PROBABILISTIC SYMMETRY AND INVARIANT NEURAL NETWORKS

Benjamin Bloem-Reddy and Yee Whye Teh

University of Oxford

In an effort to improve the performance of deep neural networks in data-scarce, non-i.i.d., or unsupervised settings, much recent research has been devoted to encoding invariance under symmetry transformations into neural network architectures. We treat the neural network input and output as random variables, and consider group invariance from the perspective of probabilistic symmetry. Drawing on tools from probability and statistics, we establish a link between functional and probabilistic symmetry, and obtain generative functional representations of joint and conditional probability distributions that are invariant or equivariant under the action of a compact group. Those representations completely characterize the structure of neural networks that can be used to model such distributions and yield a general program for constructing invariant stochastic or deterministic neural networks. We develop the details of the general program for exchangeable sequences and arrays, recovering a number of recent examples as special cases.

1. Introduction. Neural networks and deep learning methods have found success in a wide variety of applications. Much of the success has been attributed to a confluence of trends, including the increasing availability of data; advances in specialized hardware such as GPUs and TPUs; and open-source software like Theano (Theano Development Team, 2016), Tensorflow (Abadi et al., 2015), and PyTorch (Paszke et al., 2017), that enable rapid development of neural network models through automatic differentiation and high-level interfaces with specialized hardware. Neural networks have been most successful in settings with massive amounts of i.i.d. labeled training data. In recent years, a concerted research effort has aimed to improve the performance of deep learning systems in data-scarce and semi-supervised or unsupervised problems, and for structured, non-i.i.d. data. In that effort, there has been a renewed focus on novel neural network architectures: attention mechanisms (Vaswani et al., 2017), memory networks (Sukhbaatar et al., 2015), dilated convolutions (Yu and Koltun, 2016), residual networks (He et al., 2016), and graph neural networks (Scarselli et al., 2009) are a few recent examples from the rapidly expanding literature.

The focus on novel architectures reflects a basic fact of machine learning: in the presence of data scarcity, whether due to small sample size or unobserved labels, or of complicated structure, the model must pick up the slack. Amid the flurry of model-focused innovation, there is a growing need for a framework for encoding modeling assumptions and checking their validity, and for assessing the training stability and generalization potential of an architecture. In short, a principled theory of neural network design is needed.

This paper represents one small step in that direction. It concerns the development of neural network architectures motivated by symmetry considerations, typically described by invariance or equivariance with respect to the action of a group. The most well-known examples of such architectures are convolutional networks (CNNs) (LeCun et al., 1989), which ensure invariance of the output Y under translations of an input image X. Other examples include neural networks that encode rotational invariance (Cohen et al., 2018) or permutation invariance (Zaheer et al., 2017; Hartford et al., 2018; Herzig et al., 2018; Lee et al., 2018).

Within this growing body of work, symmetry is most often addressed by designing a specific neural network architecture for which the relevant invariance can be verified. A more general approach aims to answer the question:

> *For a particular symmetry property, can all invariant neural network architectures be characterized?*

General results have been less common in the literature; important exceptions include characterizations of feed-forward networks that are invariant under the action of discrete groups (Shawe-Taylor, 1989; Ravanbakhsh, Schneider and Póczos, 2017), finite linear groups (Wood and Shawe-Taylor, 1996), or compact groups (Kondor and Trivedi, 2018).

In the probability and statistics literature, there is a long history of probabilistic model specification motivated by symmetry considerations. In particular, if a random variable X is to be modeled as respecting some symmetry property, then a model should only contain distributions P_X that are invariant. The relevant theoretical question is:

> *For a particular symmetry property of X, can all invariant distributions P_X be characterized?*

Work in this area dates at least to the 1930s, and the number of general results reflects the longevity of the field. The most famous example is de Finetti's theorem (de Finetti, 1930), which is a cornerstone of Bayesian statistics and machine learning. It shows that all infinitely exchangeable (i.e., permutation-invariant) sequences of random variables have distributional representations that are conditionally i.i.d., conditioned on a random probability measure. Other examples include rotational invariance, translation invariance, and a host of others.

In the present work, we approach the question of invariance in neural network architectures from the perspective of probabilistic symmetry. In particular, we seek to understand the symmetry properties of probability distributions that are necessary and sufficient for a *generative functional representation* $Y = f(\eta, X)$, with generic noise variable η, such that f obeys the relevant functional symmetries. Our approach sheds light on the core statistical issues involved, and provides a broader view of the questions posed above: from considering classes of deterministic functions to stochastic ones; and from invariant marginal distributions to invariant conditional distributions.

We develop a general program to answer these questions, putting functional symmetries in correspondence with probabilistic symmetries. As a result, we characterize the structure of invariant stochastic neural networks that correspond to a distributional symmetry. In doing so, we recover a number of results for invariant deterministic neural networks as special cases. To demonstrate, we apply the general program to the detailed design of invariant neural networks for exchangeable sequences and arrays.

Outline. The remainder of this section provides further background and related work, and introduces the necessary measure theoretic technicalities and notation. Section 2 defines the relevant functional and probabilistic notions of symmetries; the similarities between the two suggests that there is a mathematical link. Section 3 makes that link by establishing generative functional representations of conditional distributions that are invariant or equivariant under the action of a compact group. Those results fit in a larger statistical framework; Section 4 relates them to the concepts of statistical sufficiency and adequacy. Section 5 provides the general program for designing invariant neural networks, and Sections 6 and 7 develop the details of the program for exchangeable sequences and arrays.

1.1. *Symmetry in deep learning.* Interest in neural networks that are invariant to discrete groups acting on the network nodes dates back at least to the text of Minsky and Papert (1988) on single-layer perceptrons (SLPs), who used their results to demonstrate a limitation of SLPs. Shawe-Taylor (1989, 1993) extended the theory to multi-layer perceptrons, under the name Symmetry Networks. The main findings of that theory, that invariance is achieved by weight-preserving automorphisms of the neural network, and that the connections between layers must be partitioned into weight-sharing orbits, were rediscovered by Ravanbakhsh, Schneider and Póczos (2017), who proposed novel architectures and new applications.

Wood and Shawe-Taylor (1996) extended the theory to invariance of feed-forward networks under the action of finite linear groups. Some of their results overlap with results for compact groups found in Kondor and Trivedi (2018), including the characterization of equivariance in feed-forward networks in terms of group theoretic convolution.

The most widely applied invariant neural architecture is the CNN for input images. Recently, there has been a surge of interest in generalizing the idea of invariant architectures to other data domains such as sets and graphs, with most work belonging to either of two categories:

(i) properly defined convolutions (Bruna et al., 2014; Duvenaud et al., 2015; Niepert, Ahmed and Kutzkov, 2016); or
(ii) equivariance under the action of groups that lead to weight-tying schemes (Cohen and Welling, 2016; Ravanbakhsh, Schneider and Póczos, 2017).

Both of these approaches rely on group theoretic structure in the set of symmetry transformations, and Kondor and Trivedi (2018) used group representation theory to show that the two approaches are the same.

Specific instantiations of invariant architectures abound in the literature; they are too numerous to collect here. However, we give a number of examples in Sections 6 and 7 in the context of sequence- and graph-valued input data that is invariant under permutations.

1.2. *Symmetry in probability and statistics.* The study of probabilistic symmetries has a long history. Laplace's "rule of succession" dates to 1774; it is the conceptual precursor to exchangeability (see Zabell, 2005, for a historical and philosophical account). Other examples include invariance under rotation and stationarity in time (Freedman, 1963); the former has roots in Maxwell's work in statistical mechanics in 1875 (see, for example, the historical notes in Kallenberg, 2005).

The canonical probabilistic symmetry is exchangeability. A sequence of random variables, $\mathbf{X}_n = (X_1, \ldots, X_n)$, is exchangeable if its distribution is invariant under all permutations of the elements. If that is true for every $n \in \mathbb{N}$ in an infinite sequence $\mathbf{X}_{\mathbb{N}}$, then the sequence is said to be *infinitely exchangeable.* de Finetti's theorem (de Finetti, 1930) shows that infinitely exchangeable distributions have particularly simple structure. Specifically, $\mathbf{X}_{\mathbb{N}}$ is infinitely exchangeable if and only if there exists some random distribution Q, such that the elements of $\mathbf{X}_{\mathbb{N}}$ are conditionally i.i.d. with distribution Q. Therefore, each infinitely exchangeable distribution P has an integral decomposition: there is a unique (to P) distribution ν on the set $\mathcal{M}_1(\mathcal{X})$ of all probability measures on $\mathcal{X}$, such that

$$P(\mathbf{X}_{\mathbb{N}}) = \int_{\mathcal{M}_1(\mathcal{X})} Q^{\infty}(\mathbf{X}_{\mathbb{N}})\nu(dQ) \quad \text{with} \quad Q^{\infty}(\mathbf{X}_{\mathbb{N}}) = \prod_{i=1}^{\infty} Q(X_i) \ . \tag{1}$$

The simplicity is useful: by assuming the data are infinitely exchangeable, only models that have a conditionally i.i.d. structure need to be considered.

de Finetti's theorem is a special case of a more general mathematical result, the ergodic decomposition theorem, which puts probabilistic symmetries in correspondence with integral decompositions like (1); see Orbanz and Roy (2015) for an accessible overview. de Finetti's results inspired a large body of work on other symmetries in the probability literature; Kallenberg (2005) gives a comprehensive treatment and Kallenberg (2017) contains further results. Applications of group symmetries in statistics include equivariant estimation and testing; see, for example, (Lehmann and Romano, 2005, Ch. 6); Eaton (1989); Wijsman (1990); Giri (1996).

