y_3, y_4\}$, Corollary 36 produces a $\mathbb{Z}/4$-equivariant learner predicting y_1 at every input on a dataset with $Y_D = \{y_3, y_4\}$. The Young subgroup $\mathrm{Sym}(\{y_1, y_2\}) \times \mathrm{Sym}(\{y_3, y_4\})$ is likewise not strongly constraining: $(y_1\ y_3)$ is missing, and taking $T := \{y_2, y_4\}$ produces a Young-equivariant learner predicting y_1 on a dataset with $Y_D = \{y_2, y_4\}$.

We verified the theorem by exhaustive enumeration at $|Y| = 4$ across all 30 subgroups of Sym(4): only Sym(4) itself is strongly constraining, and every proper subgroup admits a fabricating witness via Corollary 36.

The constraint theorem and rigidity theorem together characterize full label-symmetry by its no-fabrication consequence: label-symmetry is exactly what implies $L(D)(x) \in Y_D$ on the generic case, and any strictly weaker symmetry admits a fabricating witness. This clarifies the role of the no-fabrication axiom. Under full label-symmetry, no-fabrication is redundant on the generic case (Corollary 23); under any proper subgroup $G \subsetneq \mathrm{Sym}(Y)$, no-fabrication regains independent content, since the rigidity witnesses produce fabricating G-equivariant learners. The collapse of no-fabrication into label-symmetry is specific to the full symmetric group.

Two caveats deserve mention. The theorem assumes $|Y| \geq 3$; at $|Y| = 2$ the generic case $|Y_D| \leq 0$ is vacuous on $\mathcal{D}_u \setminus \{\emptyset\}$, and strong constraint holds vacuously for every subgroup. The theorem is also stated on the full domain $\mathcal{D}_u \setminus \{\emptyset\}$; restricting the domain (for instance to total datasets or single-label datasets) can make smaller subgroups strongly constraining, since fewer generic datasets need to be checked.

5 The Characterization Theorem

The constraint and rigidity theorems together characterize universal label-symmetric learners by their no-fabrication consequence on the generic case. The characterization theorem of this section completes the picture by exhibiting the entire space of such learners as an explicit product over $\mathrm{Sym}(Y)$-orbits, with a closed-form count. A general G-equivariant version (Theorem 33, proved in Section 4 as part of the rigidity apparatus) was established earlier because the rigidity witness construction requires it; the theorem of this section is its $\mathrm{Sym}(Y)$-specialization, stated here in its natural position after the spine theorems and with its proof reduced to a one-line deferral to the general version. The practical motivation is the aggregation problem: to combine multiple universal label-symmetric learners, one needs to know which choices are available for each, and the orbit-product structure makes those choices explicit.

5.1 Orbits and Orbit Representatives

The relabeling action of $\mathrm{Sym}(Y)$ partitions $\mathcal{D}_u \setminus \{\emptyset\}$ into orbits. Two datasets are in the same orbit if and only if one is obtained from the other by applying a single $\sigma \in \mathrm{Sym}(Y)$.

Lemma 37 (Orbit Structure)**.** *The $\mathrm{Sym}(Y)$-orbits in $\mathcal{D}_u \setminus \{\emptyset\}$ correspond bijectively to pairs (S, π), where $S \subseteq X$ is a nonempty subset and π is a set-partition of S into nonempty blocks of total count at most $|Y|$. We call this data the* shape *of any dataset in the orbit.*

Proof An unambiguous dataset $D \in \mathcal{D}_u \setminus \{\emptyset\}$ is determined by its input domain $S = \mathrm{dom}(D)$, the partition of S into preimages $\mathrm{dom}(D, y) := \{x \in S \mid D(x) = y\}$ for $y \in Y_D$, and the specific assignment of labels in Y_D to blocks. Two such datasets differ by a choice of $\sigma \in \mathrm{Sym}(Y)$ if and only if they share the input domain and the set-partition, differing only in label assignments. Hence, orbits are in bijection with (S, π) where π has $|Y_D| \leq |Y|$ blocks. ■

The shape is the combinatorial data of an orbit, independent of specific label identities. We write $\mathcal{O}$ for the set of orbits and choose, once and for all, a representative D for each orbit.

5.2 The Characterization Theorem

Theorem 38 (Characterization Theorem)**.** *The space of universal label-symmetric learners on $\mathcal{D}_u \setminus \{\emptyset\}$ is in canonical bijection with the product*

$$\prod_{[D] \in \mathcal{O}} \big(\mathrm{Fix}(D)\big)^{|X|},$$

over $\mathrm{Sym}(Y)$*-orbits of nonempty unambiguous datasets, where* D *is the chosen representative of each orbit. The bijection sends a learner* L *to the tuple* $\big(L(D)\big)_{[D]\in\mathcal{O}}$ *of predictors at the chosen representatives.*

Proof This is the special case $G = \mathrm{Sym}(Y)$ of Theorem 33. The G-orbit of D becomes the $\mathrm{Sym}(Y)$-orbit $[D] \in \mathcal{O}$, the stabilizer $\mathrm{Stab}_G(D)$ becomes $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)$ as identified in Lemma 19, and the fixpoint set $Y^{\mathrm{Stab}_G(D)}$ becomes $\mathrm{Fix}(D)$ as in Definition 17. ■

The proof is the orbit-decomposition principle for $\mathrm{Sym}(Y)$-equivariant set maps, specialized to our setting. Indeed, the key fact is Lemma 19, which identifies the stabilizers explicitly. The bijection is canonical in the sense that, although it is presented relative to a fixed choice of orbit representatives, transport along the $\mathrm{Sym}(Y)$-action carries the bijection consistently to any other choice; this is the standard well-definedness check for orbit-decomposition formulas.

5.3 The Size of the Parameter Space

The characterization theorem gives an explicit formula for the size of the parameter space.

Corollary 39 (Counting Universal Label-Symmetric Learners)**.** *The number of universal label-symmetric learners on* $\mathcal{D}_u \setminus \{\emptyset\}$ *is*

$$N(X,Y) := \prod_{[D]\in\mathcal{O}} \big|\mathrm{Fix}(D)\big|^{|X|}.$$

By Lemma 20, the factor for an orbit with $|Y_D| = m$ *is*

$$\big|\mathrm{Fix}(D)\big|^{|X|} = \begin{cases} m^{|X|} & \text{if } m \le |Y| - 2, \\ |Y|^{|X|} & \text{if } m \ge |Y| - 1. \end{cases}$$

The count splits the orbits into generic ($m \le |Y| - 2$) and extremal ($m \ge |Y| - 1$) contributions. Indeed, generic orbits contribute $m^{|X|}$, the size of Y_D^X. Extremal orbits contribute $|Y|^{|X|}$, the size of the full predictor space Y^X. The generic-versus-extremal dichotomy of the constraint theorem becomes a quantitative split of the parameter space.

5.4 Closed-Form Count via Stirling Numbers

The orbit count itself has a closed form. Recall the Stirling number of the second kind, $S(n,k)$, which counts the number of partitions of an n-element set into exactly k nonempty blocks.

Lemma 40 (Number of Orbits)**.** *The number of* $\mathrm{Sym}(Y)$*-orbits in* $\mathcal{D}_u \setminus \{\emptyset\}$ *with input domain size* s *and label support size* m *is*

$$\binom{|X|}{s} \cdot S(s,m),$$

provided $1 \le m \le \min(s, |Y|)$*. The total number of orbits is*

$$|\mathcal{O}| = \sum_{s=1}^{|X|} \binom{|X|}{s} \sum_{m=1}^{\min(s,|Y|)} S(s,m).$$

Proof By Lemma 37, each orbit corresponds to a pair (S,π) with $S \subseteq X$ nonempty and π a partition of S into $\le |Y|$ blocks. For fixed $s = |S|$, the choices of S are $\binom{|X|}{s}$. For fixed S, the partitions of S into exactly m nonempty blocks are counted by $S(s,m)$. Summing over m gives the count of partitions into at most $\min(s,|Y|)$ blocks. ■

Remark 41 (Bell Numbers in the Saturated Regime)**.** *When $|Y| \geq |X|$, every set-partition of $S \subseteq X$ has at most $|S| \leq |X| \leq |Y|$ blocks, so the constraint $m \leq |Y|$ is automatically satisfied and the inner sum becomes*

$$\sum_{m=1}^{s} S(s,m) \;=\; B_s,$$

the s-th Bell number, which counts all set-partitions of an s-element set. The total orbit count then simplifies to $|\mathcal{O}| = \sum_{s=1}^{|X|} \binom{|X|}{s} B_s$, a closed form in Bell numbers alone. This is the regime advertised in the abstract; the Stirling formulation in the lemma is the more general version covering the generic case $|Y| < |X|$ as well.

Combining the orbit count with the size formula gives a closed form for the total number of universal label-symmetric learners.

Corollary 42 (Closed-Form Count)**.** *The number of universal label-symmetric learners on $\mathcal{D}_u \setminus \{\emptyset\}$ is*

$$N(|X|,|Y|) \;=\; \prod_{s=1}^{|X|} \prod_{m=1}^{\min(s,|Y|)} \kappa(m,|Y|)^{\,e(s,m)}, \qquad e(s,m) \;:=\; |X| \cdot \binom{|X|}{s} \cdot S(s,m),$$

where the factor base is

$$\kappa(m,n) \;:=\; \begin{cases} m & \text{if } m \leq n-2, \\ n & \text{if } m \geq n-1. \end{cases}$$

Equivalently, each of the $\binom{|X|}{s} \cdot S(s,m)$ orbits of shape (s,m) contributes a factor $\kappa(m,|Y|)^{|X|}$, and the formula aggregates these identical contributions across all shapes.

Proof Combine Lemma 40 with Corollary 39. By Corollary 39, the factor contributed by an orbit depends only on its shape parameter $m = |Y_D|$, not on the specific orbit, so all $\binom{|X|}{s} \cdot S(s,m)$ orbits of shape (s,m) contribute the same factor base $\kappa(m,|Y|)^{|X|}$. Aggregating these identical factors gives a single combined factor with exponent $|X| \cdot \binom{|X|}{s} \cdot S(s,m)$, as displayed. ■

5.5 Worked Small Cases

We compute the formula for small $|X|, |Y|$ to illustrate.

Example 2 ($|X| = 2$, $|Y| = 2$)**.** *The orbits in $\mathcal{D}_u \setminus \{\emptyset\}$ are: two singleton orbits (one per input, $s = 1, m = 1$; the two are distinct because $\mathrm{Sym}(Y)$ acts only on labels, so it cannot relate singletons at different inputs); one single-label two-input orbit ($s = 2, m = 1$); one two-label orbit ($s = 2, m = 2$). Counts: $\binom{2}{1} S(1,1) = 2$ for $s = 1$, $\binom{2}{2} S(2,1) = 1$ for $s = 2, m = 1$, $\binom{2}{2} S(2,2) = 1$ for $s = 2, m = 2$, totalling 4 orbits.*

For $|Y| = 2$: every orbit is in the extremal case since $m \in \{1,2\} \geq |Y| - 1 = 1$. The factor per orbit is $|Y|^{|X|} = 2^2 = 4$. The total is

$$N(2,2) = 4^4 = 256.$$

Example 3 ($|X| = 2$, $|Y| = 3$)**.** *Same four orbits as above. For $|Y| = 3$: singleton orbits and single-label orbits have $m = 1 = |Y| - 2$, the generic boundary, contributing $m^{|X|} = 1^2 = 1$ per orbit. The two-label orbit has $m = 2 = |Y| - 1$, the extremal case, contributing $|Y|^{|X|} = 3^2 = 9$. The total is*

$$N(2,3) = 1^3 \cdot 9 = 9.$$

This matches the explicit enumeration: at $|X| = 2, |Y| = 3$, the universal label-symmetric learners are determined by their single choice on the two-label orbit, of which there are $3^2 = 9$ possibilities.

Example 4 ($|X| = 2$, $|Y| = 4$)**.** *Same four orbits. For $|Y| = 4$, every orbit has $m \leq 2 = |Y| - 2$, so all orbits are in the generic case. Factors $1^2 = 1$ for each of the three single-label orbits and $2^2 = 4$ for the two-label orbit give total $N(2, 4) = 1^3 \cdot 4 = 4$. Computational verification: at $|X| = 2, |Y| = 4$, only four universal label-symmetric learners exist, all of which agree on single-label data (broadcasting the single label) and differ only in their choice between the two used labels on two-label data.*

The example shows the qualitative dynamics. At $|Y| = 2$, every orbit is extremal, and the parameter space is maximally large ($|Y|^{|X|}$ per orbit). At $|Y| \geq 4$, the non-extremal orbits dominate, and the parameter space contracts substantially. The transition happens at $|Y| = 3$, where the smallest nontrivial extremal case appears (the two-label data on three labels).

5.6 Consistency and the Count of Consistent Learners

The characterization theorem (Theorem 38) parametrizes universal label-symmetric learners orbit by orbit, with no consistency requirement. Imposing consistency restricts each orbit factor: on a generic orbit $[D]$, the constraint theorem yields $L(D)(x) \in Y_D$ for every x, and consistency further pins $L(D) \equiv D$ on $\mathrm{dom}(D)$, so the orbit factor shrinks from the free cube $Y_D^{|X|}$ to its consistent slice

$$\mathcal{F}(D) := \left\{ f \colon X \to Y_D \;\middle|\; f|_{\mathrm{dom}(D)} = D|_{\mathrm{dom}(D)} \right\},$$

of cardinality $|Y_D|^{|X \setminus \mathrm{dom}(D)|}$. On an extremal orbit the fixpoint set is all of Y and consistency restricts only on $\mathrm{dom}(D)$, leaving $|Y|^{|X \setminus \mathrm{dom}(D)|}$ free choices per orbit. One sub-case of the generic restriction is especially clean.