1.3. *Measure theoretic technicalities and notation.* In order to keep the presentation as clear as possible, we aim to minimize measure theoretic technicalities. Throughout, it is assumed that there is a background probability space $(\Omega, \mathcal{A}, \mathbb{P})$ that is rich enough to support all required random variables. All random variables are assumed to take values in standard Borel spaces (spaces that are Borel isomorphic to a Borel subset of the unit interval (see, e.g., Kallenberg, 2002)). For example,

X is a $\mathcal{X}$-valued random variable in $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, where $\mathcal{B}_{\mathcal{X}}$ is the Borel σ-algebra of $\mathcal{X}$. Alternatively, we may say that X is a random element of $\mathcal{X}$. For notational convenience, for a $\mathcal{Y}$-valued random variable Y, we write $P(Y \in \bullet \mid X)$ as shorthand for, "for all sets $A \in \mathcal{B}_{\mathcal{Y}}$, $P(Y \in A \mid \sigma(X))$", where $\sigma(X)$ is the σ-algebra generated by X. We use $\mathcal{M}(\mathcal{X})$ to denote the set of measures on $\mathcal{X}$, and $\mathcal{M}_1(\mathcal{X})$ to denote the set of probability measures on $\mathcal{X}$. Many of our results pertain to conditional independence relationships; $Y \perp\!\!\!\perp_Z X$ means that Y and X are conditionally independent, given $\sigma(Z)$. Finally, $\overset{\text{d}}{=}$ denotes equality in distribution, and $\overset{\text{a.s.}}{=}$ denotes almost sure equality.

2. Functional and probabilistic symmetries.

We consider the relationship between two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, with Y being a predicted output based on input X. For example, in image classification, X might be an image and Y a class label; in sequence prediction, $X = (X_i)_{i=1}^n$ might be a sequence and Y the next element, X_{n+1}, to be predicted; and in a variational autoencoder, X might be an input vector and Y the corresponding latent variable whose posterior is to be inferred using the autoencoder. Throughout, we denote the joint distribution of both variables as $P_{X,Y}$, and the conditional distribution $P_{Y|X}$ is the primary object of interest.

Two basic notions of symmetry are considered, as are the connections between them. The first notion, defined in Section 2.1, is functional, and is most relevant when Y is a deterministic function of X, say $Y = f(X)$; the symmetry properties pertain to the function f. The second notion, defined in Section 2.2, is probabilistic, and pertains to the conditional distribution of Y given X.

In both cases, the symmetries are induced by the action of a group $\mathcal{G}$.[1] Let $\Phi_{\mathcal{X}} : \mathcal{G} \times \mathcal{X} \to \mathcal{X}$ be the left-action of $\mathcal{G}$ on the input space $\mathcal{X}$, such that $\Phi_{\mathcal{X}}(e, x) = x$ is the identity mapping for all $x \in \mathcal{X}$, and $\Phi_{\mathcal{X}}(g, \Phi_{\mathcal{X}}(h, x)) = \Phi_{\mathcal{X}}(g \cdot h, x)$ for $g, h \in \mathcal{G}$, $x \in \mathcal{X}$. For convenience, we write $g \cdot x = \Phi_{\mathcal{X}}(g, x)$. Similarly, let $\Phi_{\mathcal{Y}}$ be the action of $\mathcal{G}$ on the output space $\mathcal{Y}$, and $g \cdot y = \Phi_{\mathcal{Y}}(g, y)$ for $g \in \mathcal{G}$, $y \in \mathcal{Y}$. A group $\mathcal{G}$, along with a σ-algebra $\sigma(\mathcal{G})$, is said to be measurable if the group operations of inversion $g \mapsto g^{-1}$ and composition $(g, g') \mapsto g \cdot g'$ are $\sigma(\mathcal{G})$-measurable. $\mathcal{G}$ *acts measurably* on $\mathcal{X}$ if $\Phi_{\mathcal{X}}$ is a measurable function from $\sigma(\mathcal{G}) \otimes \mathcal{B}_{\mathcal{X}} \to \mathcal{B}_{\mathcal{X}}$ (Kallenberg, 2017).

2.1. *Functional symmetry in neural networks.*

In many machine learning settings, a prediction Y based on an input X is modeled as a deterministic function, $Y = f(X)$, where f belongs to some function class $\mathcal{F} = \{f; f : \mathcal{X} \to \mathcal{Y}\}$, often satisfying some further conditions. For example, f might belong to a Reproducing Kernel Hilbert Space, or to a subspace of all strongly convex functions. Alternatively, f is a neural network parameterized by weights and biases collected into a parameter vector $\boldsymbol{\theta}$, in which case the function class $\mathcal{F}$ corresponds to the chosen

[1] A group $\mathcal{G}$ is a set and a composition operator $\cdot$ with three properties: for each $g, h \in \mathcal{G}$ the composition $g \cdot h \in \mathcal{G}$ is in the group; there is an identity element $e \in \mathcal{G}$; and for each $g \in \mathcal{G}$ there is an inverse $g^{-1} \in \mathcal{G}$ such that $g^{-1} \cdot g = e$. See, e.g., Rotman (1995).

network architecture, and a particular f corresponds to a particular set of values for $\boldsymbol{\theta}$. We are concerned with implications on the choice of the network architecture (equivalently, the function class $\mathcal{F}$) due to symmetry properties we impose on the input-output relationship. In this deterministic setting, the conditional distribution $P_{Y|X}$ is simply a point mass at $f(X)$.

Two properties, invariance and equivariance, formalize the relevant symmetries. A function $f : \mathcal{X} \to \mathcal{Y}$ is *invariant under* $\mathcal{G}$, or $\mathcal{G}$*-invariant*, if the output is unchanged by transformations of the input induced by the group:

$$f(g \cdot x) = f(x) \quad \text{for all } g \in \mathcal{G},\ x \in \mathcal{X}\ . \tag{2}$$

Alternatively, a function $f : \mathcal{X} \to \mathcal{Y}$ is *equivariant under* $\mathcal{G}$, or $\mathcal{G}$*-equivariant*, if

$$f(g \cdot x) = g \cdot f(x) \quad \text{for all } g \in \mathcal{G},\ x \in \mathcal{X}\ . \tag{3}$$

The action of $\mathcal{G}$ commutes with the application of an equivariant f; transforming the input is the same as transforming the output. Note that the action of $\mathcal{G}$ on $\mathcal{X}$ and $\mathcal{Y}$ may be different. In particular, invariance is a special case of equivariance, whereby the group action in the output space is trivial: $g \cdot y = y$ for each $g \in \mathcal{G}$ and $y \in \mathcal{Y}$. Invariance imposes stronger restrictions on the functions satisfying it, as compared to equivariance.

These properties and their implications on network architectures are illustrated with examples from the literature. For notational convenience, let $[n] = \{1, \dots, n\}$ and $\mathbf{X}_n = (X_1, \dots, X_n) \in \mathcal{X}^n$. Finally, denote the finite symmetric group of a set of n elements (i.e., the set of all permutations of $[n]$) by $\mathbb{S}_n$.

Example 1 (Deep Sets: $\mathbb{S}_n$-invariant functions of sequences). Zaheer et al. (2017) considered a model $Y = f(\mathbf{X}_n)$, where the input $\mathbf{X}_n$ was treated as a set, i.e., the order among its elements did not matter. Those authors required that the output of f be unchanged under all permutations of the elements of $\mathbf{X}_n$, i.e., that f is $\mathbb{S}_n$-invariant. They found that f is $\mathbb{S}_n$-invariant if and only if it can be represented as $f(\mathbf{X}_n) = \tilde{f}(\sum_{i=1}^n \phi(X_i))$ for some functions $\tilde{f}$ and ϕ. Clearly, permutations of the elements of $\mathbf{X}_n$ leave such a function invariant: $f(\mathbf{X}_n) = f(\pi \cdot \mathbf{X}_n)$ for $\pi \in \mathbb{S}_n$. The fact that *all* $\mathbb{S}_n$-invariant functions can be expressed in such a form was proved by Zaheer et al. (2017) in two different settings: (i) for sets of arbitrary size when $\mathcal{X}$ is countable; and (ii) for sets of fixed size when $\mathcal{X}$ is uncountable. In both cases, ϕ uniquely encodes the elements of $\mathcal{X}$. Essentially, ϕ is a generalization of a one-hot encoding, which gets sum-pooled and passed through $\tilde{f}$. The authors call neural architectures satisfying such structure *Deep Sets*. See Figure 1, left panel, for an example diagram.

In Section 6.1, we give a short and intuitive proof using basic properties of exchangeable sequences.

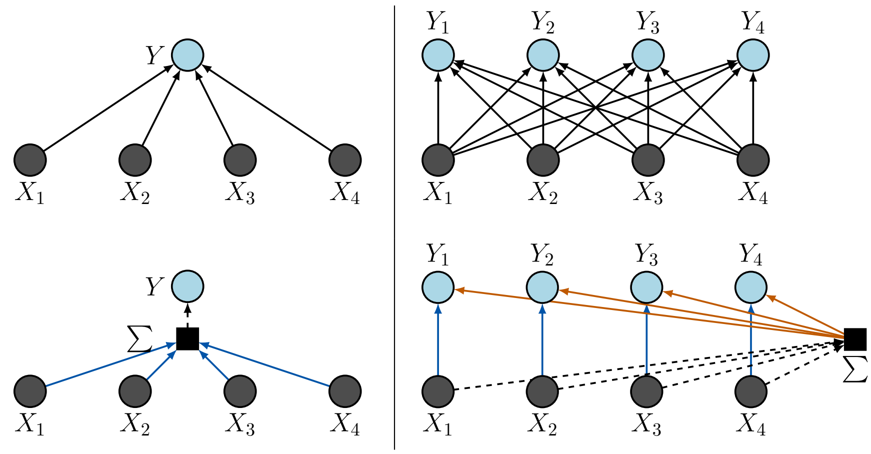


FIG 1. *Computation diagrams for the neural networks in Examples 1 and 2. Black solid arrows indicate general, uncoupled weights; black dashed arrows indicate a fixed function, i.e., an activation function; colored arrows indicate shared weights between arrows of the same color.* Left panel*: A general output layer with a different weight from each X_i to Y is shown on top, and a simple implementation of a $\mathbb{S}_n$-invariant architecture that shares weights is on the bottom.* Right panel*: A fully connected MLP with n^2 weights is shown on the top, and the $\mathbb{S}_n$-equivariant architecture corresponding to (4), with two weights, is on the bottom.*

EXAMPLE 2 ($\mathbb{S}_n$-equivariant neural network layers). Let $\mathbf{X}_n$ and $\mathbf{Y}_n$ represent adjacent layers of a standard feed-forward neural network, such that the nodes are indexed by $[n]$, with $X_i \in \mathbb{R}$ the ith node in layer $\mathbf{X}_n$, and similarly for Y_i. In a feed-forward layer, $\mathbf{Y}_n = \sigma(\Theta \mathbf{X}_n)$ where σ is an element-wise nonlinearity and Θ is a weight matrix (we ignore biases for simplicity). Shawe-Taylor (1989); Wood and Shawe-Taylor (1996); Zaheer et al. (2017) showed that the only weight matrices that lead to $\mathbb{S}_n$-equivariant layers are the sum of a diagonal matrix, $\theta_0 \mathbb{I}_n$, $\theta_0 \in \mathbb{R}$, and a constant one, $\theta_1 \mathbf{1}_n \mathbf{1}_n^T$, $\theta_1 \in \mathbb{R}$, such that

$$[\Theta \mathbf{X}_n]_i = \theta_0 X_i + \theta_1 \sum_{j=1}^{n} X_j \ . \tag{4}$$

Figure 1, right panel, shows the weight-sharing patterns of the connections.

These examples demonstrate that equivariance is less restrictive than invariance, and thus allows for more expressive parameterizations; in Example 2, invariance would require that $w_0 = 0$. At the same time, equivariance appears to be strong enough to greatly reduce the dimension of the parameter space: generic fully connected layers contain n^2 weights, compared to the two used by the $\mathbb{S}_n$-equivariant architecture.

At a high level, equivariance in feed-forward neural networks ensures that transformations of the input lead to predictable, symmetry-preserving transformations of higher layers, which allows the network to exploit the symmetry in all layers through weight-sharing (Cohen and Welling, 2016), and to capture structure at multiple scales through pooling (Kondor and Trivedi, 2018). In addition to being theoretically interesting, group invariance often indicates simplified architectures through parameter-sharing (e.g., Ravanbakhsh, Schneider and Póczos, 2017; Cohen and Welling, 2017). These in turn lead to simpler, stabler training and may lead to better generalization (Shawe-Taylor, 1991). (See also the discussion in Section 8.)

2.2. *Symmetry in conditional probability distributions.* An alternative approach to the deterministic models of Section 2.1 is to model the relationship between input X and output Y as stochastic, either by directly parameterizing the conditional distribution $P_{Y|X}$, or by defining a procedure for generating samples from $P_{Y|X}$. For example, the encoder network of a variational autoencoder computes an approximate posterior distribution over the latent variable Y given observation X; in classification, the use of a soft-max output layer is interpreted as a network which predicts a distribution over labels; in implicit models like Generative Adversarial Networks (Goodfellow et al., 2014) or simulation-based models without a likelihood (e.g., Gutmann and Corander, 2016), and in many probabilistic programming languages (e.g., van de Meent et al., 2018), $P_{Y|X}$ is not explicitly represented, but samples are generated and used to evaluate the quality of the model.

The relevant properties that encode symmetry in such settings are therefore probabilistic. One way to define symmetry properties for conditional distributions is by adding noise to invariant or equivariant functions. For example, if f is $\mathcal{G}$-invariant and η is a standard normal random variable independent of X, then $Y = \tilde{f}(X) + \eta$ corresponds to a conditional distribution $P_{Y|X}$ that is $\mathcal{G}$-invariant. This type of construction, with $Y = f(\eta, X)$ satisfying invariance (equivariance) in its second argument, will lead to invariant (equivariant) conditional distributions. While intuitive and constructive, the approach does not specify what probabilistic properties must be satisfied by random variables X and Y in order for such functions to exist. Furthermore, it leaves open the question of whether there are other approaches that may be used. To avoid these ambiguities, we define notions of symmetries for probability models directly in terms of the distributions.

The discussion on exchangeability in Section 1 pertains only to invariance of the marginal distribution P_X under $\mathbb{S}_n$. Suitable notions of probabilistic symmetry under more general groups are needed for the conditional distribution of Y given X. Let $\mathcal{G}$ be a group acting measurably on $\mathcal{X}$ and on $\mathcal{Y}$. The conditional distribution $P_{Y|X}$ of Y given X is $\mathcal{G}$-invariant if $Y|X \overset{\text{d}}{=} Y|g \cdot X$. More precisely, for all $A \in \mathcal{B}_{\mathcal{Y}}$ and $B \in \mathcal{B}_{\mathcal{X}}$ such that $P_X(B) > 0$,

$$P_{Y|X}(Y \in A \mid X \in B) = P_{Y|X}(Y \in A \mid g \cdot X \in B) \quad \text{for all } g \in \mathcal{G} \,. \tag{5}$$

On the other hand, $P_{Y|X}$ is $\mathcal{G}$-equivariant if $Y|X \overset{\text{d}}{=} g \cdot Y|g \cdot X$ for all $g \in \mathcal{G}$. That is, transforming X by g leads to the same conditional distribution of Y except that it is also transformed by g. More precisely, for all $A \in \mathcal{B}_{\mathcal{Y}}$ and $B \in \mathcal{B}_{\mathcal{X}}$ such that $P_X(B) > 0$,

$$P_{Y|X}(Y \in A \mid X \in B) = P_{Y|X}(g \cdot Y \in A \mid g \cdot X \in B) \quad \text{for all } g \in \mathcal{G} \,. \tag{6}$$

Typically, conditional invariance or equivariance is desired because of symmetries in the marginal distribution P_X. In Example 1, it was assumed that the ordering among the elements of the input sequence is unimportant, and therefore it is reasonable to assume that the marginal distribution of the input sequence is exchangeable. In general, we assume throughout the present work that X is marginally $\mathcal{G}$-invariant:

$$P_X(X \in A) = P_X(g \cdot X \in A) \quad \text{for all } g \in \mathcal{G}, A \in \mathcal{B}_{\mathcal{X}} \,. \tag{7}$$

Clearly, $\mathcal{G}$-invariance of P_X and of $P_{Y|X}$ will result in the joint invariance $P_{X,Y}(X,Y) = P_{X,Y}(g \cdot X, Y)$; similarly, if P_X is $\mathcal{G}$-invariant and $P_{Y|X}$ is $\mathcal{G}$-equivariant, then $P_{X,Y}(X,Y) = P_{X,Y}(g \cdot X, g \cdot Y)$. The converse is also true for sufficiently nice groups and spaces. Therefore, we may work with the joint distribution of X and Y, which is often more convenient than working with the marginal and conditional distributions. These ideas are summarized in the following proposition, which is a special case of more general results on invariant measures found in, for example, Kallenberg (2017, Ch. 7).

Proposition 1. *For a group $\mathcal{G}$ acting measurably on Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, if P_X is marginally $\mathcal{G}$-invariant then*

(i) $P_{Y|X}$ is conditionally $\mathcal{G}$-invariant if and only if $(X,Y) \overset{\mathrm{d}}{=} (g \cdot X, Y)$ for all $g \in \mathcal{G}$.
(ii) $P_{Y|X}$ is conditionally $\mathcal{G}$-equivariant if and only if $(X,Y) \overset{\mathrm{d}}{=} g \cdot (X,Y) := (g \cdot X, g \cdot Y)$ for all $g \in \mathcal{G}$, i.e., X and Y are jointly $\mathcal{G}$-invariant.

There is a simple algebra of compositions of equivariant and invariant functions. Specifically, compositions of equivariant functions are equivariant (equivariance is transitive under function composition), and composing an equivariant function with an invariant one yields an invariant function (Cohen and Welling, 2016; Kondor and Trivedi, 2018). Such compositions generate a grammar with which to construct elaborate functions with the desired symmetry properties. Such functions abound in the deep learning literature. These compositional properties carry over to the probabilistic case as well.

Proposition 2. *Let X, Y, Z be random variables such that $X \perp\!\!\!\perp_Y Z$. Suppose X is marginally $\mathcal{G}$-invariant.*

(i) If Y is conditionally $\mathcal{G}$-equivariant given X, and Z is conditionally $\mathcal{G}$-equivariant given Y, then Z is conditionally $\mathcal{G}$-equivariant given X.
(ii) If Y is conditionally $\mathcal{G}$-equivariant given X, and Z is conditionally $\mathcal{G}$-invariant given Y, then Z is conditionally $\mathcal{G}$-invariant given X.

PROOF. (i) Proposition 1 shows that X, Y are jointly $\mathcal{G}$-invariant, implying that Y is marginally $\mathcal{G}$-invariant. Proposition 1 applied to Y and Z shows that Y, Z are jointly $\mathcal{G}$-invariant as well. Joint invariance of X, Y, along with $X \perp\!\!\!\perp_Y Z$, implies that X, Y, Z are jointly $\mathcal{G}$-invariant. Marginalizing out Y, X, Z are jointly $\mathcal{G}$-invariant. Proposition 1 shows that Z is conditionally $\mathcal{G}$-equivariant given X.