Corollary 43 (Total-Data Uniqueness)**.** *Let D be a total dataset, that is, $\mathrm{dom}(D) = X$, in the generic case. Every universal label-symmetric consistent learner satisfies $L(D) = D$.*

Proof Consistency on the total domain gives $L(D)(x) = D(x)$ for every $x \in \mathrm{dom}(D) = X$, hence $L(D) = D$ as functions $X \to Y$. The constraint theorem is consistent with this and adds no restriction beyond what consistency already forces. ■

The total generic case is exactly the regime in which the learner has nothing to do beyond memorization, and the corollary records that the framework forces this. Indeed, different universal label-symmetric consistent learners (such as nearest-neighbor procedures with different distance functions) must agree on every total dataset on which fewer than $|Y|-1$ labels appear. Aggregating the orbit factors gives the closed-form count.

Corollary 44 (Count of Universal Label-Symmetric Consistent Learners)**.** *Let X and Y be finite nonempty sets, of inputs and labels respectively, and write $\mathcal{O}_{\mathrm{gen}}$ and $\mathcal{O}_{\mathrm{ext}}$ for the sets of generic and extremal $\mathrm{Sym}(Y)$-orbits in $\mathcal{D}_u \setminus \{\emptyset\}$, partitioning $\mathcal{O} = \mathcal{O}_{\mathrm{gen}} \sqcup \mathcal{O}_{\mathrm{ext}}$. The number of universal label-symmetric consistent learners on $\mathcal{D}_u \setminus \{\emptyset\}$ is*

$$N_{\mathsf{C}}(|X|, |Y|) \;=\; \prod_{[D] \in \mathcal{O}_{\mathrm{gen}}} |Y_D|^{|X \setminus \mathrm{dom}(D)|} \;\times\; \prod_{[D] \in \mathcal{O}_{\mathrm{ext}}} |Y|^{|X \setminus \mathrm{dom}(D)|}.$$

Proof Multiply the orbit factors from above: each generic orbit contributes $|\mathcal{F}(D)| = |Y_D|^{|X \setminus \mathrm{dom}(D)|}$, and each extremal orbit contributes $|Y|^{|X \setminus \mathrm{dom}(D)|}$. ■

The pair $(N(|X|, |Y|), N_{\mathsf{C}}(|X|, |Y|))$ makes the effect of consistency quantitative. At $|X| = 2$, $|Y| = 4$ every orbit is generic, $\mathcal{F}(D)$ has cardinality 1 on each orbit since $\mathrm{dom}(D) = X$ throughout, and so $N = 4$ collapses to $N_{\mathsf{C}} = 1$: consistency eliminates all freedom.

6 Classical Impossibilities as Corollaries

The three theorems of the previous sections give a complete structural account of universal label-symmetric learners. The exceptional phenomena of axiomatic learning theory live on the extremal case; the generic case is fully constrained. Classical impossibility results follow as corollaries, each an application of the master constraint to a specific dataset and group action. We work through each in turn.

6.1 The Locality Impossibility

Strict locality (Definition 10) requires, for every unseen input $x \notin \mathrm{dom}(D)$, that $L(D)(x) = L(\emptyset)(x)$ and that $\emptyset \in \mathrm{dom}(L)$. The learner's prediction at an unseen input therefore depends only on the empty dataset; in particular, L must be defined on the empty dataset.

Theorem 45 (Locality Impossibility)**.** *If $|Y| \geq 2$, no learner is simultaneously universal, label-symmetric, and strictly local on any domain containing the empty dataset. More generally, no learner is universal, label-symmetric, and strictly local on a domain containing a dataset D with $\mathrm{dom}(D) \neq X$, since the locality axiom on such D requires L to be defined on the empty dataset.*

Proof Suppose $\emptyset \in \mathrm{dom}(L)$. The stabilizer of $\emptyset$ in $\mathrm{Sym}(Y)$ is all of $\mathrm{Sym}(Y)$ (every permutation fixes the empty multiset). By Lemma 18, $L(\emptyset)(x) \in Y^{\mathrm{Sym}(Y)}$. For $|Y| \geq 2$, $Y^{\mathrm{Sym}(Y)}$ is empty: no element of Y is fixed by every permutation. This contradicts the existence of $L(\emptyset)(x)$.

For the general statement: if L is universal and strictly local, then for any $D \in \mathrm{dom}(L)$ with $x \notin \mathrm{dom}(D)$, strict locality requires $\emptyset \in \mathrm{dom}(L)$ and $L(D)(x) = L(\emptyset)(x)$. The first requirement places $\emptyset$ in $\mathrm{dom}(L)$, which then reduces to the first case; alternatively, $\mathrm{dom}(L)$ consists only of total datasets, which violates universality. ■

6.2 The Double-Symmetry Impossibility

Joint equivariance under both $\mathrm{Sym}(X)$ and $\mathrm{Sym}(Y)$ produces a larger stabilizer on suitable datasets, with correspondingly fewer fixed points.

Theorem 46 (Double-Symmetry Impossibility)**.** *Suppose $|X| \geq 3$ and that $|Y| = 2$ or $|Y| \geq 4$. Then no learner is simultaneously universal, label-symmetric, and input-symmetric.*

Proof We may pick distinct $x_1, x_2 \in X$ and $x_3 \in X \setminus \{x_1, x_2\}$ (possible since $|X| \geq 3$), and distinct $y_1, y_2 \in Y$. Set

$$D := \{(x_1, y_1), (x_2, y_2)\}.$$

Step 1. By input-symmetry under permutations of $\mathrm{Sym}(X)$ fixing $\{x_1, x_2\}$ pointwise, the value $L(D)(x)$ is the same constant c for every $x \in X \setminus \{x_1, x_2\}$. Explicitly: for any $\tau \in \mathrm{Sym}(X)$ fixing x_1 and x_2 pointwise, input-symmetry gives $L(D)(\tau(x)) = L(D)(x)$ for every x, and the subgroup $\mathrm{Sym}(X \setminus \{x_1, x_2\})$ embedded as such permutations acts transitively on $X \setminus \{x_1, x_2\}$, so $L(D)$ is constant on this set.

Step 2. By label-symmetry under permutations of $\mathrm{Sym}(Y)$ fixing $\{y_1, y_2\}$ pointwise, c is fixed by every such permutation. The pointwise stabilizer of $\{y_1, y_2\}$ in $\mathrm{Sym}(Y)$ is $\mathrm{Sym}(Y \setminus \{y_1, y_2\})$, embedded as permutations fixing y_1 and y_2. Its fixed points in Y are $\{y_1, y_2\}$ (fixed trivially) together with the elements of $Y \setminus \{y_1, y_2\}$ fixed by every permutation of $Y \setminus \{y_1, y_2\}$, which is empty when $|Y \setminus \{y_1, y_2\}| \geq 2$, i.e., when $|Y| \geq 4$. We split on $|Y|$:

- For $|Y| = 2$: $\{y_1, y_2\} = Y$, and the pointwise stabilizer is trivial. Its fixed points are all of Y, so $c \in Y$ is unconstrained at this step.
- For $|Y| = 3$: $Y \setminus \{y_1, y_2\} = \{y_3\}$ is a singleton, so $\mathrm{Sym}(Y \setminus \{y_1, y_2\}) = \mathrm{Sym}(\{y_3\}) = \{e\}$ is trivial, the fixed points are all of Y, so $c \in Y$.

- For $|Y| \geq 4$: pointwise stabilizer is nontrivial, fixed points are $\{y_1, y_2\}$, so $c \in \{y_1, y_2\}$.

Step 3. Let $\tau \in \mathrm{Sym}(X)$ swap x_1, x_2 and fix x_3. Let $\sigma \in \mathrm{Sym}(Y)$ swap y_1, y_2 and fix every other element of Y. Then $\tau \cdot D = \sigma D$, since both are $\{(x_2, y_1), (x_1, y_2)\}$. By input-symmetry,

$$L(\tau \cdot D)(x_3) = L(D)(\tau^{-1}(x_3)) = L(D)(x_3) = c.$$

By label-symmetry, $L(\sigma D)(x_3) = \sigma(L(D)(x_3)) = \sigma(c)$. Equating, $c = \sigma(c)$, so c is fixed by $\sigma = (y_1\ y_2)$. The fixed points of σ in Y are $Y \setminus \{y_1, y_2\}$, so $c \in Y \setminus \{y_1, y_2\}$.

Combining the steps. For $|Y| = 2$: Step 3 gives $c \in Y \setminus \{y_1, y_2\} = \emptyset$. Contradiction.

For $|Y| \geq 4$: Step 2 gives $c \in \{y_1, y_2\}$, while Step 3 gives $c \in Y \setminus \{y_1, y_2\}$. Intersection is empty. Contradiction.

For $|Y| = 3$: Step 2 gives no constraint, Step 3 gives $c \in \{y_3\}$. The conclusion $c = y_3$ is consistent and no impossibility follows. The joint $\mathrm{Sym}(X) \times \mathrm{Sym}(Y)$-equivariance at $|Y| = 3$ implies $L(D)(x_3) = y_3$ but does not produce an impossibility on its own. The existence of a universal label-symmetric input-symmetric learner realizing this assignment at $|Y| = 3$ is established by the missing-label compositional learner of Section 7. ■

6.3 The Ambiguity Impossibility

On the multiset (ambiguous) domain, additional permutations stabilize a dataset by swapping equally-labeled examples. The stabilizer of an ambiguous dataset can be strictly larger than the pointwise stabilizer of Y_D, and the fixpoint set correspondingly smaller.

Theorem 47 (Ambiguity Impossibility)**.** *If $|Y| = 2$ or $|Y| \geq 4$, no learner is simultaneously universal and label-symmetric on any domain that contains a dataset of the form $S = \{(x_0, y_1), (x_0, y_2)\}$ with $y_1 \neq y_2$.*

Proof Any $\sigma \in \mathrm{Sym}(Y)$ fixing the set $\{y_1, y_2\}$ *setwise* sends S to itself, since σ either fixes both pairs (when σ fixes y_1 and y_2 pointwise) or swaps them (when σ swaps y_1 and y_2). The stabilizer of S in $\mathrm{Sym}(Y)$ is thus the setwise stabilizer of $\{y_1, y_2\}$, which is $\mathrm{Sym}(\{y_1, y_2\}) \times \mathrm{Sym}(Y \setminus \{y_1, y_2\})$.

By Lemma 18, $L(S)(x_0) \in Y^{\mathrm{Stab}(S)}$. The joint fixed points of $\mathrm{Sym}(\{y_1, y_2\}) \times \mathrm{Sym}(Y \setminus \{y_1, y_2\})$ in Y are: elements of $\{y_1, y_2\}$ fixed by $\mathrm{Sym}(\{y_1, y_2\})$ (none, since the swap of y_1, y_2 acts nontrivially), and elements of $Y \setminus \{y_1, y_2\}$ fixed by $\mathrm{Sym}(Y \setminus \{y_1, y_2\})$ (none when $|Y \setminus \{y_1, y_2\}| \geq 2$, i.e., when $|Y| \geq 4$).

Thus, for $|Y| = 2$ or $|Y| \geq 4$, the joint fixed set is empty, contradicting the existence of $L(S)(x_0)$. ■

6.4 Unification via the Master Constraint

The three impossibility theorems of this section have been proved separately, each by its own fixpoint computation. We record the observation that they are not independent results: each applies the master constraint at a different triple (D, x, H), with the impossibility following in each case from an empty joint fixpoint set. The compositional trichotomy of Section 7 joins them as a fourth instance, unified through the core lemma at $k = 2$.

Remark 48 (Unification via the Master Constraint)**.** *The three impossibility theorems of §6 differ only in the triple (D, x, H) at which they apply the master constraint; the compositional trichotomy is the fourth instance, unified through the core lemma on symmetric $\mathrm{Sym}(Y)$-equivariant functions on $Y^{(k)}$ at $k = 2$ (Remark 54). The following table records each instantiation. The Trichotomy row is schematic: the impossibility operates at the aggregator Φ rather than at a specific dataset, and the core lemma replaces the direct fixpoint computation.*

Impossibility	D	x	H	*Joint fixpoint set in* Y
Locality	$\emptyset$	*any* $x \in X$	$\mathrm{Sym}(Y)$	$\emptyset$ *for* $\lvert Y\rvert \geq 2$
Double-symmetry	$\{(x_1, y_1), (x_2, y_2)\}$	$x_3 \notin \{x_1, x_2\}$	$\mathrm{Sym}(X) \times \mathrm{Sym}(Y)$	$\emptyset$ *for* $\lvert Y\rvert = 2$ *or* $\lvert Y\rvert \geq 4$
Ambiguity	$\{(x_0, y_1), (x_0, y_2)\}$	x_0	$\mathrm{Sym}(Y)$	$\emptyset$ *for* $\lvert Y\rvert = 2$ *or* $\lvert Y\rvert \geq 4$
Trichotomy	*(aggregator* Φ*, Lemma 52)*	*any unseen* x	$\mathrm{Sym}(Y)$ *on* $Y^{(2)}$	*no* Φ *exists for* $\lvert Y\rvert = 2$ *or* $\lvert Y\rvert \geq 4$ *(via core lemma)*

In each row the master constraint $L(D)(x) \in Y^{\mathrm{Stab}_H(D,x)}$ *(Lemma 56) yields the impossibility when the joint fixpoint set is empty; for the trichotomy, the core lemma supplies the empty-fixpoint conclusion at the level of the aggregator* Φ *itself. Locality applies at* $D = \emptyset$ *with* $H = \mathrm{Sym}(Y)$*, whose stabilizer is all of* $\mathrm{Sym}(Y)$ *and has no fixed points in* Y *for* $\lvert Y\rvert \geq 2$*. Double-symmetry applies at the two-label dataset with* $H = \mathrm{Sym}(X) \times \mathrm{Sym}(Y)$*, where input-symmetry yields a constant prediction at unseen inputs, the diagonal swap excludes* $\{y_1, y_2\}$*, the pointwise stabilizer of* $\{y_1, y_2\}$ *excludes* $Y \setminus \{y_1, y_2\}$ *for* $\lvert Y\rvert \geq 4$*, and the two constraints together leave nothing for* $\lvert Y\rvert = 2$ *or* $\lvert Y\rvert \geq 4$*. Ambiguity applies at the ambiguous dataset* S *with* $H = \mathrm{Sym}(Y)$*, whose setwise stabilizer of* $\{y_1, y_2\}$ *has no fixed points in* Y *for* $\lvert Y\rvert = 2$ *or* $\lvert Y\rvert \geq 4$*. The framework is therefore unified: every impossibility in the paper is one instance of the same fixpoint principle.*

Remark 49 (The $\lvert Y\rvert = 3$ Extremal Exception Is Uniform)**.** *Every impossibility above admits a* $\lvert Y\rvert = 3$ *exception: the missing-label learner of Section 7 is universal, label-symmetric, and input-symmetric (satisfying the double-symmetry hypotheses), and any label-symmetric universal learner is* forced *to predict the missing label* y_3 *on the ambiguous dataset* S *when* $\lvert Y\rvert = 3$*. These are not independent anomalies; they are the same extremal phenomenon in each case. The no-fabrication axiom (Section 6.5) eliminates all three exceptions uniformly, restoring each impossibility for all* $\lvert Y\rvert \geq 2$*.*

6.5 No-Fabrication Revisited

The constraint theorem showed that label-symmetry forces no-fabrication on the generic case; the rigidity theorem showed that label-symmetry is the unique symmetry hypothesis producing this consequence. The no-fabrication axiom NF (Definition 13) is therefore not an independent axiom of the framework, but a derived property whose content depends on regime and symmetry hypothesis.