(ii) A similar argument as above shows that $(g \cdot X, g \cdot Y, Z) \stackrel{\text{d}}{=} (X, Y, Z)$ for each $g \in \mathcal{G}$. Marginalizing out Y, we see that $(g \cdot X, Z) \stackrel{\text{d}}{=} (X, Z)$, so that Z is conditionally $\mathcal{G}$-invariant given X. □

3. Generative functional representations of probabilistic symmetries. The functional and probabilistic notions of symmetries described in Sections 1 and 2 represent two different approaches to achieving the same goal: a principled framework for constructing models from symmetry considerations. Despite their apparent similarities, it is not immediately clear whether there is a precise mathematical link. The connection between the probabilistic and functional notions of symmetry is explored in this section. The connection relies on the concept of *noise outsourcing* (Section 3.1), which can be understood as guaranteeing the existence of generative functional representations for distributions, and which has a particular refinement when conditional independence is present. That refinement is applied to the *disintegration over maximal invariants* of an invariant distribution P_X in order to establish generative functional representations of invariant and equivariant conditional probability distributions (Section 3.2).

3.1. *Noise outsourcing and conditional independence.* Noise outsourcing is a standard technical tool from measure theoretic probability, where it is also known by other names such as *transfer* (Kallenberg, 2002). For any two random variables X and Y taking values in "nice" spaces (e.g., Borel spaces), noise outsourcing says that there exists a generative functional representation of the conditional distribution $P_{Y|X}$ in terms of X and independent noise: $Y \stackrel{\text{a.s.}}{=} f(\eta, X)$. It can be viewed as a general version of the so-called reparameterization trick (Kingma and Welling, 2014; Rezende, Mohamed and Wierstra, 2014) for random variables taking values in general measurable spaces (not just $\mathbb{R}$).[2] The noise variable η acts as a generic source of randomness that is "outsourced", a term borrowed from Austin (2015). The relevant property of η is its independence from X, and the uniform distribution is not special in this regard. η could be replaced by any other random variable taking values in a Borel space, for example a standard normal, and the result would still hold, albeit with a different f.

Basic noise outsourcing can be refined in the presence of conditional independence. Let $S : \mathcal{X} \to \mathcal{S}$ be a statistic such that Y is conditionally independent from X, given

[2]Noise outsourcing is sometimes written as $Y = f(\eta, X) = F(X)$, for some *random* function F. The two are equivalent, but we write $f(\eta, X)$ to emphasize that the stochastic part of the relationship between X and Y can be outsourced to η.

$S(X)$: $Y \perp\!\!\!\perp_{S(X)} X$. Borrowing terminology from the graphical models literature, we say that S *d-separates* X and Y (Lauritzen, 1996). The following basic result, which we use extensively, says that if there is a statistic that d-separates X and Y, then it is possible to represent Y as a noise-outsourced function of S.

LEMMA 3. *Let X and Y be random variables with joint distribution $P_{X,Y}$. Let $\mathcal{S}$ be a standard Borel space and $S : \mathcal{X} \to \mathcal{S}$ a measurable map. Then $S(X)$ d-separates X and Y if and only if there is a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} (X, f(\eta, S(X))) \quad \textit{where} \quad \eta \sim \text{Unif}[0,1] \quad \textit{and} \quad \eta \perp\!\!\!\perp X \, . \tag{8}$$

In particular, conditionally on X, $Y \overset{\text{a.s.}}{=} f(\eta, S(X))$.

The proof is a straightforward application of a standard result from measure theoretic probability, given in Appendix A.1. Note that in general, f is measurable but need not be differentiable or otherwise have desirable properties, although for modeling purposes f can be limited to functions belonging to a tractable class (e.g., differentiable, parameterized by a neural network). Note also that the identity map $S(X) = X$ trivially d-separates X and Y, so that $Y \overset{\text{a.s.}}{=} f(\eta, X)$, which is standard noise outsourcing (e.g., Austin, 2015, Lem. 3.1).

3.2. *Maximal invariants, orbit laws, and functional representations of $\mathcal{G}$-symmetric conditional distributions.* For a group $\mathcal{G}$ acting on a set $\mathcal{X}$, the *orbit* of any $x \in \mathcal{X}$ is the set of elements in $\mathcal{X}$ that can be generated by applying the elements of $\mathcal{G}$. It is denoted $\mathcal{G} \cdot x = \{g \cdot x; g \in \mathcal{G}\}$. The *stabilizer*, or isotropy subgroup, of $x \in \mathcal{X}$ is the subgroup of $\mathcal{G}$ that leaves x unchanged: $\mathcal{G}_x = \{g \in \mathcal{G}; g \cdot x = x\}$. An *invariant statistic* $S : \mathcal{X} \to \mathcal{S}$ is a measurable map that satisfies $S(x) = S(g \cdot x)$ for all $g \in \mathcal{G}$ and $x \in \mathcal{X}$. A *maximal invariant statistic*, or maximal invariant, is an invariant statistic $M : \mathcal{X} \to \mathcal{S}$ such that $M(x_1) = M(x_2)$ implies $x_2 = g \cdot x_1$ for some $g \in \mathcal{G}$; equivalently, M takes a different constant value on each orbit.

The last property is useful when working with invariant distributions. By definition, an invariant distribution P_X is constant on any particular orbit. Consider the conditional distribution $P_X(X \mid M(X) = m)$. Conditioning on the maximal invariant taking a particular value is equivalent to conditioning on X being in a particular orbit; for invariant P_X, the conditional distribution is zero outside the orbit, and "uniform" on the orbit, modulo fixed points (see Footnote 4). Lemma 5 below makes that intuition precise; it asserts that any $\mathcal{G}$-invariant distribution P_X on $\mathcal{X}$ can be disintegrated into a distribution *over* orbits and a fixed conditional distribution *on* each orbit. Defining the appropriate distributions requires the following definition, taken from Eaton (1989).

DEFINITION 4. A *measurable cross-section* is a set $\mathcal{C} \subset \mathcal{X}$ with the following properties:

(i) $\mathcal{C}$ is measurable.
(ii) For each $x \in \mathcal{X}$, $\mathcal{C} \cap \mathcal{G} \cdot x$ consists of exactly one point, say $c(x)$.
(iii) The function $t : \mathcal{X} \to \mathcal{C}$ defined by $t(x) = c(x)$ is $\mathcal{B}_{\mathcal{X}}$-measurable when $\mathcal{C}$ has σ-algebra $\mathcal{B}_{\mathcal{C}} = \{B \cap \mathcal{C} | B \in \mathcal{B}_{\mathcal{X}}\}$.

One may think of a measurable cross-section as consisting of a single representative from each orbit of $\mathcal{X}$ under $\mathcal{G}$. For a compact group acting measurably on a Borel space, there always exists a (not necessarily unique) measurable cross-section. For non-compact groups or more general spaces, the existence of a measurable cross-section is not guaranteed, and more powerful technical tools are required (e.g., Andersson (1982); Schindler (2003); Kallenberg (2017)).

Orbit laws. Intuitively, given that X lies on a particular orbit, it should be possible to sample a realization of X by first sampling a random group element G, and then applying G to a representative element of the orbit. More precisely, let $\lambda_{\mathcal{G}}$ be the normalized Haar measure of $\mathcal{G}$, which is the unique left- and right-invariant measure on $\mathcal{G}$ such that $\lambda_{\mathcal{G}}(\mathcal{G}) = 1$ (Kallenberg, 2002, Ch. 2). Let $M : \mathcal{X} \to \mathcal{S}$ be a maximal invariant. For any $m \in \mathcal{S}$, let x_m be the element of $\mathcal{C}$ for which $M(x_m) = m$, if such an element exists. Note that the set of elements in $\mathcal{S}$ that do not correspond to an orbit of $\mathcal{X}$, $M_{\emptyset} := \{m; M^{-1}(m) = \emptyset\}$, has measure zero under P_X.

For any $B \in \mathcal{B}_{\mathcal{X}}$ and $m \in \mathcal{S}$, define the *orbit law* as[3]

$$\mathbb{U}^{\mathcal{G}}_m(B) = \int_{\mathcal{G}} \delta_{g \cdot x_m}(B) \lambda_{\mathcal{G}}(dg) = \lambda_{\mathcal{G}}(\{g; g \cdot x_m \in B\}) \;, \tag{9}$$

with $\mathbb{U}^{\mathcal{G}}_m(\bullet) = 0$ for any $m \in M_{\emptyset}$. Observe that, in agreement with the intuition above, a sample from the orbit law can be generated by sampling a random group element $G \sim \lambda_{\mathcal{G}}$, and applying it to x_m. The orbit law inherits the invariance of $\lambda_{\mathcal{G}}$ and acts like the uniform distribution on the elements of the orbit, up to fixed points.[4] For any function $\varphi : \mathcal{X} \to \mathbb{R}$, the expectation with respect to the orbit law is

$$\mathbb{U}^{\mathcal{G}}_m[\varphi] = \int_{\mathcal{X}} \varphi(x) \mathbb{U}^{\mathcal{G}}_m(dx) = \int_{\mathcal{G}} \varphi(g \cdot x_m) \lambda_{\mathcal{G}}(dg) \;. \tag{10}$$

The orbit law arises in the disintegration of any $\mathcal{G}$-invariant distribution P_X as the fixed distribution on each orbit.

[3] $\delta_x(A) = 1$ if $x \in A$ and is 0 otherwise, for any measurable set A.

[4] The orbit law only coincides with the uniform distribution on the orbit if the action of $\mathcal{G}$ on the orbit is *regular*, which corresponds to the action being transitive and *free* (or fixed-point free). Since by definition the action of $\mathcal{G}$ is transitive on each orbit, the orbit law is equivalent to the uniform distribution on the orbit if and only if $\mathcal{G}$ acts freely on each orbit; if the orbit has any fixed points (i.e., $g \cdot x = x$ for some x in the orbit and $g \neq e$), then the fixed points will have higher probability mass.