Proposition 50 (Regime Structure of No-Fabrication)**.** *Let* L *be a universal label-symmetric learner on* $\mathcal{D}_u \setminus \{\emptyset\}$*.*

(i) *On the generic case* $\lvert Y_D\rvert \leq \lvert Y\rvert - 2$*,* NF *is automatic:* $L(D)(x) \in Y_D$ *holds for every* x*.*

(ii) *On the missing-label stratum where* $\lvert Y_D\rvert = \lvert Y\rvert - 1$*,* NF *is content-bearing: at each extremal orbit, the characterization theorem leaves a free choice of* $\lvert Y\rvert^{\lvert X\rvert}$ *predictors, and* NF *restricts this to* $(\lvert Y\rvert - 1)^{\lvert X\rvert}$*.*

(iii) *On the total stratum* $\lvert Y_D\rvert = \lvert Y\rvert$*,* NF *is vacuous, since* $Y_D = Y$*.*

Under G*-equivariance with* $G \subsetneq \mathrm{Sym}(Y)$ *and* $\lvert Y\rvert \geq 3$*, item (i) fails: by Corollary 36, there exists a universal* G*-equivariant learner violating* NF *on some generic dataset. Under proper symmetry,* NF *regains independent content on every regime.*

Proof Item (i) is Theorem 21. Item (ii) is Theorem 38: when $\lvert Y_D\rvert = \lvert Y\rvert - 1$, the set $Y \setminus Y_D$ is a singleton, so the relabeling stabilizer $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D) = \mathrm{Sym}(Y \setminus Y_D)$ is trivial, $\mathrm{Fix}(D) = Y$, and the predictor at the orbit representative is free; NF then restricts it to Y_D^X. Item (iii) is immediate, since $\lvert Y_D\rvert = \lvert Y\rvert$ gives $Y_D = Y$. The G-equivariant statement is Corollary 36. ■

One further observation motivated the exclusion of the empty dataset from the domain throughout the paper.

Proposition 51 (No-Fabrication on the Empty Dataset)**.** *For $|Y| \geq 2$, no learner L satisfies universality including $\emptyset$, label-symmetry, and* NF *on $\emptyset$ simultaneously.*

Proof NF on $\emptyset$ requires $L(\emptyset)(x) \in Y_\emptyset$. But $Y_\emptyset = \emptyset$ vacuously (the empty dataset has no examples, hence no labels appearing), so no value is available, contradiction. ■

Together, the two propositions locate the content of NF precisely: it concentrates on the missing-label stratum $|Y_D| = |Y| - 1$ where the constraint theorem is silent, becomes content-bearing throughout the generic case under any proper subgroup, and clashes outright with universality plus label-symmetry on the empty dataset.

7 The Core Lemma and the Compositional Trichotomy

The compositional trichotomy is the structural anomaly at $|Y| = 3$: the unique cardinality at which a universal, label-symmetric, coordinate-wise compositional learner can exist nontrivially. The same combinatorial fact that drives the trichotomy, a counting lemma for symmetric $\mathrm{Sym}(Y)$-equivariant functions on tuples of distinct labels, governs the extremal case more generally. We prove the lemma first, then apply it to the trichotomy, and finally show how it produces the recurring $|Y| = k + 1$ exception across multiple axioms.

7.1 The Core Lemma on Symmetric Equivariant Functions

Let $Y^{(k)}$ denote the set of k-tuples of pairwise distinct elements of Y. A function $f\colon Y^{(k)} \to Y$ is *symmetric* if $f(y_{\pi(1)}, y_{\pi(2)}, \ldots, y_{\pi(k)}) = f(y_1, y_2, \ldots, y_k)$ for every permutation π of $\{1, 2, \ldots, k\}$, and $\mathrm{Sym}(Y)$*-equivariant* if $f(\sigma(y_1), \sigma(y_2), \ldots, \sigma(y_k)) = \sigma(f(y_1, y_2, \ldots, y_k))$ for every $\sigma \in \mathrm{Sym}(Y)$.

Lemma 52 (Core Lemma)**.** *Let $|Y| = n$ and $k \geq 2$. The number of symmetric* $\mathrm{Sym}(Y)$*-equivariant functions $f\colon Y^{(k)} \to Y$ is:*

(i) *exactly one (the empty function) if $n < k$, since $Y^{(k)} = \emptyset$;*

(ii) *zero if $n = k$;*

(iii) *exactly one if $n = k + 1$, the* missing-label function *$f(y_1, y_2, \ldots, y_k) =$ the unique element of $Y \setminus \{y_1, y_2, \ldots, y_k\}$;*

(iv) *zero if $n \geq k + 2$.*

Proof If $n < k$, the source set $Y^{(k)}$ is empty, so there is a unique function from the empty set to Y. We assume $n \geq k$ for the rest of the proof.

Fix a k-tuple $(y_1, y_2, \ldots, y_k) \in Y^{(k)}$ and set $T := \{y_1, y_2, \ldots, y_k\}$, a k-element subset of Y. Symmetry of f means $f(y_1, y_2, \ldots, y_k)$ depends only on T, not on the order. We may write $f = \hat{f} \circ \pi$ for a function $\hat{f}$ defined on k-subsets of Y and the forgetful projection π from tuples to underlying sets. Equivariance of f translates to

$$\hat{f}(\sigma(T)) = \sigma(\hat{f}(T)) \quad \text{for every } \sigma \in \mathrm{Sym}(Y).$$

Note that $\mathrm{Sym}(Y)$ acts transitively on the set of k-subsets of Y, so $\hat{f}$ is determined by its value at a single representative T_0. This value must lie in the joint fixpoint set in Y of the setwise stabilizer of T_0 in $\mathrm{Sym}(Y)$. The setwise stabilizer is $\mathrm{Sym}(T_0) \times \mathrm{Sym}(Y \setminus T_0)$ (factor on each side).

A fixed point of $\mathrm{Sym}(T_0)$ in Y lies outside T_0 in case $|T_0| = k \geq 2$, since $\mathrm{Sym}(T_0)$ has transitive action on T_0. A fixed point of $\mathrm{Sym}(Y \setminus T_0)$ in Y lies in T_0 when $|Y \setminus T_0| \geq 2$, and is unconstrained when $|Y \setminus T_0| \leq 1$. The joint fixed set is the intersection.

Three cases.

If $n = k$, we have $Y \setminus T_0 = \emptyset$, so any fixed point lies in T_0. Since $k \geq 2$, $\mathrm{Sym}(T_0)$ acts transitively on the at-least-two-element set T_0, so no element of T_0 is fixed by $\mathrm{Sym}(T_0)$. The intersection is empty.

If $n = k + 1$, we have $|Y \setminus T_0| = 1$, say $Y \setminus T_0 = \{y^*\}$. Being fixed by $\mathrm{Sym}(T_0)$ requires $y = y^*$. Being fixed by $\mathrm{Sym}(\{y^*\}) = \{e\}$ is vacuous. The intersection is $\{y^*\}$, the unique missing element.

If $n \geq k + 2$, we have $|Y \setminus T_0| \geq 2$. Being fixed by $\mathrm{Sym}(T_0)$ requires $y \notin T_0$. Being fixed by $\mathrm{Sym}(Y \setminus T_0)$ requires $y \notin Y \setminus T_0$, since $|Y \setminus T_0| \geq 2$ and $\mathrm{Sym}(Y \setminus T_0)$ acts transitively on $Y \setminus T_0$; equivalently, $y \in T_0$. The two conditions together require y both inside and outside T_0, so the intersection is empty.

The count of valid functions equals the count of valid values at T_0, giving the lemma. ■

7.2 Statement and Setup

Recall that coordinate-wise compositionality (Definition 15) requires an aggregator $\Phi\colon Y \times Y \to Y$ such that for disjoint $D_1, D_2 \in \mathcal{D}_u \setminus \{\emptyset\}$ with $D_1 \cup D_2 \in \mathrm{dom}(L)$, and any $x \notin \mathrm{dom}(D_1) \cup \mathrm{dom}(D_2)$,

$$L(D_1 \cup D_2)(x) = \Phi(L(D_1)(x), L(D_2)(x)).$$

The aggregator $\Phi\colon Y \times Y \to Y$ is a binary operation on label predictions. Under universality plus label-symmetry, Φ inherits strong structural constraints. The trichotomy below is closely related to classification-aggregation impossibilities studied in the axiomatic decision-theory literature, particularly the Maniquet-Mongin characterization of compromise functions (Maniquet and Mongin, 2016) and the classification-aggregation impossibility of Cailloux et al. (2024); the present version is the specialization to a single learner's behavior on disjoint datasets, rather than to aggregating multiple classifiers' outputs.

Theorem 53 (Compositional Trichotomy)**.** *Let $|X| \geq 3$ and let L be a universal, label-symmetric, and coordinate-wise compositional learner on $\mathcal{D}_u \setminus \{\emptyset\}$ with aggregator Φ. Then Φ is* symmetric *and* $\mathrm{Sym}(Y)$-equivariant*. For $|Y| \geq 3$, Φ is also* diagonal-fixing *($\Phi(a, a) = a$ for every $a \in Y$). The cardinality $|Y|$ controls existence as follows.*

(i) *If $|Y| = 1$, the trivial learner is the unique example, and Φ is the unique constant function on the singleton.*

(ii) *If $|Y| = 2$ or $|Y| \geq 4$, no such L exists with Φ nontrivial on distinct labels (no symmetric $\mathrm{Sym}(Y)$-equivariant function on the set $Y^{(2)}$ of pairs of distinct labels exists).*

(iii) *If $|Y| = 3$, the aggregator on distinct labels is uniquely determined as $\Phi(a, b) =$ the third label, that is, the unique element of $Y \setminus \{a, b\}$, and $\Phi(a, a) = a$ for the diagonal.*

7.3 Proof of the Structural Constraints on Φ

Proof [Proof of structural constraints, $|Y| \geq 3$] We prove that Φ is symmetric and $\mathrm{Sym}(Y)$-equivariant for $|Y| \geq 3$; the case $|Y| = 2$ is handled directly in the proof of case (ii) below.

Symmetry. Since $D_1 \cup D_2 = D_2 \cup D_1$ as multisets, $L(D_1 \cup D_2) = L(D_2 \cup D_1)$. Applying coordinate-wise compositionality at any $x \notin \mathrm{dom}(D_1) \cup \mathrm{dom}(D_2)$,

$$\Phi(L(D_1)(x), L(D_2)(x)) = \Phi(L(D_2)(x), L(D_1)(x)).$$

By varying D_1 and D_2, the values $L(D_1)(x)$ and $L(D_2)(x)$ can each be made to range over Y: take $D_1 = \{(x_1, y)\}$ and $D_2 = \{(x_2, y')\}$ for any $y, y' \in Y$ and apply Corollary 25 (singleton triviality,

valid for $|Y| \geq 3$), which gives $L(D_1)(x) = y$ and $L(D_2)(x) = y'$. Hence, $\Phi(a, b) = \Phi(b, a)$ for every $a, b \in Y$.

Sym(Y)*-equivariance.* Apply $\sigma \in \mathrm{Sym}(Y)$ to both sides of the compositionality identity. By label-symmetry, $L(\sigma D_1 \cup \sigma D_2)(x) = \sigma L(D_1 \cup D_2)(x) = \sigma\Phi(L(D_1)(x), L(D_2)(x))$. Applying compositionality to the relabeled datasets,

$$L(\sigma D_1 \cup \sigma D_2)(x) = \Phi(L(\sigma D_1)(x), L(\sigma D_2)(x)) = \Phi(\sigma L(D_1)(x), \sigma L(D_2)(x)).$$

Equating the two expressions yields $\Phi(\sigma a, \sigma b) = \sigma\Phi(a, b)$ for every $\sigma \in \mathrm{Sym}(Y)$ and every $a, b \in Y$, which is Sym(Y)-equivariance. ■

7.4 Proof of the Trichotomy

Proof [Proof of the cardinality cases] *(i).* For $|Y| = 1$, it holds that Y^X has one element, the constant function. The only learner is the trivial constant predictor, and the aggregator Φ is the unique function on a singleton.

(ii), case $|Y| \geq 4$. We have already established that Φ is symmetric and Sym(Y)-equivariant. Restricted to $Y^{(2)} := \{(a, b) \in Y \times Y \mid a \neq b\}$, it is a symmetric Sym($Y$)-equivariant function $Y^{(2)} \to Y$. By Lemma 52 in the case $n \geq k + 2$, no such function exists, contradicting the existence of L.