LEMMA 5. *Let $\mathcal{X}$ and $\mathcal{S}$ be Borel spaces, $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$, and $M : \mathcal{X} \to \mathcal{S}$ a maximal invariant on $\mathcal{X}$ under $\mathcal{G}$. If X is a random element of $\mathcal{X}$, then its distribution P_X is $\mathcal{G}$-invariant if and only if*

$$P_X(X \in \bullet \mid M(X) = m) = \mathbb{U}^{\mathcal{G}}_m(\bullet) = q(\bullet, m) \,, \tag{11}$$

for some Markov kernel $q : \mathcal{B}_{\mathcal{X}} \times \mathcal{S} \to \mathbb{R}_+$. If P_X is $\mathcal{G}$-invariant and Y is any other random variable, then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if $Y \perp\!\!\!\perp_{M(X)} X$.

The proof is given in Appendix A.2. Note that (11) is equivalent to the decomposition

$$P_X(X \in \bullet) = \int_{\mathcal{S}} \mathbb{U}^{\mathcal{G}}_m(\bullet)\nu(dm) = \int_{\mathcal{C}} \mathbb{U}^{\mathcal{G}}_{M(x)}(\bullet)\mu(dx) \,,$$

for some maximal invariant $M : \mathcal{X} \to \mathcal{S}$ and probability measures ν on $\mathcal{S}$ and μ on the measurable cross-section $\mathcal{C}$. This type of statement is more commonly encountered (e.g., Eaton, 1989, Ch. 4-5), but the form of (11) emphasizes the role of the orbit law as the conditional distribution of X given $M(X)$, which is central to the development of ideas in Section 4.

Invariant conditional distributions. With the conditional independence relationship $Y \perp\!\!\!\perp_{M(X)} X$ established in Lemma 5, Lemma 3 can be used to obtain a generative functional representation of invariant conditional distributions.

THEOREM 6. *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random elements of Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, and $\mathcal{G}$ a compact group acting measurably on $\mathcal{X}$. Assume that P_X is $\mathcal{G}$-invariant, and pick a maximal invariant $M : \mathcal{X} \to \mathcal{S}$, with $\mathcal{S}$ another Borel space. Then $P_{Y|X}$ is $\mathcal{G}$-invariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X \,. \tag{12}$$

PROOF. The proof follows directly from Lemmas 3 and 5. □

A maximal invariant always exists for sufficiently nice $\mathcal{G}$ and $\mathcal{X}$ (Hall, Wijsman and Ghosh, 1965); for example, define $M : \mathcal{X} \to \mathbb{R}$ to be a function that takes a unique value on each orbit. Therefore, the orbit law can always be defined in such cases. The assumptions made in the present work, that $\mathcal{G}$ acts measurably on the Borel space $\mathcal{X}$, allow for the existence of maximal invariants, and in many settings of interest a maximal invariant is straightforward to construct. For example, Sections 6 and 7 rely on constructing maximal invariants for applications of the results of this section to specific exchangeable structures.

Versions of Theorem 6 may hold for non-compact groups and more general topological spaces, but require considerably more technical details. At a high level,

$\mathcal{G}$-invariant measures on $\mathcal{X}$ may be decomposed into product measures on suitable spaces $\mathcal{S} \times \mathcal{Z}$ under fairly general conditions, though extra care is needed in order to ensure that statements of conditional independence such as (11) are well-defined. See, for example, Andersson (1982); Eaton (1989); Wijsman (1990); Schindler (2003); and especially Kallenberg (2017).

Maximal equivariants. The orbit law, based on *any* maximal invariant, was used to establish a generative functional representation of $\mathcal{G}$-invariant conditional distributions. If a particular type of maximal invariant exists and is measurable, then it can be used to establish a generative functional representation of equivariant conditional distributions. Let $\tau : \mathcal{X} \to \mathcal{G}$ be a function mapping elements of $\mathcal{X}$ to elements of $\mathcal{G}$, such that it is equivariant:

$$\tau(g \cdot x) = g \cdot \tau(x) \ , \quad \text{for all } g \in \mathcal{G}, x \in \mathcal{X} \ . \tag{13}$$

For ease of notation, let $\tau_x = \tau(x)$, and denote by τ_x^{-1} the left-inverse of τ_x in $\mathcal{G}$, such that $\tau_x^{-1}\tau_x = e$. We call such a function a *maximal equivariant*, after Wijsman (1990, Remark 12.2), because it can be used to construct a maximal invariant that is a representative element from each orbit in $\mathcal{X}$ and, more importantly, to construct a $\mathcal{G}$-equivariant function from an arbitrary function. The proof of the following lemma is given in Appendix A.3.

LEMMA 7. *For a group $\mathcal{G}$ acting measurably on Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, a maximal equivariant $\tau : \mathcal{X} \to \mathcal{G}$, as defined in (13), has the following properties:*

(i) The function $M_\tau : \mathcal{X} \to \mathcal{X}$ defined by $M_\tau(x) = \tau_x^{-1} \cdot x$ is a maximal invariant.
(ii) For any mapping $b : \mathcal{X} \to \mathcal{Y}$, the function

$$f(x) = \tau_x \cdot b(\tau_x^{-1} \cdot x), \quad x \in \mathcal{X} \tag{14}$$

satisfies $\mathcal{G}_x \subseteq \mathcal{G}_{f(x)} \Rightarrow f(g \cdot x) = g \cdot f(x)$.

Equivariant conditional distributions. The properties from Lemma 7 are used to establish the equivariant counterpart of Theorem 6. In essence, for $\mathcal{G}$-invariant P_X, Y is conditionally $\mathcal{G}$-equivariant given X if and only if there exists a function such that

$$g \cdot Y \overset{\text{a.s.}}{=} f(\eta, g \cdot X) \ , \quad \text{for each } g \in \mathcal{G}, \text{ with } \eta \sim \text{Unif}[0,1] \text{ and } \eta \perp\!\!\!\perp X \ . \tag{15}$$

Observe that this is equivalent to f being $\mathcal{G}$-equivariant (3) in the second argument: let e denote the identity element of $\mathcal{G}$; then $Y = e \cdot Y = f(\eta, e \cdot X) = f(\eta, X)$ and therefore

$$f(\eta, g \cdot X) = g \cdot Y = g \cdot f(\eta, X) \ , \quad \text{for each } g \in \mathcal{G} \ .$$

Equation (15) holds for compact groups acting on a space $\mathcal{X}$ such that τ as in (13) exists and is measurable, with random elements X, Y satisfying the regularity condition $\mathcal{G}_X \subseteq \mathcal{G}_Y$ almost surely.

THEOREM 8. *Let $\mathcal{G}$ be a compact group acting measurably on Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, such that there exists a measurable maximal equivariant $\tau : \mathcal{X} \to \mathcal{G}$ satisfying (13), and consider random elements $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ such that $\mathcal{G}_X \subseteq \mathcal{G}_Y$ almost surely. Suppose P_X is $\mathcal{G}$-invariant. Then $P_{Y|X}$ is $\mathcal{G}$-equivariant if and only if there exists a measurable function $f : [0,1] \times \mathcal{X} \to \mathcal{Y}$ such that (15) is true.*

Theorem 8 is an adaptation of Kallenberg (2005, Lem. 7.11). The proof is fairly technical and only sketched here; it is given in full in Appendix A.4. It relies on establishing the conditional independence relationship $\tau_X^{-1} \cdot Y \perp\!\!\!\perp_{M_\tau(X)} X$. Once that is established, noise outsourcing (Lemma 3) implies that $\tau_X^{-1} \cdot Y = b(\eta, M_\tau(X))$ for some $b : [0,1] \times \mathcal{X} \to \mathcal{Y}$; Lemma 7 then implies that

$$Y = f(\eta, X) := \tau_X \cdot b(\eta, M_\tau(X)) = \tau_X \cdot b(\eta, \tau_X^{-1} \cdot X)$$

is equivariant in the second argument.

Observe that if the action of $\mathcal{G}$ on $\mathcal{Y}$ is trivial, $\tau_X^{-1} \cdot Y \perp\!\!\!\perp_{M_\tau(X)} X$ reduces to $Y \perp\!\!\!\perp_{M_\tau(X)} X$, which is precisely the relationship needed to establish the invariant representation in Theorem 6. However, in the invariant case, *any* maximal invariant will d-separate X and Y, and therefore Theorem 6 is more general than simply applying Theorem 8 to the case with trivial group action on $\mathcal{Y}$. Intuitively, the additional assumptions in Theorem 8 account for the non-trivial action of $\mathcal{G}$ on $\mathcal{Y}$, which requires setting up a fixed frame of reference through τ and the orbit representatives $M_\tau(X)$.

We note that Theorem 8 may hold for non-compact groups, but the proof for compact groups makes use of the normalized Haar measure; extending to non-compact groups requires additional technical overhead. Moreover, the regularity condition $\mathcal{G}_X \subseteq \mathcal{G}_Y$ is a sufficient condition, but there are situations when it is not necessary; permutation-invariance (Section 6) is one such case, as Theorem 14 demonstrates.

3.3. *Computing maximal invariants and maximal equivariants.* In theory, a maximal invariant can be computed by specifying a representative element for each orbit. In practice, this can always be done when $\mathcal{G}$ or $\mathcal{X}$ is discrete because the relevant elements can be enumerated and reduced to permutation operations. Several systems for computational discrete mathematics, such as Mathematica (Wolfram Research, Inc., 2018) and GAP (The GAP Group, 2018), have built-in functions to do so. For continuous groups it may not be clear how to compute a maximal invariant. Furthermore, maximal invariants are not unique, and some may be better suited to a particular application than others. As such, they are best handled on a case-by-case basis, depending on the problem at hand. Some examples from the classical statistics literature are reviewed in Lehmann and Romano (2005, Ch. 6) and Eaton (1989, Ch. 2); Kallenberg (2017, Ch. 7) presents a generic method based on so-called projection maps. Sections 6 and 7 apply the theory of this section to exchangeable structures by explicitly constructing maximal invariants.