(ii), case $|Y| = 2$. We give a direct argument since the structural constraints proof above used Corollary 25 (valid only for $|Y| \geq 3$). Write $Y = \{y_1, y_2\}$ and $\sigma := (y_1\ y_2)$, the unique nontrivial element of Sym(Y). Pick distinct $x_1, x_2, x_3 \in X$ (this is possible since $|X| \geq 3$), and set $D_1 := \{(x_1, y_1)\}$ and $D_2 := \{(x_2, y_1)\}$. Let $a := L(D_1)(x_3)$ and $b := L(D_2)(x_3)$ in Y.

Subcase $a \neq b$. By the multiset identity $D_1 \cup D_2 = D_2 \cup D_1$, compositionality at x_3 gives $\Phi(a, b) = \Phi(b, a)$. By label-symmetry, $L(\sigma(D_1 \cup D_2))(x_3) = \sigma L(D_1 \cup D_2)(x_3) = \sigma\Phi(a, b)$, while compositionality applied to σD_1, σD_2 gives $L(\sigma(D_1 \cup D_2))(x_3) = \Phi(\sigma a, \sigma b) = \Phi(b, a)$ (the last equality since $\{a, b\} = Y$, so $\sigma a = b$ and $\sigma b = a$). Combining, $\Phi(a, b) = \sigma\Phi(a, b)$; but σ has no fixed point in Y, a contradiction.

Subcase $a = b$. Replace D_2 by $D_2' := \sigma D_2 = \{(x_2, y_2)\}$. By label-symmetry, it follows that $L(D_2')(x_3) = \sigma L(D_2)(x_3) = \sigma a$. As $\sigma a \neq a$ in $Y = \{y_1, y_2\}$, the pair $(L(D_1)(x_3), L(D_2')(x_3)) = (a, \sigma a)$ is off-diagonal. The previous subcase applied to (D_1, D_2') yields the contradiction.

(iii), case $|Y| = 3$*, distinct pairs.* By the structural constraints proof, Φ on $Y^{(2)}$ is a symmetric Sym(Y)-equivariant function. By Lemma 52 in the case $n = k + 1 = 3$, the unique such function is the missing-label function $\Phi(a, b) :=$ the unique element of $Y \setminus \{a, b\}$.

(iii), diagonal-fixing at $|Y| = 3$. Consider the single-example dataset $D_a := \{(x_0, a)\}$ for $a \in Y$ and any $x_0 \in X$. By Corollary 25 (Singleton Triviality), valid for $|Y| \geq 3$, $L(D_a)(x) = a$ for every $x \in X$. Take $D_1 = \{(x_1, a)\}$ and $D_2 = \{(x_2, a)\}$ with $x_1 \neq x_2$; both yield $L(D_i) \equiv a$. Compositionality at any $x_3 \notin \{x_1, x_2\}$ gives $L(D_1 \cup D_2)(x_3) = \Phi(a, a)$. But $D_1 \cup D_2$ is a single-label dataset with $Y_{D_1 \cup D_2} = \{a\}$, and the constraint theorem implies $L(D_1 \cup D_2) \equiv a$. Thus, $\Phi(a, a) = a$. ■

7.5 The Meta-Observation and the Extremal Case

The core lemma produces the recurring $|Y| = k + 1$ cardinality across multiple axioms.

Remark 54 (The Universal Boundary $|Y| = k + 1$)**.** *Let L be a universal label-symmetric learner, and let* A *be an axiom forcing the existence of a function $f\colon Y^{(k)} \to Y$ naturally associated to L that is symmetric and* Sym(Y)*-equivariant. "Naturally associated" is informal: in each concrete instance (compositionality, ambiguity, and double-symmetry, all three at $k = 2$), the associated function is constructed explicitly, and the application of Lemma 52 is direct. Then if $|Y| = k$ or*

$|Y| \geq k+2$, *no such* L *exists; and if* $|Y| = k+1$, f *is uniquely determined as the missing-label function on* $Y^{(k)}$. *This is immediate from Lemma 52. The remark is the unifying observation behind the recurring* $|Y| = k+1$ *exceptions throughout the paper: each such exception is the same combinatorial phenomenon (the existence of a unique missing-label function at the cardinality boundary) viewed through a different axiom* A.

The compositional trichotomy is the case $k = 2$: Φ is a symmetric $\mathrm{Sym}(Y)$-equivariant function on $Y^{(2)}$, and $|Y| = 3$ is its $k+1$. The ambiguity impossibility is also the case $k = 2$, applied to the ambiguous dataset $\{(x_0, y_1), (x_0, y_2)\}$ whose setwise stabilizer in $\mathrm{Sym}(Y)$ yields the same fixpoint computation. The pattern is structural: any axiom that entails a symmetric $\mathrm{Sym}(Y)$-equivariant function on k-tuples yields a trichotomy at $|Y| = k+1$.

The trichotomy at $|Y| = 3$ has a direct interpretation in terms of the constraint theorem.

Corollary 55 (Compositionality at the Extremal Case)**.** *The nontrivial regime* $|Y| = 3$ *is the smallest cardinality at which the extremal case* $|Y_D| = |Y| - 1$ *first admits a nontrivial dataset. The missing-label aggregator at* $|Y| = 3$ *is the unique* $\mathrm{Sym}(Y)$*-equivariant selection of an unused label permitted by the silence of the constraint theorem in the extremal case.*

Proof Let $D := \{(x_1, y_1), (x_2, y_2)\}$, with $|Y_D| = 2$. For $|Y| \geq 4$ this is the generic case $|Y_D| \leq |Y| - 2$, so the constraint theorem yields $\Phi(y_1, y_2) = L(D)(x_3) \in \{y_1, y_2\}$; combined with $\mathrm{Sym}(Y)$-equivariance and the core lemma, no nontrivial Φ exists. For $|Y| = 3$ this is the extremal case, the constraint is silent, and the missing-label function $\Phi(y_1, y_2) = y_3$ is the unique $\mathrm{Sym}(Y)$-equivariant choice. Concretely, the aggregator is

$$\Phi(\text{red}, \text{green}) = \text{blue}, \quad \Phi(\text{green}, \text{blue}) = \text{red}, \quad \Phi(\text{red}, \text{blue}) = \text{green},$$

with $\Phi(a, a) = a$ on the diagonal. The associated learner is universal, label-symmetric, and compositional, but fails no-fabrication; Section 6.5 shows that adding NF eliminates this exception, restoring the impossibility uniformly for $|Y| \geq 2$. ■

8 The General Master Constraint

Sections 3–6 develop the master constraint for $H = \mathrm{Sym}(Y)$ acting only on labels, and derive the impossibility theorems from it. The present section generalizes the constraint to any group H acting *jointly* on both X and Y, which is the natural setting for input-symmetry, the double-symmetry impossibility, and the soft-label extension.

Four things are new here relative to the earlier sections. First, Lemma 56 states the master constraint in full generality; Lemma 18 is the special case $H = \mathrm{Sym}(Y)$ acting trivially on X. Second, Theorem 57 (Joint-Symmetry Feasibility) characterizes when an H-equivariant universal learner exists at all, via a fixpoint condition on the point-stabilizers of H in Y; the four impossibility theorems emerge as instances of the obstruction. Third, Theorem 59 collects and names the general G-equivariant constraint in a form that makes the fabrication set $F_G(Y_D)$ explicit, providing a single statement for the rigidity witnesses across all proper subgroups. Fourth, Theorem 61 (Soft Constraint Theorem) in Section 8.5, previewed in the introduction and proved here: label-symmetry in the soft-label setting permits only *uniform* mass on unseen labels, forbidding biased fabrication while permitting unbiased fabrication.

8.1 The General Master Constraint

The label-symmetric and G-equivariant master constraints (Lemmas 18 and 29) were the two specializations needed earlier in the paper. We now record the constraint in full generality, for an

arbitrary group H acting jointly on X and on Y. The proof is again the same fixpoint computation; only the action has changed.

Lemma 56 (Master Constraint, H-Equivariant Form)**.** *Let H be a group equipped with actions on X and on Y, and let L be H-equivariant in the sense of Definition 16. For every $D \in \mathrm{dom}(L)$ and every $x \in X$,*

$$L(D)(x) \in Y^{\mathrm{Stab}_H(D,x)},$$

where $\mathrm{Stab}_H(D, x) := \{h \in H \mid h \cdot D = D,\ h \cdot x = x\}$.

Proof For $h \in \mathrm{Stab}_H(D, x)$, the hypothesis $h \cdot D = D$ gives $L(D) = h \cdot L(D)$ by H-equivariance, where the action on predictors is $(h \cdot f)(x') = h \cdot f(h^{-1} \cdot x')$. Evaluating at x,

$$L(D)(x) = (h \cdot L(D))(x) = h \cdot L(D)(h^{-1} \cdot x) = h \cdot L(D)(x),$$

using $h \cdot x = x$. Hence, $L(D)(x)$ is fixed by every $h \in \mathrm{Stab}_H(D, x)$, so it lies in the joint fixpoint set. ■

Lemma 18 is the special case $H = \mathrm{Sym}(Y)$ acting trivially on X.

8.2 Joint-Symmetry Feasibility

The general master constraint reveals an obstruction to the existence of H-equivariant universal learners. The following theorem characterizes when no such obstruction arises. The natural statement involves the empty-dataset case, which lies outside the standard domain $\mathcal{D}_u \setminus \{\emptyset\}$ used throughout the paper; we therefore state the theorem in two parts, one for sufficiency on the standard domain and one for necessity under the additional assumption that L is defined on $\emptyset$.

Theorem 57 (Joint-Symmetry Feasibility)**.** *Let H act on X and on Y, and let*

$$\mathrm{Stab}_H(x) := \{h \in H \mid h \cdot x = x\}$$

denote the point stabilizer of x in H. The following hold.

(i) Sufficiency. *If $Y^{\mathrm{Stab}_H(x)} \neq \emptyset$ for every $x \in X$, then there exists an H-equivariant universal consistent learner on $\mathcal{D}_u \setminus \{\emptyset\}$.*

(ii) Necessity (under extended domain). *If there exists an H-equivariant universal consistent learner whose domain includes $\emptyset$, then $Y^{\mathrm{Stab}_H(x)} \neq \emptyset$ for every $x \in X$.*

Proof *(ii) Necessity.* Suppose L is H-equivariant and $\emptyset \in \mathrm{dom}(L)$. Since $h \cdot \emptyset = \emptyset$ for every $h \in H$, Lemma 56 at $D = \emptyset$ and any $x \in X$ gives $L(\emptyset)(x) \in Y^{\mathrm{Stab}_H(x)}$, which must therefore be nonempty.

(i) Sufficiency. Choose a set $\mathcal{R} \subseteq X$ of H-orbit representatives, one per H-orbit in X. For each $x_* \in \mathcal{R}$, choose some $y_* \in Y^{\mathrm{Stab}_H(x_*)}$ (possible by hypothesis). Define $L_0 \colon X \to Y$ by: for $x \in X$, write $x = h \cdot x_*$ and set $L_0(x) := h \cdot y_*$. This is well-defined: if $x = h \cdot x_* = h' \cdot x_*$, then $h^{-1}h' \in \mathrm{Stab}_H(x_*)$, so $h \cdot y_* = h' \cdot y_*$. The function L_0 is H-equivariant: $L_0(g \cdot x) = gh \cdot y_* = g \cdot L_0(x)$. Define L on $\mathcal{D}_u \setminus \{\emptyset\}$ by $L(D)(x) := D(x)$ if $x \in \mathrm{dom}(D)$ and $L(D)(x) := L_0(x)$ otherwise. This L is universal, consistent, and H-equivariant; the construction is well-posed because H acts by bijections, preserving unambiguity. To verify equivariance: if $x \in \mathrm{dom}(h \cdot D)$, then $L(h \cdot D)(x) = h \cdot D(h^{-1} \cdot x) = h \cdot L(D)(h^{-1} \cdot x)$ by consistency; if $x \notin \mathrm{dom}(h \cdot D)$, then $L(h \cdot D)(x) = L_0(x) = h \cdot L_0(h^{-1} \cdot x) = h \cdot L(D)(h^{-1} \cdot x)$ by H-equivariance of L_0. Both cases give $L(h \cdot D)(x) = (h \cdot L(D))(x)$. ■

Remark 58 (On the Role of the Empty Dataset)**.** *The necessity direction genuinely requires the extended domain. Without $\emptyset$, the closest analogous statement on the standard domain is vacuous: indeed, for any singleton $D = \{(x_0, y_0)\}$, the stabilizer $\mathrm{Stab}_H(D, x_0)$ always fixes y_0, so $Y^{\mathrm{Stab}_H(D,x_0)} \neq \emptyset$ holds trivially, giving no obstruction. The empty-dataset hypothesis in (ii) is therefore essential; the sufficiency direction (i) is the substantive statement on the standard domain, and it is what we use in the corollaries that follow.*

8.3 Recovering the Impossibility Theorems via the General Master Constraint

Each of the four impossibility theorems is an instance of the general master constraint at a specific (D, x, H); Locality and Double-Symmetry are recovered here via the joint-symmetry feasibility framework, Ambiguity and Trichotomy are recovered in their natural sections via the same master-constraint principle. We work the two cases for which the joint-symmetry framework is the natural setting, and record the others by reference.

Locality. Take $H = \mathrm{Sym}(Y)$, acting trivially on X. Strict locality (Definition 10) at $x \notin \mathrm{dom}(D)$ requires $L(\emptyset)(x)$ to exist, so $\emptyset \in \mathrm{dom}(L)$. We have $\mathrm{Stab}_H(x) = \mathrm{Sym}(Y)$ for every $x \in X$, and $Y^{\mathrm{Sym}(Y)} = \emptyset$ for $|Y| \geq 2$. The extended-domain necessity in Theorem 57(ii) is violated, so no such L exists.