Likewise, a maximal equivariant $\tau : \mathcal{X} \to \mathcal{G}$, as used in Theorem 8, always exists for discrete groups and $\mathcal{X}$ a Borel space, a proof of which is given in Kallenberg (2005, Lem. 7.10). If $\mathcal{G}$ is even compact and acts measurably on Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, then a maximal equivariant exists; see Schindler (2003, Remark 2.46). In more general settings, τ with the desired properties may not be well-defined or it may not be measurable, in which case a $\mathcal{G}$-invariant probability kernel (called an *inversion kernel*) may by constructed from a maximal invariant to obtain similar representations; see Kallenberg (2017, Ch. 7).

4. Statistical sufficiency and adequacy. Theorems 6 and 8 show that the two approaches in Section 2 are equivalent: symmetry can be incorporated into an explicit probabilistic model by considering invariant and equivariant distributions, or it can be incorporated into an implicit model by considering invariant and equivariant functions. The deterministic functional models in Section 2.1 are special cases of the latter. Those results suggest how to specify models as families of functions that satisfy (5)-(6) for a particular group. To make this precise, and to put these ideas in wider statistical context, we appeal to the foundational statistical concepts of sufficiency and adequacy; when combined with the results of the previous section, they form a general program for identifying function classes and neural network architectures corresponding to symmetry conditions.

Sufficiency and adequacy are based on the idea that a statistic may contain all information that is needed for an inferential procedure; for example, to completely describe the distribution of a sample, parameter inference, or for prediction. The ideas go hand-in-hand with notions of symmetry: while invariance describes information that is irrelevant (for example, permutation-invariance means that the ordering among a set of data points does not matter), sufficiency and adequacy describe the information that is relevant (the values of the data points).

Whereas the results in Section 3 concerned probability distributions, sufficiency and adequacy are defined with respect to a probability *model*: a family of distributions indexed by some parameter $\theta \in \Omega$. Throughout, we consider a model for the joint distribution over X and Y, $\mathcal{P}_{X,Y} = \{P_{X,Y;\theta} : \theta \in \Omega\}$, from which there is an induced marginal model $\mathcal{P}_X = \{P_{X;\theta} : \theta \in \Omega\}$ and a conditional model $\mathcal{P}_{Y|X} = \{P_{Y|X;\theta} : \theta \in \Omega\}$. For convenience, we suppress the notational dependence on θ.

4.1. *Sufficiency.* Statistical sufficiency originates in the work of Fisher (1922); it formalizes the notion that a statistic might be used in place of the data for any statistical procedure. It has been generalized in a variety of different directions, including predictive sufficiency (Bahadur, 1954; Lauritzen, 1974a; Fortini, Ladelli and Regazzini, 2000) and adequacy (Skibinsky, 1967). There are a number of ways to formalize sufficiency, which are equivalent under certain regularity conditions; see Schervish (1995). The definition that is most convenient here is due to Halmos and Savage (1949): there is a *single* Markov kernel that gives the *same* conditional distribution of X conditioned on $S(X) = s$ for *every* distribution $P_X \in \mathcal{P}_X$.

DEFINITION 9. Let $\mathcal{S}$ be a Borel space and $S : \mathcal{X} \to \mathcal{S}$ a measurable map. S is a *sufficient statistic for* $\mathcal{P}_X$ if there is a Markov kernel $q : \mathcal{B}_{\mathcal{X}} \times \mathcal{S} \to \mathbb{R}_+$ such that for all $P_X \in \mathcal{P}_X$ and $s \in \mathcal{S}$, we have $P_X(\,\bullet \mid S(X) = s) = q(\,\bullet\,, s)$.

A canonical example is that of a sequence of n i.i.d. coin tosses. If the probability model is the family of Bernoulli distributions with probability of heads equal to p, then the number of heads N_h is sufficient: conditioned on $N_h = n_h$, the distribution of the data is uniform on all sequences with n_h heads, and the number of heads is sufficient in estimating p. For example, the maximum likelihood estimator is N_h/n.

Section 6 pertains to the more nuanced example of finite exchangeable sequences $\mathbf{X}_n \in \mathcal{X}^n$. A distribution P_X on $\mathcal{X}^n$ is *finitely exchangeable* if for all sets $A_1, \ldots, A_n \in \mathcal{B}_{\mathcal{X}}$,

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_{\pi(1)} \in A_1, \ldots, X_{\pi(n)} \in A_n)\ , \quad \text{for all } \pi \in \mathbb{S}_n\ . \tag{16}$$

Denote by $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n}$ the family of all exchangeable distributions on $\mathcal{X}^n$. The de Finetti conditional i.i.d. representation (1) may fail for a finitely exchangeable sequence (see Diaconis, 1977; Diaconis and Freedman, 1980a, for examples). However, the *empirical measure* plays a central role in both the finitely and infinitely exchangeable cases. The empirical measure of a sequence $\mathbf{X}_n \in \mathcal{X}^n$ is defined as

$$\mathbb{M}_{\mathbf{X}_n}(\,\bullet\,) = \sum_{i=1}^{n} \delta_{X_i}(\,\bullet\,)\ , \tag{17}$$

where δ_{X_i} denotes an atom of unit mass at X_i. The empirical measure discards the information about the order of the elements of $\mathbf{X}_n$, but retains all other information. A standard fact (see Section 6) is that a distribution $P_{\mathbf{X}_n}$ on $\mathcal{X}^n$ is exchangeable if and only if the conditional distribution $P_{\mathbf{X}_n}(\mathbf{X}_n \mid \mathbb{M}_{\mathbf{X}_n} = m)$ is the uniform distribution on all sequences that can be obtained by applying a permutation to $\mathbf{X}_n$. This is true for all distributions in $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n}$; therefore $\mathbb{M}_{\mathbf{X}_n}$ is a sufficient statistic for $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n}$ according to Definition 9, and we may conclude for any probability model $\mathcal{P}_{\mathbf{X}_n}$:

$\mathbb{S}_n$*-invariance of all* $P_{\mathbf{X}_n} \in \mathcal{P}_{\mathbf{X}_n}$ *is equivalent to the sufficiency of* $\mathbb{M}_{\mathbf{X}_n}$.

In this case invariance and sufficiency clearly are two sides of the same coin: a sufficient statistic captures all information that is relevant to a model for $\mathbf{X}_n$; invariance discards the irrelevant information. Section 6 explores this in further detail.

We note that in this work, there is a clear correspondence between group invariance and a sufficient statistic (see Section 4.3), as in the example of exchangeable sequences. In other situations, there is a sufficient statistic but the set of symmetry transformations may not correspond to a group. See Freedman (1962, 1963); Diaconis and Freedman (1987) for examples.

4.2. *Adequacy.* The counterpart of sufficiency for modeling the conditional distribution of Y given X is *adequacy* (Skibinsky, 1967; Speed, 1978).[5] The following definition adapts one given by Lauritzen (1974b), which is easier to interpret than the measure theoretic definition introduced by Skibinsky (1967).

DEFINITION 10. Let $\mathcal{S}$ be a Borel space, and let $S : \mathcal{X} \to \mathcal{S}$ be a measurable map. Then S is an *adequate statistic of X for Y with respect to $\mathcal{P}_{X,Y}$* if

(i) S is sufficient for $\mathcal{P}_X$; and
(ii) for all $x \in \mathcal{X}$ and $P_{X,Y} \in \mathcal{P}_{X,Y}$,

$$P_{X,Y}(Y \in \bullet \mid X = x) = P_{X,Y}(Y \in \bullet \mid S = S(x))\,. \tag{18}$$

Equation (18) amounts to the conditional independence $Y \perp\!\!\!\perp_{S(X)} X$ of Y and X given $S(X)$. To see this, note that because $S(X)$ is a measurable function of X,

$$P_{X,Y}(Y \in \bullet \mid X = x) = P_{X,Y}(Y \in \bullet \mid X = x, S = S(x)) = P_{X,Y}(Y \in \bullet \mid S = S(x))\,,$$

which is equivalent to $Y \perp\!\!\!\perp_{S(X)} X$. Therefore, adequacy is equivalent to sufficiency for $\mathcal{P}_X$ and d-separation of X and Y, for all distributions in $\mathcal{P}_{X,Y}$.

To make the connections between ideas clear, we state the following corollary of Lemma 3.

COROLLARY 11. *Let $S : \mathcal{X} \to \mathcal{S}$ be a sufficient statistic for the model $\mathcal{P}_X$. Then S is an adequate statistic for $\mathcal{P}_{X,Y}$ if and only if (8) holds.*

4.3. *Maximal invariants as adequate statistics.* The orbit-resolving property of maximal invariants is statistically useful. In particular, denote by $\mathcal{P}^{\mathcal{G}}_{X,Y}$ the family of probability measures on $\mathcal{X} \times \mathcal{Y}$ that satisfy $(g \cdot X, Y) \overset{\text{d}}{=} (X, Y)$, for all $g \in \mathcal{G}$. The induced marginal model, the family of all $\mathcal{G}$-invariant probability measures on $\mathcal{X}$, is $\mathcal{P}^{\mathcal{G}}_X$. Lemma 5 holds for any $P_{X,Y} \in \mathcal{P}^{\mathcal{G}}_{X,Y}$, and in particular any maximal invariant is an adequate statistic for $\mathcal{P}^{\mathcal{G}}_{X,Y}$.