Double-symmetry. Take $H = \mathrm{Sym}(X) \times \mathrm{Sym}(Y)$, acting jointly on X and Y. The impossibility does not follow directly from Theorem 57, since the obstruction arises not at the empty dataset but at the two-label dataset $D = \{(x_1, y_1), (x_2, y_2)\}$ with a third unseen input x_3. It is, however, a direct instance of the general master constraint (Lemma 56): the joint (D, x_3)-stabilizer in H forces $L(D)(x_3)$ to lie in the intersection of $\{y_1, y_2\}$ (imposed by the pointwise stabilizer of Y_D via label-symmetry) and $Y \setminus \{y_1, y_2\}$ (imposed by the diagonal swap (τ, σ) swapping $x_1 \leftrightarrow x_2$ and $y_1 \leftrightarrow y_2$ simultaneously). That intersection is empty for $|Y| = 2$ or $|Y| \geq 4$. The detailed argument, including the $|Y| = 3$ exception, is Theorem 46.

Ambiguity. Theorem 47 (ambiguity impossibility) instantiates the master constraint at $H = \mathrm{Sym}(Y)$ and the dataset $S = \{(x_0, y_1), (x_0, y_2)\}$, whose setwise stabilizer $\mathrm{Sym}(\{y_1, y_2\}) \times \mathrm{Sym}(Y \setminus \{y_1, y_2\})$ has empty joint fixed set in Y for $|Y| = 2$ or $|Y| \geq 4$. We refer to Section 6 for the detailed argument.

Compositional trichotomy. The compositional trichotomy (Theorem 53) is the fourth instance, recovered via the core lemma (Lemma 52) on symmetric $\mathrm{Sym}(Y)$-equivariant functions on $Y^{(2)}$: the aggregator Φ inherits symmetry and equivariance from coordinate-wise compositionality and label-symmetry, and the core lemma rules out nontrivial Φ at $|Y| = 2$ or $|Y| \geq 4$. The detailed argument is in Section 7.

Diagonal action ($X = Y$). Joint symmetry is not always obstructive; the diagonal action gives a natural feasible example. Take $X = Y = \{1, 2, \ldots, n\}$ and $H = \mathrm{Sym}(\{1, 2, \ldots, n\})$ acting diagonally. Then $\mathrm{Stab}_H(i) = \mathrm{Sym}(\{1, 2, \ldots, n\} \setminus \{i\})$, with $Y^{\mathrm{Stab}_H(i)} = \{i\}$. The condition is satisfied: by Theorem 57(i) an H-equivariant universal consistent learner exists on $\mathcal{D}_u \setminus \{\emptyset\}$, with the natural extension to $\emptyset$ predicting the identity function $X \to Y$.

8.4 The General Constraint Theorem

For a general subgroup $G \leq \mathrm{Sym}(Y)$ (with $H = G$ acting trivially on X), the analogue of the constraint theorem takes the following form.

Theorem 59 (General G-Equivariant Constraint)**.** *Let $G \leq \mathrm{Sym}(Y)$, let L be a universal G-equivariant learner on $\mathcal{D}_u \setminus \{\emptyset\}$, and let D be a nonempty unambiguous dataset. For every $x \in X$,*

$$L(D)(x) \in Y_D \cup F_G(Y_D),$$

where $F_G(Y_D) := \{y \in Y \setminus Y_D \mid \sigma(y) = y \text{ for every } \sigma \in G \cap \mathrm{Sym}(Y \setminus Y_D)\}$ *is the set of* fabricable labels permitted by G at Y_D.

Proof By Lemma 29, $L(D)(x) \in Y^{\mathrm{Stab}_G(D)}$. By Lemma 30, $\mathrm{Stab}_G(D) = G \cap \mathrm{Sym}(Y \setminus Y_D)$. The fixpoint set in Y of this subgroup is Y_D (trivially fixed) together with $F_G(Y_D)$. ■

The fabrication set $F_G(Y_D)$ measures how much fabrication a G-equivariant learner is permitted on a dataset with label support Y_D. For $G = \mathrm{Sym}(Y)$, $F_G(Y_D)$ is empty when $|Y_D| \leq |Y| - 2$ (recovering the constraint theorem) and equals $Y \setminus Y_D$ when $|Y_D| = |Y| - 1$ (recovering the extremal case). For proper subgroups, $F_G(Y_D)$ can be nonempty on the generic case, recovering the rigidity theorem's witness construction. This theorem is strictly more general than Theorem 21: the latter is the $G = \mathrm{Sym}(Y)$ specialization restricted to the generic case, while the present statement holds for arbitrary $G \leq \mathrm{Sym}(Y)$ and all $D \in \mathcal{D}_u \setminus \{\emptyset\}$, with the generic/extremal structure absorbed into $F_G(Y_D)$.

8.5 The Soft-Label Constraint Theorem

The constraint theorem is stated for hard-label learners $L(D)(x) \in Y$. Many practical classifiers, including all classifiers trained with cross-entropy loss, return soft labels, that is, probability distributions over Y. The master constraint principle extends without change to the soft-label setting, with a genuine refinement of the conclusion: a universal label-symmetric soft-label learner may leak mass to unseen labels, but only *uniformly* across those unseen labels. Fabrication of a specific unseen label is forbidden; fabrication of unseen labels as a class, in the uniform-mass form, is permitted.

Let $\Delta(Y)$ denote the simplex of probability distributions on Y. By analogy with the hard-label case, a *soft-label learner* is a partial function $L \colon \mathcal{D}(X, Y) \rightharpoonup \Delta(Y)^X$. The symmetric group $\mathrm{Sym}(Y)$ acts on $\Delta(Y)$ by permuting coordinates, $(\sigma \cdot p)(y) := p(\sigma^{-1}(y))$, and on $\Delta(Y)^X$ pointwise; a soft-label learner is *label-symmetric* if $L(\sigma D) = \sigma \cdot L(D)$. The master constraint of Lemma 18 applies to this action, since $\Delta(Y)$ is a $\mathrm{Sym}(Y)$-set in exactly the same way Y is.

Lemma 60 (Invariant Distributions)**.** *Let* $D \in \mathcal{D}_u \setminus \{\emptyset\}$ *be in the generic case, with*

$$\mathrm{Stab}_{\mathrm{Sym}(Y)}(D) \cong \mathrm{Sym}(Y \setminus Y_D)$$

acting as the pointwise stabilizer of Y_D*. A distribution* $p \in \Delta(Y)$ *is fixed by* $\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)$ *if and only if* p *is constant on the unseen labels* $Y \setminus Y_D$*. Equivalently,* p *admits the unique decomposition*

$$p = (1 - \alpha) \cdot q + \alpha \cdot u_{Y \setminus Y_D},$$

where $q \in \Delta(Y_D)$*,* $u_{Y \setminus Y_D}$ *denotes the uniform distribution on the unseen labels, and* $\alpha \in [0, 1]$ *is the total mass assigned to those unseen labels.*

Proof The orbits of $\mathrm{Sym}(Y \setminus Y_D)$ on Y are the singletons $\{y\}$ for $y \in Y_D$ (each fixed pointwise) and the single orbit $Y \setminus Y_D$ (permuted transitively when $|Y \setminus Y_D| \geq 2$). A distribution p is invariant if and only if it is constant on each orbit. On the singleton orbits this is automatic; on $Y \setminus Y_D$ this means there is a common value $c \geq 0$ with $p(y) = c$ for every $y \in Y \setminus Y_D$. Let $\alpha := \sum_{y \in Y \setminus Y_D} p(y) = (|Y| - |Y_D|) \cdot c$ be the total mass p assigns to the unseen labels, so $c = \alpha / |Y \setminus Y_D|$ when $|Y \setminus Y_D| \geq 1$. The restriction of p to $Y \setminus Y_D$ is then $\alpha \cdot u_{Y \setminus Y_D}$, and the restriction to Y_D has total mass $1 - \alpha$ and is an arbitrary $q \in \Delta(Y_D)$. The decomposition is unique because α and q are recovered from p as the total unseen mass and the conditional distribution on Y_D, respectively. ■

Theorem 61 (Soft Constraint Theorem)**.** *Let L be a universal label-symmetric soft-label learner on $\mathcal{D}_u \setminus \{\emptyset\}$. For every D in the generic case $|Y_D| \leq |Y| - 2$ and every $x \in X$, the distribution $L(D)(x)$ admits the decomposition*

$$L(D)(x) = (1 - \alpha(D, x)) \cdot q(D, x) + \alpha(D, x) \cdot u_{Y \setminus Y_D},$$

where $q(D, x) \in \Delta(Y_D)$ and $\alpha(D, x) \in [0, 1]$. In particular, the mass assigned to $Y \setminus Y_D$ is uniform across the unseen labels: for any two unseen labels $y, y' \in Y \setminus Y_D$,

$$L(D)(x)(y) = L(D)(x)(y') = \frac{\alpha(D, x)}{|Y \setminus Y_D|}.$$

No individual unseen label is favored over any other.

Proof By the master constraint (Lemma 18, applied to $\Delta(Y)$ as a $\mathrm{Sym}(Y)$-set; the proof of Lemma 18 goes through verbatim with Y replaced by $\Delta(Y)$, as noted at the start of this section), $L(D)(x) \in \Delta(Y)^{\mathrm{Stab}_{\mathrm{Sym}(Y)}(D)}$, and Lemma 60 gives the stated decomposition. ■

The parameter $\alpha(D, x)$ measures the total mass leaked onto unseen labels at the point (D, x), and label-symmetry alone places no upper bound on it: any $\alpha \in [0, 1]$ is admissible at each (D, x), including the extreme $\alpha = 1$ that places all mass uniformly on the unseen labels. In particular, a softmax classifier trained with cross-entropy on a generic dataset may assign nonzero probability to unseen classes, but only by spreading that mass uniformly across them: it cannot systematically favor one unseen class over another. This is a falsifiable empirical prediction about the average behavior of label-symmetric cross-entropy training procedures, not about any single trained checkpoint. To test it, hold out two or more classes from the training set of a softmax cross-entropy classifier, retrain the model repeatedly with different random initializations of a label-symmetric procedure, and average the output distributions on inputs whose true labels are in the held-out set. The theorem predicts that the averaged probability mass assigned to the held-out classes will be uniform across them; any systematic preference for one held-out class over another, after averaging over training-procedure randomness, would indicate that the architecture or update rule is not, in fact, label-symmetric at the level of the learning procedure. The hard constraint theorem is the deterministic specialization.

Corollary 62 (Determinism Forces the Hard Constraint)**.** *If L is universal, label-symmetric, and deterministic, that is, $L(D)(x)$ is a Dirac distribution e_y for some $y \in Y$ at every D and x, then* $\mathrm{supp}(L(D)(x)) \subseteq Y_D$ *on the generic case.*

Proof Theorem 61 gives $L(D)(x) = (1-\alpha)q + \alpha \cdot u_{Y \setminus Y_D}$ for some $q \in \Delta(Y_D)$ and $\alpha \in [0, 1]$. We split on the value of α. If $\alpha = 0$, then $L(D)(x) = q$ is a Dirac supported on Y_D, and the conclusion holds. If $\alpha = 1$, the combination collapses to $u_{Y \setminus Y_D}$, which is a Dirac only when $|Y \setminus Y_D| = 1$, ruled out on the generic case where $|Y \setminus Y_D| \geq 2$. If $0 < \alpha < 1$, the convex combination $(1 - \alpha)q + \alpha u_{Y \setminus Y_D}$ is a Dirac only if both summands are the same Dirac (any other combination has support of size at least two); but $u_{Y \setminus Y_D}$ is uniform on $Y \setminus Y_D$ with $|Y \setminus Y_D| \geq 2$, hence not a Dirac, so this case is impossible. Hence, $\alpha = 0$ and the support is contained in Y_D. ■

Corollary 63 (Soft No-Fabrication Recovers the Hard Support Constraint)**.** *Adding the* soft no-fabrication *axiom $\alpha(D, x) = 0$ for every generic D and every $x \in X$ to label-symmetry reduces Theorem 61 to the hard support constraint* $\mathrm{supp}(L(D)(x)) \subseteq Y_D$*.*

Proof The soft no-fabrication axiom implies $\alpha(D, x) = 0$ pointwise. An application of Theorem 61 then gives $L(D)(x) = q(D, x) \in \Delta(Y_D)$. ■

Soft no-fabrication is therefore a content-bearing axiom in the soft setting, even though its hard analogue is redundant under label-symmetry on the generic case (Corollary 23). Indeed, the soft setting is permissive enough that the no-fabrication constraint must be imposed separately to recover the hard support condition.

9 The Axiom-Consistency Lattice

The set $\mathcal{A} = \{\mathsf{U}, \mathsf{C}, \mathsf{LS}, \mathsf{IS}, \mathsf{Loc}, \mathsf{Idem}, \mathsf{Mon}, \mathsf{NF}, \mathsf{CL}, \mathsf{Comp}\}$ is a catalogue of ten properties a learner may or may not possess. We do not propose $\mathcal{A}$ as a unified theory of classification: as the analysis below makes precise, the full set is inconsistent. Rather, the catalogue is a menu, and the substantive axiomatic systems are the consistent subsets $S \subseteq \mathcal{A}$.

Each consistent subset S defines a coherent design space, namely the class of learners satisfying every axiom in S simultaneously. Each inconsistent subset records a combination of properties no learner can possess at once. The structural questions of the catalogue are therefore the following. *Which subsets are consistent? Which inconsistencies are minimal, in the sense that dropping any single axiom yields a consistent subset? Which consistent subsets are maximal, in the sense that adding any further axiom yields an inconsistent subset?* This section answers all three.

The answer has a tight shape. Three minimal impossibilities (subsets that are inconsistent, but whose proper subsets are all consistent) account for every inconsistency in $\mathcal{A}$. Three maximal consistent subsets (consistent subsets that no further axiom can be added to) classify the consistent combinations: every consistent subset of $\mathcal{A}$ is contained in one of these three. The list of consistent axiomatic systems built from $\mathcal{A}$ is therefore exactly the set of subsets of three explicit lists.

9.1 Definitions and Working Regime

We fix the regime in which the lattice analysis takes place.

Definition 64 (Working Regime B_2)**.** *The* working regime *of this section is*

$$B_2 \;:=\; \{(|X|, |Y|) \mid |X| \geq 3,\ |Y| \geq 2,\ |Y| \neq 3\}.$$

This is the regime at which all three minimal impossibilities of Theorem 67 bite simultaneously. The cardinality $|X| \geq 3$ ensures that the compositionality and input-symmetry impossibilities are non-vacuous: each of these uses an input outside the union of two disjoint singleton domains, which exists only when $|X| \geq 3$. The boundary case $|Y| = 3$, where the compositional trichotomy admits the missing-label exception, is treated separately in the explicit enumeration of Section 9.5 at $|X| = 2$, $|Y| = 3$, which is outside B_2.

Definition 65 (Consistency of a Subset)**.** *A subset $S \subseteq \mathcal{A}$ is* consistent *if there exists a learner L satisfying every axiom in S simultaneously, and* inconsistent *otherwise.*

Definition 66 (Minimal Impossibility, Maximal Consistent Subset)**.** *A subset $S \subseteq \mathcal{A}$ is a* minimal impossibility *if S is inconsistent and every proper subset of S is consistent. A subset $S \subseteq \mathcal{A}$ is a* maximal consistent subset *if S is consistent and every strict superset of S in $\mathcal{A}$ is inconsistent.*

Consistency is downward-closed: if S is consistent and $S' \subseteq S$, then any witness for S also witnesses S'. Inconsistency is upward-closed: if S is inconsistent and $S \subseteq S'$, then S' has no witness either. The consistent subsets therefore form a downward-closed family in the powerset $2^{\mathcal{A}}$, fully determined by their maximal elements.

9.2 Minimal Impossibilities

Three minimal impossibilities organize the lattice.

Theorem 67 (Minimal Impossibilities)**.** *On the regime B_2, the minimal impossibilities in $\mathcal{A}$ are exactly:*

(1) $\{\mathsf{U}, \mathsf{Loc}\}$*: locality impossibility (Theorem 45).*

(2) $\{\mathsf{U}, \mathsf{LS}, \mathsf{IS}\}$*: double-symmetry impossibility (Theorem 46).*

(3) $\{\mathsf{U}, \mathsf{LS}, \mathsf{Comp}\}$*: the compositional trichotomy impossibility on B_2 (Theorem 53).*

The boundary case $|Y| = 3$, where item (3) admits the missing-label compositional exception and so is not an impossibility, is treated separately in Section 9.5.

Proof Items (1), (2), (3) are restatements of the corresponding theorems. For minimality, each proper subset of every listed impossibility is witnessed by one of the learners $L_<$, $L_\prec$, L_{tot} of Theorem 68 below. Specifically, the proper subsets of $\{\mathsf{U}, \mathsf{Loc}\}$ are consistent because $\{\mathsf{U}\}$ is witnessed by $L_<$ (which satisfies U and fails Loc) and $\{\mathsf{Loc}\}$ by L_{tot} (which is vacuously strictly local on total data, and fails U). The proper subsets of $\{\mathsf{U}, \mathsf{LS}, \mathsf{IS}\}$ are consistent because $\{\mathsf{U}, \mathsf{LS}\}$ is witnessed by $L_<$ (drops IS), $\{\mathsf{U}, \mathsf{IS}\}$ by $L_\prec$ (drops LS), and $\{\mathsf{LS}, \mathsf{IS}\}$ by L_{tot}, which is the identity on total datasets and is therefore both label-symmetric and input-symmetric, vacuously on its domain (drops U). The proper subsets of $\{\mathsf{U}, \mathsf{LS}, \mathsf{Comp}\}$ are consistent because $\{\mathsf{U}, \mathsf{LS}\}$ is again witnessed by $L_<$ (drops Comp), $\{\mathsf{U}, \mathsf{Comp}\}$ by $L_\prec$ (drops LS, satisfies Comp with aggregator $\min_\prec$), and $\{\mathsf{LS}, \mathsf{Comp}\}$ by L_{tot} (drops U). Full verification of which axioms each witness satisfies is in Theorem 68. ■

9.3 Maximal Consistent Subsets

We turn from the impossibilities at the top of the lattice to its consistent core: the maximal subsets of $\mathcal{A}$ on which all axioms can be satisfied simultaneously.

Theorem 68 (Maximal Consistent Subsets)**.** *On the regime B_2, the maximal consistent subsets of the catalogue $\mathcal{A}$ are exactly three:*

(1) $M_1 = \{\mathsf{U}, \mathsf{C}, \mathsf{LS}, \mathsf{Idem}, \mathsf{Mon}, \mathsf{NF}, \mathsf{CL}\}$ *(excludes $\mathsf{IS}, \mathsf{Loc}, \mathsf{Comp}$): the universal label-symmetric regime, witnessed by the nearest-by-order learner $L_<$.*

(2) $M_2 = \{\mathsf{U}, \mathsf{C}, \mathsf{IS}, \mathsf{Idem}, \mathsf{Mon}, \mathsf{NF}, \mathsf{CL}, \mathsf{Comp}\}$ *(excludes $\mathsf{LS}, \mathsf{Loc}$): the universal input-symmetric compositional regime, witnessed by the minimum-of-data learner $L_\prec$.*

(3) $M_3 = \{\mathsf{C}, \mathsf{LS}, \mathsf{IS}, \mathsf{Loc}, \mathsf{Idem}, \mathsf{Mon}, \mathsf{NF}, \mathsf{CL}, \mathsf{Comp}\}$ *(excludes U): the total-data identity regime, witnessed by L_{tot} on total datasets.*

The boundary case $|Y| = 3$ is treated separately in Section 9.5, where the nine universal label-symmetric learners are enumerated and the maximal consistent strata redescribed.

Proof By Theorem 67, the minimal impossibilities in $\mathcal{A}$ on B_2 are $\{\mathsf{U}, \mathsf{Loc}\}$, $\{\mathsf{U}, \mathsf{LS}, \mathsf{IS}\}$, and $\{\mathsf{U}, \mathsf{LS}, \mathsf{Comp}\}$. A subset $S \subseteq \mathcal{A}$ is consistent if and only if it contains no minimal impossibility, equivalently, if and only if the complement $\mathcal{A} \setminus S$ is a hitting set for the three minimal impossibilities. The maximal consistent subsets are therefore the complements of the minimal hitting sets.

The minimal hitting sets for $\{\mathsf{U}, \mathsf{Loc}\}$, $\{\mathsf{U}, \mathsf{LS}, \mathsf{IS}\}$, $\{\mathsf{U}, \mathsf{LS}, \mathsf{Comp}\}$ are:

- $\{\mathsf{U}\}$: hits all three (since U appears in each).
- $\{\mathsf{LS}, \mathsf{Loc}\}$: hits all three.
- $\{\mathsf{IS}, \mathsf{Comp}, \mathsf{Loc}\}$: hits all three.

Complementing in $\mathcal{A}$ gives the subsets M_3, M_2, M_1 in the theorem statement. Since NF does not appear in any minimal impossibility, it lies in every maximal consistent subset.

We now verify that each M_i is consistent via the stated witness learner.

(1) M_1 witnessed by $L_<$. Fix a total order on X and set $L_<(D)(x)$ to be the label of the nearest input to x in $\mathrm{dom}(D)$ under this order. The learner is universal, consistent, label-symmetric, idempotent, monotone (if $L_<(D)(x) = y'$, the nearest input to x in $\mathrm{dom}(D)$ carries y'; adding (x', y') to D either leaves this nearest input unchanged, or replaces it with x', which itself carries y', so the prediction at x remains y'), nonfabricating, and self-training closed. It fails input-symmetry (the order on X breaks it), strict locality (it uses data at other inputs), and compositionality. The self-training closure verification is given below.

(2) M_2 witnessed by $L_\prec$. Fix a total order $\prec$ on Y, and set $L_\prec(D)(x) := D(x)$ for $x \in \mathrm{dom}(D)$ and $L_\prec(D)(x) := \min_\prec Y_D$ otherwise. The learner is universal, consistent, input-symmetric (for any $\tau \in \mathrm{Sym}(X)$, the relabeling $\tau \cdot D$ has $Y_{\tau \cdot D} = Y_D$ since τ permutes inputs without changing which labels appear, hence $\min_\prec Y_{\tau \cdot D} = \min_\prec Y_D$ and the off-domain default is invariant), idempotent (adding a duplicate (x, y) to D changes neither $\mathrm{dom}(D)$ nor Y_D, and since $L_\prec(D)(x) = D(x)$ on $\mathrm{dom}(D)$ and $L_\prec(D)(x) = \min_\prec Y_D$ off-domain, the predictor at every input is unchanged, hence $L_\prec(D \uplus \{(x, y)\}) = L_\prec(D)$), monotone, nonfabricating (since $\min_\prec Y_D \in Y_D$, both on $\mathrm{dom}(D)$ inputs by consistency and on off-domain inputs by definition), self-training closed, and coordinate-wise compositional with aggregator $\Phi(a, b) := \min_\prec(a, b)$, the meet under $\prec$. It fails label-symmetry, since $\min_\prec$ is not $\mathrm{Sym}(Y)$-equivariant, and fails strict locality, since Loc requires $\emptyset \in \mathrm{dom}(L_\prec)$ but universality places the domain of $L_\prec$ at $\mathcal{D}_u \setminus \{\emptyset\}$, so $\emptyset \notin \mathrm{dom}(L_\prec)$. The compositionality verification is given below.

(3) M_3 witnessed by L_{tot}. Defined only on total datasets $D \colon X \to Y$, with $L_{\mathrm{tot}}(D)(x) := D(x)$. The learner is consistent, label-symmetric, input-symmetric, strictly local (vacuously: $\mathrm{dom}(D) = X$ for all $D \in \mathrm{dom}(L_{\mathrm{tot}})$), idempotent, monotone, nonfabricating, self-training closed, and coordinate-wise compositional with the trivial identity-on-singletons aggregator. It fails universality (defined only on total data). ■

Corollary 69 (Classification of Consistent Subsets)**.** *A subset $S \subseteq \mathcal{A}$ is consistent on the regime B_2 if and only if $S \subseteq M_i$ for some $i \in \{1, 2, 3\}$.*

Proof ($\Leftarrow$) If $S \subseteq M_i$, the witness for M_i also witnesses S, since consistency is downward-closed.

($\Rightarrow$) Suppose S is consistent. We extend S to a maximal consistent subset M by adding axioms one at a time, preserving consistency at each step. Note that this process necessarily terminates since $\mathcal{A}$ is finite. By Theorem 68, $M \in \{M_1, M_2, M_3\}$, so $S \subseteq M_i$ for some i. ■

The consistent axiomatic systems that can be built from the catalogue $\mathcal{A}$ are therefore exactly the subsets of M_1, M_2, or M_3. The catalogue admits three coherent design spaces and no others.
For each of the three witness learners, the proof above lists the axioms it satisfies. We supply the two nonobvious verifications explicitly.

Self-training closure for $L_<$. For the fixed total order on X and any $T \subseteq X$, $x \in X$, define $\nu_T(x) := \min\{t \in T \mid t \geq x\}$, with the convention $\nu_T(x) := \max T$ when no element of T exceeds x (treating the order on X cyclically, so that "next element at or above x" wraps around to the maximum of T when no such element exists). The learner is then $L_<(D)(x) := D(\nu_{\mathrm{dom}(D)}(x))$. For $T \subseteq X' \subseteq X$ and any $x \in X$, the identity

$$\nu_T(\nu_{X'}(x)) = \nu_T(x) \tag{$*$}$$

holds: $\nu_T(x) \in T \subseteq X'$ and $\nu_T(x) \geq x$, so $\nu_{X'}(x) \in [x, \nu_T(x)]$ (the smallest element of X' at or above x lies at or below $\nu_T(x)$); and no element of T sits in $[x, \nu_T(x))$, so $\nu_T(\nu_{X'}(x)) = \nu_T(x)$.

In the wraparound case (no element of X' at or above x), the convention $\nu_{X'}(x) := \max X'$ kicks in. Since $T \subseteq X'$ and no element of T is at or above x (because $T \subseteq X'$ inherits the absence), the outer $\nu_T(x)$ also wraps around: $\nu_T(x) = \max T$. The inner $\nu_T(\nu_{X'}(x)) = \nu_T(\max X')$ also wraps, since no element of T exceeds $\max X'$ either: $\nu_T(\max X') = \max T$. Both sides of $(*)$ therefore equal $\max T$, and the identity holds.

Given $(*)$, the closure property is immediate. The closure $D_S^{L_<}$ assigns to each $z \in S$ the label $D(\nu_{\mathrm{dom}(D)}(z))$, so for any $x \in X$,

$$L_<(D_S^{L_<})(x) = D_S^{L_<}(\nu_S(x)) = D\big(\nu_{\mathrm{dom}(D)}(\nu_S(x))\big) = D\big(\nu_{\mathrm{dom}(D)}(x)\big) = L_<(D)(x),$$

where the third equality applies $(*)$ with $X' = S, T = \mathrm{dom}(D) \subseteq S$.

Coordinate-wise compositionality for $L_\prec$. For D_1, D_2 with disjoint domains satisfying in addition $x \notin \mathrm{dom}(D_1) \cup \mathrm{dom}(D_2)$, the definition gives $L_\prec(D_i)(x) = \min_\prec Y_{D_i}$ for $i = 1, 2$, as well as $L_\prec(D_1 \cup D_2)(x) = \min_\prec Y_{D_1 \cup D_2}$. Since D_1, D_2 have disjoint domains, $Y_{D_1 \cup D_2} = Y_{D_1} \cup Y_{D_2}$, and the minimum under $\prec$ of a union is the minimum of the two minima. It therefore follows that $L_\prec(D_1 \cup D_2)(x) = \min_\prec(\min_\prec Y_{D_1}, \min_\prec Y_{D_2}) = \Phi(L_\prec(D_1)(x), L_\prec(D_2)(x))$, where we used the notation $\Phi(a, b) := \min_\prec(a, b)$, the meet under $\prec$.