THEOREM 12. *Let $\mathcal{G}$ be a compact group acting measurably on standard Borel spaces $\mathcal{X}$ and $\mathcal{Y}$, and let $\mathcal{S}$ be another Borel space. Then any maximal invariant $M : \mathcal{X} \to \mathcal{S}$ on $\mathcal{X}$ under $\mathcal{G}$ is a sufficient statistic for any $\mathcal{P}_X \subseteq \mathcal{P}^{\mathcal{G}}_X$. Moreover, M is an adequate statistic of X for Y with respect to any model $\mathcal{P}_{X,Y} \subseteq \mathcal{P}^{\mathcal{G}}_{X,Y}$. In particular, to each $P_{X,Y} \in \mathcal{P}^{\mathcal{G}}_{X,Y}$ corresponds a measurable function $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that*

$$(X, Y) \overset{\text{a.s.}}{=} \big(X, f(\eta, M(X))\big) \quad \textit{with } \eta \sim \mathrm{Unif}[0,1] \textit{ and } \eta \perp\!\!\!\perp X\,. \tag{19}$$

[5] When Y is the prediction of X_{n+1} from $\mathbf{X}_n$, adequacy is also known as *predictive sufficiency* (Fortini, Ladelli and Regazzini, 2000) and, under certain conditions, it is equivalent to sufficiency and transitivity (Bahadur, 1954) or to total sufficiency (Lauritzen, 1974a). See Lauritzen (1974b) for a precise description of how these concepts are related.

PROOF. Let $M : \mathcal{X} \to \mathcal{S}$ be any maximal invariant on $\mathcal{X}$ under $\mathcal{G}$. From Lemma 5, the conditional distribution of X given $M(X) = m$ is equal to the orbit law $\mathbb{U}^{\mathcal{G}}_m$. $\mathcal{S}$ is assumed to be Borel, so there is a Markov kernel (e.g., Kallenberg (2002, Thm. 6.3)) $q_M : \mathcal{B}_{\mathcal{X}} \times \mathcal{S} \to \mathbb{R}_+$ such that $q_M(\bullet, m) = \mathbb{U}^{\mathcal{G}}_m(\bullet)$, and therefore M is a sufficient statistic for $\mathcal{P}^{\mathcal{G}}_X$. Moreover, also by Lemma 5, M d-separates X and Y under any $P_{X,Y} \in \mathcal{P}^{\mathcal{G}}_{X,Y}$, and therefore M is also an adequate statistic. Lemma 3 implies the identity (19). □

To illustrate, consider again the example of an exchangeable $\mathcal{X}$-valued sequence.

EXAMPLE 3 (Maximal invariant of an exchangeable sequence). Let $\mathbf{X}_n$ be an exchangeable $\mathcal{X}$-valued sequence. $\mathbb{S}_n$ acts on any point $\mathbf{x}_n \in \mathcal{X}^n$ by permuting its indices, which defines an action on $\mathcal{X}^n$; the orbit of $\mathbf{x}_n$ is the set of sequences that can be obtained from $\mathbf{x}_n$ by applying a permutation. That the empirical measure is a sufficient statistic for $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n}$ is not a coincidence; it is easy to see that the empirical measure is a maximal invariant. If $\mathcal{X} = \mathbb{R}$ (or any other set with a total order), then so too is the vector of order statistics, $\mathbf{x}^{\uparrow}_n = (x_{(1)}, \ldots, x_{(n)})$, with $x_{(1)} \leq \cdots \leq x_{(n)}$. A maximal equivariant may be obtained by defining $\tau^{-1}_{\mathbf{x}_n}$ as any permutation (not necessarily unique) that satisfies $\pi \cdot \mathbf{x}_n = \mathbf{x}^{\uparrow}_n$.

5. A program for obtaining symmetric functional representations.

Section 4 suggest that for a model consisting of $\mathcal{G}$-invariant distributions, if an adequate statistic can be found then all distributions in the conditional model, $P_{Y|X} \in \mathcal{P}_{Y|X}$ have a noise-outsourced functional representation in terms of the adequate statistic, as in (8). Furthermore, Theorem 12 says that any maximal invariant is an adequate statistic. Therefore, a program for functional model specification through distributional symmetry is as follows:

(i) Specify a group $\mathcal{G}$ under which the input-output relationship is assumed to be distributionally invariant:

$$(g \cdot X, Y) \stackrel{\mathrm{d}}{=} (X, Y) \, , \quad \text{for all } g \in \mathcal{G} \, . \tag{20}$$

(ii) Determine a statistic $M : \mathcal{X} \to \mathcal{S}$ that is a maximal invariant under $\mathcal{G}$.

(iii) Specify the model with a function class $\mathcal{F}_M = \{f | f : [0,1] \times \mathcal{S} \to \mathcal{Y}\}$. That is, model the input-output relationship by a class of noise-outsourced functions: $Y = f(\eta, M(X))$.

Embellishments on this basic program can be used for increased model flexibility. For example, Proposition 2 implies that multiple noise-outsourced invariant functions can be composed to obtain another distributionally invariant random variable. If a maximal equivariant can be found, then the same proposition implies that the model's function class can be enlarged to include compositions of noise-outsourced equivariant functions, composed with a final noisy invariant function.

Exchangeable random structures.. The remainder of the paper applies the program outlined above to various exchangeable random structures, and relates the resulting representations to some recent (Zaheer et al., 2017; Gilmer et al., 2017; Hartford et al., 2018) and less recent (Shawe-Taylor, 1989) work in the neural networks literature. Exchangeability is distributional invariance under the action of $\mathbb{S}_n$ (or other groups defined by composing $\mathbb{S}_n$ in different ways for other data structures). Example 3 showed that the empirical measure is a maximal invariant of $\mathbb{S}_n$ acting on $\mathcal{X}^n$; suitably defined generalizations of the empirical measure are maximal invariants for other exchangeable structures. With these maximal invariants, obtaining a functional representation of an invariant conditional distribution is straightforward.

With an additional conditional independence assumption that is satisfied by most neural network architectures, Theorem 8 can be refined to obtain a detailed functional representation of the relationship between $\mathbb{S}_n$-equivariant random variables (Theorem 14). That refinement relies on the particular subgroup structure of $\mathbb{S}_n$, and it raises the question, not pursued here, of whether, and under what conditions, other groups may yield similar refinements.

6. Learning from finitely exchangeable sequences. In this section, the program described in Section 5 is fully developed for the case where the conditional distribution of Y given X has invariance or equivariance properties with respect to $\mathbb{S}_n$. Deterministic examples have appeared in the neural networks literature (Shawe-Taylor, 1989; Zaheer et al., 2017; Ravanbakhsh, Schneider and Póczos, 2017), and the theory developed in this section establishes the necessary and sufficient functional forms for permutation invariance and equivariance, in both the stochastic and deterministic cases.

Throughout this section, the input $X = \mathbf{X}_n$ is a sequence of length n, and Y is an output variable, whose conditional distribution given $\mathbf{X}_n$ is to be modeled.[6] Recall from Section 2.2 that $P_{Y|\mathbf{X}_n}$ is $\mathbb{S}_n$-invariant if $Y|\mathbf{X}_n \overset{\text{d}}{=} Y|\pi \cdot \mathbf{X}_n$ for each permutation $\pi \in \mathbb{S}_n$ of the input sequence. Alternatively, if $Y = \mathbf{Y}_n$ is also a sequence of length n,[7] then we say that $\mathbf{Y}_n$ given $\mathbf{X}_n$ is $\mathbb{S}_n$-equivariant if $\mathbf{Y}_n|\mathbf{X}_n \overset{\text{d}}{=} \pi\cdot\mathbf{Y}_n|\pi\cdot\mathbf{X}_n$. In both cases these symmetry properties stem from the assumption that the ordering in the input sequence $\mathbf{X}_n$ does not matter; that is, the distribution of $\mathbf{X}_n$ is finitely exchangeable: $\mathbf{X}_n \overset{\text{d}}{=} \pi\cdot\mathbf{X}_n$ for each $\pi \in \mathbb{S}_n$. Recall that $P_{\mathbf{X}_n}$ and $P_{Y|\mathbf{X}_n}$ denote the marginal and conditional distributions respectively, $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n}$ is the family of distributions on $\mathcal{X}^n$ that are $\mathbb{S}_n$-invariant, and $\mathcal{P}^{\mathbb{S}_n}_{Y|\mathbf{X}_n}$ is the family of conditional distributions on $\mathcal{Y}$ given $\mathbf{X}_n$ that are $\mathbb{S}_n$-invariant.

The crux of the matter is establishing the central role of the empirical measure $\mathbb{M}_{\mathbf{X}_n}$. Exchangeable sequences have been studied in great detail, and the suffi-

[6]Note that $\mathbf{X}_n$ may represent a set (i.e., there are no repeated values) or a multi-set (there may be repeated values). It depends entirely on whether $P_{\mathbf{X}_n}$: if $P_{\mathbf{X}_n}$ places all of its probability mass on sequences $x_n \in \mathcal{X}^n$ that do not have repeated values, then $\mathbf{X}_n$ represents a set almost surely. Otherwise, $\mathbf{X}_n$ represents a multi-set. The results of this section hold in either case.

[7]In general, $\mathbf{Y}$ need not be of length n, but the results are much simpler when it is; see Section 6.4.

ciency of the empirical measure for $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n}$ is well-known (e.g., Diaconis and Freedman (1980a); Kallenberg (2005, Prop. 1.8)). It is also straightforward to show the adequacy of the empirical measure for $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n,Y}$ using methods that are not explicitly group theoretic. Alternatively, it is enough to show that the empirical measure is a maximal invariant of $\mathcal{X}^n$ under $\mathbb{S}_n$ and then apply Theorem 12. In either case, the results of the previous sections imply a noise-outsourced functional representation of $\mathbb{S}_n$-invariant conditional distributions (Section 6.1). The previous sections also imply a representation for $\mathbb{S}_n$-equivariant conditional distributions, but under an additional conditional independence assumption a more detailed representation can be obtained due to the structure of $\mathbb{S}_n$ (Section 6.2).