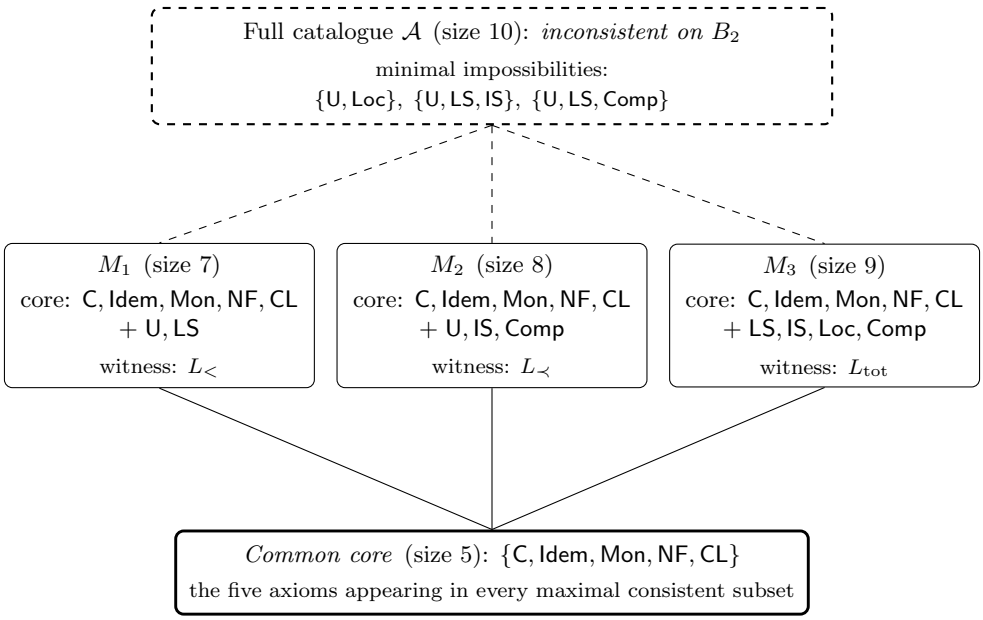


Figure 1: Consistency structure of the full catalogue $\mathcal{A} = \{\mathsf{U}, \mathsf{C}, \mathsf{LS}, \mathsf{IS}, \mathsf{Loc}, \mathsf{Idem}, \mathsf{Mon}, \mathsf{NF}, \mathsf{CL}, \mathsf{Comp}\}$. The full catalogue (shown dashed) is inconsistent on the regime B_2, with three minimal impossibilities $\{\mathsf{U}, \mathsf{Loc}\}, \{\mathsf{U}, \mathsf{LS}, \mathsf{IS}\}, \{\mathsf{U}, \mathsf{LS}, \mathsf{Comp}\}$. The three maximal consistent subsets M_1, M_2, M_3 (Theorem 68) are pairwise incomparable, with shared core $M_1 \cap M_2 \cap M_3$ shown at the bottom.

9.4 Implications and the Open Independence Question

The lattice of Section 9 settles which subsets of $\mathcal{A}$ are consistent. We close the section by recording two further structural facts about the catalogue: which axioms imply which others, and what remains unresolved about independence within each maximal consistent subset.

Proposition 70 (Relations Among the Axioms)**.** *Let* $\mathcal{A} := \{\mathsf{U}, \mathsf{C}, \mathsf{LS}, \mathsf{IS}, \mathsf{Loc}, \mathsf{Idem}, \mathsf{Mon}, \mathsf{NF}, \mathsf{CL}, \mathsf{Comp}\}$ *denote the catalogue.*

(i) Implications. *The only nontrivial* unconditional *implication between distinct axioms in* $\mathcal{A}$ *within the framework is the constraint theorem:*

$$\{\mathsf{U}, \mathsf{LS}\} \implies \mathsf{NF} \text{ on every generic dataset} \quad (\textit{Theorem 21}).$$

The implication fails on extremal datasets and on the empty dataset, and on every regime under G*-equivariance with* $G \subsetneq \mathrm{Sym}(Y)$ *(Proposition 50). Corollary 25 (Singleton Triviality) is the singleton specialization of this implication: at* $|Y_D| = 1$ *the constraint* $L(D)(x) \in Y_D = \{y_0\}$ *pins the prediction completely, but this is the singleton case of* $\{\mathsf{U}, \mathsf{LS}\} \Rightarrow \mathsf{NF}$ *and not a separate implication between catalogue axioms. A restricted local implication holds within the closure interval: under* C*, the axiom* CL *implies* Mon *on* $[D, c_L(D)]$ *(Proposition 74); globally,* CL *does not imply* Mon*.*

(ii) Inconsistency of $\bigwedge \mathcal{A}$. *The full catalogue is inconsistent on the regime* B_2*, with three minimal impossibilities*

$$\{\mathsf{U}, \mathsf{Loc}\}, \qquad \{\mathsf{U}, \mathsf{LS}, \mathsf{IS}\}, \qquad \{\mathsf{U}, \mathsf{LS}, \mathsf{Comp}\}$$

(Theorem 67). The classical formulation "A *is independent of* $\mathcal{A} \setminus \{\mathsf{A}\}$*" is therefore meaningful only for* $\mathsf{A} = \mathsf{U}$*, witnessed by* L_{tot}*; for every other axiom* A*, the set* $\mathcal{A} \setminus \{\mathsf{A}\}$ *already contains a minimal impossibility and is itself inconsistent. The meaningful independence questions are accordingly within each maximal consistent subset* M_i *(Theorem 68), discussed in Remark 71 below.*

Remark 71 (Within-M_i Irredundancy: A Natural Follow-Up)**.** *Theorem 68 establishes that each maximal consistent subset* M_i *is nonempty, witnessed respectively by* $L_{\prec}$*,* $L_{\succ}$*, and* L_{tot}*. A finer structural question is whether each* M_i *is* irredundantly *axiomatized: for each axiom* $\mathsf{A} \in M_i$*, does there exist a learner satisfying* $M_i \setminus \{\mathsf{A}\}$ *but failing* A*? A positive answer means no axiom in* M_i *is forced by the remaining axioms; a negative answer would identify a strictly smaller minimal core.*

The question is nontrivial and the answer depends on $|X|$ *and* $|Y|$ *within* B_2*. For an explicit instance, at* $|X| = 3$ *define a learner* L *on total datasets only, sending datasets with* $|Y_D| \in \{1, 3\}$ *to themselves and sending datasets with* $|Y_D| = 2$ *(necessarily of the form: two inputs carrying one label, the remaining input carrying another) to the constant predictor returning the minority label at every input. This learner satisfies every axiom of* M_3 *except* C*: the minority label is* $\mathrm{Sym}(X)$*-and* $\mathrm{Sym}(Y)$*-equivariant; the resulting predictor on* $|Y_D| = 2$ *inputs is constant and label-symmetric; and the remaining axioms of* M_3 *hold either by construction on totals or vacuously since* L *has no non-total datasets in its domain. Hence,* C *is irredundant in* M_3 *at these cardinalities.*

The complete case analysis across B_2*, sensitive to* $|X|$ *and* $|Y|$*, is a natural follow-up; the explicit instance above verifies irredundancy of* C *in* M_3 *at* $|X| = 3$*.*

9.5 Explicit Enumeration at $|X| = 2$, $|Y| = 3$

The smallest nontrivial case is $|X| = 2, |Y| = 3$, where by Example 3 the characterization theorem gives exactly nine universal label-symmetric learners. We describe the parameter space in detail.

The $\mathrm{Sym}(Y)$-orbits of nonempty unambiguous datasets are:

- Two singleton orbits, one per input, $[\{(x_i, y)\}]$ for $i = 1, 2$. Each has $|Y_D| = 1$, generic case, forced prediction broadcasting the single label.
- One single-label two-input orbit $[\{(x_1, y), (x_2, y)\}]$, also generic case, with the prediction forced to equal y at every input.
- One two-label orbit $[\{(x_1, y_1), (x_2, y_2)\}]$. Here $|Y_D| = 2 = |Y| - 1$, extremal case, and the predictor at the representative can be any function $\{x_1, x_2\} \to Y$, giving $|Y|^{|X|} = 9$ choices.

The nine universal label-symmetric learners are parametrized by the nine predictors at the two-label orbit representative. Three are worth singling out (see Table 2 below for the full 3×3 grid, where rows index $L(D)(x_1) \in Y$ and columns index $L(D)(x_2) \in Y$).

Data-completion. Predictor $(x_1 \mapsto y_1, x_2 \mapsto y_2)$, agreeing with the data. Satisfies C, NF, IS, CL. This is the unique learner satisfying all four among the nine: consistency implies $L(D)(x_1) = y_1$ and $L(D)(x_2) = y_2$ at the two-label representative $D = \{(x_1, y_1), (x_2, y_2)\}$, leaving no free choice.

Missing-label compositional. Predictor $(x_1 \mapsto y_3, x_2 \mapsto y_3)$, the constant missing label. Satisfies IS and CL (and exhibits the missing-label aggregator structure that becomes non-vacuous compositionality at $|X| \geq 3$). Violates NF and C, by construction predicting outside Y_D on the extremal case.

Label-swap. Predictor $(x_1 \mapsto y_2, x_2 \mapsto y_1)$, swapping the labels. Satisfies NF and IS. Violates C, since the prediction at x_1 should be y_1 but is y_2.

	$L(D)(x_2) =$		
$L(D)(x_1) =$	y_1	y_2	y_3
y_1	(y_1, y_1)	(y_1, y_2) *(DC)*	(y_1, y_3)
y_2	(y_2, y_1) *(Sw)*	(y_2, y_2)	(y_2, y_3)
y_3	(y_3, y_1)	(y_3, y_2)	(y_3, y_3) *(ML)*

Legend: *(DC)* data-completion; *(Sw)* label-swap; *(ML)* missing-label compositional.

Table 2: The nine universal label-symmetric learners at $|X| = 2$, $|Y| = 3$, indexed by the predictor at the two-label orbit representative $D = \{(x_1, y_1), (x_2, y_2)\}$. The three highlighted cells are the learners singled out in the text above; the remaining six cells are the partial-axiom learners discussed in the following paragraph.

The remaining six learners realize partial axiom combinations (NF + CL, NF + IS, IS + CL, and so on); together they witness the partial-subset consistencies that organize the lattice. Four of the nine satisfy none of the axioms beyond universality, label-symmetry, and what is automatic at $|X| = 2$; these are the "pure" label-symmetric learners with no extra structure. This is easily verified by checking each cell against the non-automatic axioms C, IS, NF, CL (Idem, Mon, and Comp are automatic at $|X| = 2$, and Loc is incompatible with universality plus label-symmetry); the four "pure" cells are exactly the off-diagonal predictors mixing one seen label with the unseen y_3, namely (y_1, y_3), (y_2, y_3), (y_3, y_1), (y_3, y_2).

The enumeration illustrates how Theorem 38 stratifies under the eight axioms beyond universality and label-symmetry: each combination of axioms picks out a (possibly empty) stratum of learners, and the consistency lattice of Section 9 records exactly which combinations are nonempty. Together, the framework admits exactly three coherent learner profiles, each with a concrete witness.

9.6 Self-Training Closure

The self-training closure axiom CL captures a fixpoint property of empirical self-training: retraining on the learner's own predictions yields the same predictor (Lee, 2013). This section gives a positive characterization of self-training-closed learners on the unambiguous domain.

Definition 72 (Self-Training Closure Operator)**.** *For a universal learner L on $\mathcal{D}_u \setminus \{\emptyset\}$ and a nonempty unambiguous D, the* closure *of D under L is*

$$c_L(D) := \{(x, L(D)(x)) \mid x \in X\},$$

the total dataset on X obtained by completing D with L's predictions at unseen inputs.

For nonempty unambiguous D and $S \subseteq X$ with $\mathrm{dom}(D) \subseteq S$, the dataset

$$D_S^L := \{(x, L(D)(x)) \mid x \in S\}$$

sits in between: $D \subseteq D_S^L \subseteq c_L(D)$ as partial functions, with $D_S^L = D$ at $S = \mathrm{dom}(D)$ (by consistency) and $D_S^L = c_L(D)$ at $S = X$.

Theorem 73 (Characterization of Self-Training Closure)**.** *Let L be a universal consistent learner on $\mathcal{D}_u \setminus \{\emptyset\}$. The following are equivalent.*

(i) *L is closed under self-training: $L(D_S^L) = L(D)$ for every nonempty unambiguous D and every $S \subseteq X$ with* $\mathrm{dom}(D) \subseteq S$.

(ii) *The closure operator c_L is* extensive *($D \subseteq c_L(D)$) and* idempotent *($c_L(c_L(D)) = c_L(D)$), and L is* constant on the closure interval*: $L(D') = L(D)$ for every nonempty unambiguous D' with $D \subseteq D' \subseteq c_L(D)$.*

Proof *(i) ⇒ (ii).* For extensiveness, consistency gives $L(D)(x) = D(x)$ on $\mathrm{dom}(D)$, so $c_L(D)$ contains D.

For the interval property, let $D \subseteq D' \subseteq c_L(D)$ be nonempty unambiguous and set $S := \mathrm{dom}(D')$. We claim $D' = D_S^L$. For $x \in \mathrm{dom}(D)$, $D'(x) = D(x) = L(D)(x)$ by $D \subseteq D'$ and consistency. For $x \in S \setminus \mathrm{dom}(D)$, $D'(x) = c_L(D)(x) = L(D)(x)$ by $D' \subseteq c_L(D)$. Hence, $D' = D_S^L$, and (i) gives $L(D') = L(D)$. (Note that $D' = D_S^L$ is indeed unambiguous, since $D' \subseteq c_L(D)$ and $c_L(D)$ is a function.)

For idempotence, taking $S = X$ in (i) yields $L(c_L(D)) = L(D)$; then

$$c_L(c_L(D)) = \{(x, L(c_L(D))(x)) \mid x \in X\} = \{(x, L(D)(x)) \mid x \in X\} = c_L(D).$$

(ii) ⇒ (i). For any $S \subseteq X$ with $\mathrm{dom}(D) \subseteq S$, we verify that the dataset D_S^L sits in the interval $[D, c_L(D)]$. For $x \in \mathrm{dom}(D) \subseteq S$, consistency gives $D_S^L(x) = L(D)(x) = D(x)$, so $D \subseteq D_S^L$; and $D_S^L \subseteq c_L(D)$ since $S \subseteq X$. The set D_S^L is unambiguous because $L(D)$ is a function. The interval property gives $L(D_S^L) = L(D)$. ■

The theorem identifies self-training closure with a lattice property of L on the partial-function poset $\mathcal{D}_u$: the structural content lives on the entire closure interval $[D, c_L(D)]$ above each D, not merely at its endpoint $c_L(D)$. Equivalently, L factors through the closure operator c_L. Endpoint self-stability $L(c_L(D)) = L(D)$ alone is strictly weaker than the full closure axiom; the structural content of CL is the constancy of L on the entire interval.

9.7 Monotonicity

The monotonicity axiom Mon is a stability property: adding an example (x', y') to D should not remove y' from any prediction. We record its place within the framework.

Proposition 74 (Under Closure, Monotonicity Is Vacuous on the Closure Interval)**.** *Let L be a universal consistent learner satisfying self-training closure. Then for every D, D' with $D \subseteq D' \subseteq c_L(D)$, $L(D') = L(D)$, and the labels added in passing from D to any such D' match $L(D)$'s predictions exactly. In particular, monotonicity is* vacuously satisfied *on the closure interval $[D, c_L(D)]$, since every added example carries a label that $L(D)$ already predicts: no separate monotonicity assumption is needed.*

Proof By Theorem 73, self-training closure makes L constant on the closure interval $[D, c_L(D)]$. Each example added in passing from D to a $D' \in [D, c_L(D)]$ has the form $(x, L(D)(x))$. The added label is therefore exactly $L(D)$'s own prediction at x, and the monotonicity hypothesis is vacuous on the closure interval, since the labels added are precisely those $L(D)$ already predicts. ■

Monotonicity in full generality remains a constraint relating $L(D)$ and $L(D \cup \{(x', y')\})$ for arbitrary (x', y'), but on the specific interval $[D, c_L(D)]$ it is vacuous under closure. This is the precise compatibility statement between the two axioms: CL does not just coexist with Mon, it makes Mon vacuous on the closure interval where it would otherwise need to be checked. Both axioms are satisfied by the witness learners of Theorem 68.

10 Related Work

The paper draws on equivariant machine learning and axiomatic decision theory. What is new here is a complete structural account of classification under the single most natural symmetry hypothesis (label-permutation), without any probabilistic or distributional assumption, recovering four impossibility theorems as instances of a single fixpoint calculation.

10.1 Equivariant Machine Learning

Group equivariance as a structural constraint on learners has a long history in machine learning, beginning with Cohen and Welling (2016) and the DeepSets formalism of Zaheer et al. (2017) for permutation-invariant set functions; a general theoretical scaffolding for equivariant architectures is given by Kondor and Trivedi (2018), the probabilistic version by Bloem-Reddy and Teh (2020), and a broad unified treatment by Bronstein et al. (2022). The relationship between group actions, parameter-sharing, and equivariance is studied in Ravanbakhsh et al. (2017), and universality of invariant networks in Maron et al. (2019). The modern literature also spans the question of when equivariance bottlenecks expressivity (Finkelshtein et al., 2025).

The standard direction is equivariance under groups acting on *input* spaces, with the goal of parameter-efficient architectures exploiting input symmetries. We study instead the equivariance of the entire learner $L\colon \mathcal{D} \to Y^X$ under the label-permutation action $\mathrm{Sym}(Y)$ on the *output* side. The Bloem-Reddy-Teh framework comes closest: their probabilistic representation theorem characterizes G-equivariant conditional distributions for compact G acting on X. Our setting specializes to finite Y, finite $G = \mathrm{Sym}(Y)$ acting by permutation, and deterministic learners; the no-fabrication phenomenon is a discrete cardinality obstruction with no probabilistic counterpart, and the soft-label extension of Section 8.5 bridges the two frameworks. Theorem 38 is an instance of the standard orbit-decomposition of G-equivariant set maps (Serre, 1977, Chapter 1), with stabilizers identified as Young subgroups of $\mathrm{Sym}(Y)$; see James and Kerber (1981) for the underlying representation theory. The closed-form count involves Bell and Stirling numbers (Stanley, 2011, § 1.4 and § 1.9). A practical reading of the constraint theorem is that it gives a group-theoretic explanation for why label-symmetric architectures avoid fabricating unseen classes, the phenomenon formalized as open-set recognition by Scheirer et al. (2013).

10.2 Axiomatic Decision Theory

Axiomatic decision theory characterizes specific decision rules by exhibiting lists of properties that they, and they alone, satisfy: Gilboa and Schmeidler (1995, 2003) on case-based decision making, Maniquet and Mongin (2016) on compromise functions, Pennock et al. (2000) on aggregation, and the classification-aggregation impossibility of Cailloux et al. (2024); the social choice framework of Arrow (1951) and May's characterization of majority rule (May, 1952) are the foundational instances. Our setting differs in two ways. The axioms are structural rather than rational-choice motivated; we do not assume the learner optimizes any utility. The results characterize the full *space* of learners satisfying given axioms, rather than selecting a unique learner. Within these two differences, the two traditions remain in contact: the rigidity theorem is an axiomatic characterization of full label-symmetry by its no-fabrication consequence on the generic case, exactly in the spirit of axiomatic decision theory. The self-training closure axiom captures the fixpoint property of iterative pseudo-labeling (Lee, 2013); we characterize it (Theorem 73) as constancy of the learner on closure intervals in the partial-function poset.

The classification-aggregation results of Maniquet and Mongin (2016) and Cailloux et al. (2024) are the closest precedents in this tradition: both study aggregation functions on classifications under conditions analogous to label-symmetry, and both obtain impossibility-type conclusions. Our framework connects to these results, since aggregation can be cast as the special case in which the input is a tuple of classifier outputs rather than a labeled dataset, and the generic constraint

theorem is related to the Maniquet-Mongin dictatorship characterization. The Maniquet-Mongin dictatorship-or-constant result on Y^2 at $|Y| \geq 4$ is precisely the zero-count case of Lemma 52 at $k = 2$: no symmetric $\mathrm{Sym}(Y)$-equivariant function $Y^{(2)} \to Y$ exists outside $|Y| = 3$. The same fixpoint engine underlies the author's companion work on tie-breaking (Feys, 2026), where the joint stabilizer of the input determines the canonical orbit partition into which any reasonable strict tie-breaking rule decomposes.

11 Discussion

11.1 Principal Moral

The principal moral is that *label-symmetry is the right axiom for classification without external label structure*. The constraint theorem says nothing stronger is needed for no-fabrication on the generic case; the rigidity theorem says nothing weaker suffices; the characterization theorem makes the entire space of such classifiers explicit, as a product over orbits with a closed-form count. Strengthening the symmetry produces impossibility (the double-symmetry corollary); weakening it produces fabrication (the rigidity witnesses). Label-symmetry sits at the structural balance point: necessary, sufficient, and complete for no-fabrication on the generic case. Four impossibility theorems collapse into a single fixpoint principle, and the compositional trichotomy at $|Y| = 3$ emerges as the unique extremal exception, governed by the core lemma on $\mathrm{Sym}(Y)$-equivariant symmetric functions. The total parameter space of universal label-symmetric learners is explicitly computable: a closed form in Bell and Stirling numbers (Corollary 42) enumerates it for every finite $|X|, |Y|$, and at $|X| = 2$, $|Y| = 3$ there are exactly nine such learners, tabulated in Table 2.

The practical reach of the constraint theorem is broader than the generic hypothesis suggests: in most multi-class applications $|Y_D| \leq |Y| - 2$ is satisfied by construction, since real classification problems have label sets of size in the hundreds or thousands; ImageNet has 1000 classes, and any dataset missing two or more of them is in the generic regime. For binary classification ($|Y| = 2$), the generic case is empty and the constraint is vacuous; the framework's main content concerns multi-class settings, and applies as soon as the label count exceeds the observed labels by two or more.

11.2 What the Theorems Do and Do Not Show

The formal scope is at once broad in one direction and narrow in another. Broad: the procedures covered include softmax cross-entropy, k-nearest-neighbors, decision trees with categorical splits, random forests, and kernel methods with label-symmetric kernels; each is universal and label-symmetric in the senses used here. Narrow: the framework requires unambiguous training data, finite label set Y, and treats deterministic learning procedures (data $\to$ predictor) rather than trained predictors in isolation.

The theorems do not establish several things one might initially infer. They are *pre-statistical*: they say nothing about generalization, sample complexity, training dynamics, or optimization landscape, only what predictions are formally permitted to do. They have limits in their formal hypotheses: they do not apply on the extremal case, on the empty dataset, on ambiguous training data, or to continuous or semantically structured label spaces (an order, a metric, an embedding, a hierarchy), each of which breaks label-symmetry and produces an explicit fabricating witness by the rigidity theorem. The theorems also characterize learning *procedures*, not the predictor obtained on a particular run; see Section 2 for the precise sense in which label-symmetry is a property of the procedure, not of any individual stochastic realization.

11.3 Scope, Limitations, and Future Directions

The framework is discrete, finite, and combinatorial; probability, generalization, sample complexity, and optimization are deliberately outside its scope.

Three natural extensions go beyond this scope, each suggested by a specific feature of the framework.

Representation-theoretic decomposition of the soft-label space. The soft-label constraint theorem (Theorem 61) characterizes the $\mathrm{Sym}(Y)$-invariant slice of $\Delta(Y)$ at each orbit but does not exploit the full representation theory of $\mathrm{Sym}(Y)$. A finer decomposition using Schur-Weyl duality for $\mathrm{Sym}(Y)$ and the partition algebra acting on the predictor space would yield an irreducible-component decomposition of the parameter space of universal label-symmetric soft-label learners, refining the present uniform-mass decomposition; the Fourier-analytic machinery for such an analysis is developed in Diaconis (1988).

Aggregation of learners. When given universal label-symmetric learners $L_1, L_2, \ldots, L_n$, which aggregators $A(L_1, \ldots, L_n)$ are again universal and label-symmetric? The discrete combinatorics of the present framework, together with the rigidity theorem as a non-dictatorship analogue, suggests an Arrow-style axiomatic characterization (Arrow, 1951). This is the natural next paper.

Subgroup-parametrized framework. The rigidity theorem identifies, for every proper subgroup $G \subsetneq \mathrm{Sym}(Y)$, the obstruction that lets a G-equivariant universal learner fabricate. The substantive regime, in which the framework has most to say, is thus partial label-symmetry: hierarchical, ordinal, and embedded label sets, each carrying a strict subgroup of $\mathrm{Sym}(Y)$. A systematic study of these partial-symmetry regimes, with the present framework as the $\mathrm{Sym}(Y)$-saturated baseline, is the program the paper is meant to start.

Acknowledgments and Disclosure of Funding

The author declares no funding sources and no competing interests.

References

K. J. Arrow. *Social Choice and Individual Values*. John Wiley & Sons, 1951.

B. Bloem-Reddy and Y. W. Teh. Probabilistic Symmetries and Invariant Neural Networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*. MIT Press, 2022. Also available as arXiv:2104.13478.

O. Cailloux, M. Hervouin, A. I. Özkes, and M. R. Sanver. Classification aggregation without unanimity. *Mathematical Social Sciences*, 128:6–9, 2024.

T. Cohen and M. Welling. Group Equivariant Convolutional Networks. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 2990–2999, 2016.

P. Diaconis. *Group Representations in Probability and Statistics*. Institute of Mathematical Statistics Lecture Notes, Monograph Series, Vol. 11. Institute of Mathematical Statistics, Hayward, CA, 1988.

F. M. V. Feys. An Axiomatic Theory of Tie-Breaking: Impossibility, Characterization, and Decomposition. Preprint, arXiv:2605.22846. Submitted to *Social Choice and Welfare*, 2026.

B. Finkelshtein, İ. İ. Ceylan, M. Bronstein, and R. Levie. Equivariance Everywhere All At Once: A Recipe for Graph Foundation Models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

I. Gilboa and D. Schmeidler. Case-Based Decision Theory. *The Quarterly Journal of Economics*, 110(3):605–639, 1995.

I. Gilboa and D. Schmeidler. Inductive Inference: An Axiomatic Approach. *Econometrica*, 71(1):1–26, 2003.

G. D. James and A. Kerber. *The Representation Theory of the Symmetric Group*. Encyclopedia of Mathematics and Its Applications, Vol. 16. Addison-Wesley, Reading, MA, 1981.

R. Kondor and S. Trivedi. On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2747–2755, 2018.

D.-H. Lee. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *ICML 2013 Workshop on Challenges in Representation Learning*, 2013.

F. Maniquet and P. Mongin. A theorem on aggregating classifications. *Mathematical Social Sciences*, 79:6–10, 2016.

H. Maron, E. Fetaya, N. Segol, and Y. Lipman. On the Universality of Invariant Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4363–4371, 2019.

K. O. May. A Set of Independent Necessary and Sufficient Conditions for Simple Majority Decision. *Econometrica*, 20(4):680–684, 1952.

D. M. Pennock, P. Maynard-Reid II, C. L. Giles, and E. Horvitz. A Normative Examination of Ensemble Learning Algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pages 735–742, 2000.

S. Ravanbakhsh, J. Schneider, and B. Póczos. Equivariance Through Parameter-Sharing. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 2892–2901, 2017.

W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.

J.-P. Serre. *Linear Representations of Finite Groups*. Graduate Texts in Mathematics 42. Springer, 1977.

R. P. Stanley. *Enumerative Combinatorics, Volume 1*. Cambridge University Press, 2nd ed., 2011.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 3391–3401, 2017.