6.1. *$\mathbb{S}_n$-invariant conditional distributions.* The orbit law from (9) has a special interpretation when $\mathcal{G} = \mathbb{S}_n$; it is also known as the *urn law* $\mathbb{U}^{\mathbb{S}_n}_m$, defined as a probability measure on $\mathcal{X}^n$ corresponding to permuting the elements of $\mathbb{M}_{\mathbf{X}_n}$ according to a uniformly sampled random permutation:

$$\mathbb{U}^{\mathbb{S}_n}_{\mathbb{M}_{\mathbf{X}_n}}(\,\bullet\,) = \frac{1}{n!}\sum_{\pi\in\mathbb{S}_n} \delta_{\pi\cdot\mathbf{X}_n}(\,\bullet\,)\ . \tag{21}$$

The urn law is so called because it computes the probability of generating any sequence that may be obtained by sampling without replacement from the elements of $\mathbb{M}_{\mathbf{X}_n}$.[8] The summand does not depend on $\mathbf{X}_n$ beyond $\mathbb{M}_{\mathbf{X}_n}$ and only is written as in (21) for convenience; any sequence with $\mathbb{M}_{\mathbf{X}_n}$ as its empirical measure can be used on the right-hand side of the equation. The following special case of Theorem 12 establishes a functional representation for all $\mathbb{S}_n$-invariant conditional distributions.

THEOREM 13. *Suppose $\mathbf{X}_n \in \mathcal{X}^n$ for some $n \in \mathbb{N}$. Then $\mathbf{X}_n$ is exchangeable if and only if*

$$P_X(\mathbf{X}_n \in \bullet \mid \mathbb{M}_{\mathbf{X}_n} = m) = \mathbb{U}^{\mathbb{S}_n}_m(\,\bullet\,)\ . \tag{22}$$

If $\mathbf{X}_n$ is exchangeable and Y is any other random variable such that $Y|\mathbf{X}_n \overset{\mathrm{d}}{=} Y|\pi\cdot\mathbf{X}_n$ for each $\pi \in \mathbb{S}_n$, then $Y \perp\!\!\!\perp_{\mathbb{M}_{\mathbf{X}_n}} \mathbf{X}_n$, and therefore $\mathbb{M}_{\mathbf{X}_n}$ is sufficient for the family $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n}$, and adequate for the family $\mathcal{P}^{\mathbb{S}_n}_{\mathbf{X}_n,Y}$. In particular, Y is conditionally $\mathbb{S}_n$-invariant given $\mathbf{X}_n$ if and only if there is a measurable function $f : [0,1]\times\mathcal{M}(\mathcal{X}) \to \mathcal{Y}$ such that

$$(\mathbf{X}_n, Y) \overset{\text{a.s.}}{=} (\mathbf{X}_n, f(\eta, \mathbb{M}_{\mathbf{X}_n})) \quad \textit{where} \quad \eta \sim \mathrm{Unif}[0,1] \quad \textit{and} \quad \eta \perp\!\!\!\perp \mathbf{X}_n\ . \tag{23}$$

[8] The metaphor is that of an urn with n balls, each labeled by an element of $\mathbb{M}_{\mathbf{X}_n}$; a sequence is constructed by repeatedly picking a ball uniformly at random from the urn, without replacement. Variations of such a scheme can be considered, for example sampling with replacement, or replacing a sampled ball with two of the same label. See Mahmoud (2008) for an overview of the extensive probability literature studying such processes, including the classical Pólya urn and its generalizations (e.g., Blackwell and MacQueen, 1973), which play important roles in Bayesian nonparametrics.

PROOF. As mentioned above, the proof of (22) can be found in Diaconis and Freedman (1980a) or Kallenberg (2005, Proposition 1.8); the proofs by those authors rely on the urn law. However, with Lemma 5 and Theorem 12, we need only prove that $\mathbb{M}_{\mathbf{X}_n}$ is a maximal invariant of $\mathcal{X}^n$ under $\mathbb{S}_n$. Clearly, $\mathbb{M}_{\mathbf{X}_n}$ is $\mathbb{S}_n$-invariant. Now, let $\mathbf{X}'_n$ be another sequence such that $\mathbb{M}_{\mathbf{X}'_n} = \mathbb{M}_{\mathbf{X}_n}$. Then $\mathbf{X}'_n$ and $\mathbf{X}_n$ contain the same elements of $\mathcal{X}$, and therefore $\mathbf{X}'_n = \pi \cdot \mathbf{X}_n$ for some $\pi \in \mathbb{S}_n$, so $\mathbb{M}_{\mathbf{X}_n}$ is a maximal invariant. Finally, observe that the Haar measure for a discrete group is the counting measure, and therefore (22) is (11) for the special case $\mathcal{G} = \mathbb{S}_n$. □

Modeling $\mathbb{S}_n$-invariance with neural networks. Theorem 13 is a general characterization of $\mathbb{S}_n$-invariant conditional distributions. It says that all such conditional distributions must have a noise-outsourced functional representation given by $Y = f(\eta, \mathbb{M}_{\mathbf{X}_n})$. Recall that $\mathbb{M}_{\mathbf{X}_n} = \sum_{i=1}^n \delta_{X_i}$. An atom δ_X can be thought of as a measure-valued generalization of a one-hot encoding to arbitrary measurable spaces, so that $\mathbb{M}_{\mathbf{X}_n}$ is a sum-pooling of encodings of the inputs (which removes information about the ordering of $\mathbf{X}_n$), and the output Y is obtained by passing that, along with independent outsourced noise η, through a function f. In case the conditional distribution is deterministic, the outsourced noise is unnecessary, and so we simply have $Y = f(\mathbb{M}_{\mathbf{X}_n})$.

From a modeling perspective, one choice for (stochastic) neural network architectures that are $\mathbb{S}_n$-invariant is

$$Y = f(\eta, \sum_{i=1}^{n} \phi(X_i)) \,, \tag{24}$$

where f and ϕ are arbitrary neural network modules, with ϕ interpreted as an embedding function of input elements into a high-dimensional space (see first panel of Fig. 2). These embeddings are sum-pooled, and passed through a second neural network module f. This architecture can be made to approximate arbitrarily well any $\mathbb{S}_n$-invariant conditional distribution (c.f., Hornik, Stinchcombe and White, 1989). Roughly, $\phi(X)$ can be made arbitrarily close to a one-hot encoding of X, which can in turn be made arbitrarily close to an atom δ_X by increasing its dimensionality, and similarly the neural module f can be made arbitrarily close to any desired function. Below, we revisit an earlier example and give some new ones.

EXAMPLE 4 (Deep Sets: $\mathbb{S}_n$-invariant functions of sequences, revisited). The architecture derived above is exactly the one described in Example 1. Theorem 13 generalizes the result in Zaheer et al. (2017), from deterministic functions to conditional distributions. The proof technique is also significantly simpler and sheds light on the core concepts underlying the functional representations of permutation invariance.[9]

[9]The result in Zaheer et al. (2017) holds for sets of arbitrary size when $\mathcal{X}$ is countable and for

In general, a function of $\mathbb{M}_{\mathbf{X}_n}$ is a function of $\mathbf{X}_n$ that discards the order of its elements. That is, functions of $\mathbb{M}_{\mathbf{X}_n}$ are permutation-invariant functions of $\mathbf{X}_n$. The sum-pooling in (24) gives rise to one such class of functions. Other permutation invariant pooling operations can be used; for example, product, maximum, minimum, log-sum-exp, mean, median, and percentiles have been used in various neural network architectures. Any such function can be written in the form $f(\eta, \mathbb{M}_{\mathbf{X}_n})$, by absorbing the pooling operation into f itself. Note that using these other pooling operators need not give universal approximators of $\mathbb{S}_n$-invariant conditional distributions or functions, since the resulting f is restricted to include the pooling operation.

Exchangeability plays a central role in a growing body of work in the deep learning literature, particularly when deep learning methods are combined with Bayesian ideas. Examples include Edwards and Storkey (2017); Garnelo et al. (2018); Korshunova et al. (2018), and the following.

Example 5 (Neural networks for exchangeable genetic data). In Chan et al. (2018), $X_i \in \{0,1\}^d$ is a binary d-dimensional vector indicating the presence or absence of d single nucleotide polymorphisms in individual i. The individuals are treated as being exchangeable, forming an exchangeable sequence $\mathbf{X}_n$ of $\{0,1\}^d$-valued random variables. Chan et al. (2018) analyze the data using a neural network where each vector X_i is embedded into $\mathbb{R}^{d'}$ using a convolutional network, the pooling operation is the element-wise mean of the top decile, and the final function is parameterized by a fully-connected network. They demonstrated empirically that encoding permutation invariance into the network architecture led to faster training and higher test set accuracy.

Example 6 (Pooling using Abelian groups or semigroups). A group $(\mathcal{G}, \oplus)$ is Abelian, or commutative, if its elements commute: $g \oplus h = h \cdot g$ for all $g, h \in \mathcal{G}$. Examples are $(\mathbb{R}_+, \times)$ and $(\mathbb{Z}, +)$. A semigroup has the same structure as a group, without the requirements for inverse elements and identity. Examples are $(\mathbb{R}_+, \vee)$ ($\vee$ denotes maximum: $x \vee y = \max\{x, y\}$) and $(\mathbb{R}_+, \wedge)$ ($\wedge$ denotes minimum). For a map $\phi : \mathcal{X} \to \mathcal{G}$, a $\mathbb{S}_n$-invariant conditional distribution of Y given a sequence $\mathbf{X}_n$ can be constructed with $f : [0,1] \times \mathcal{G} \to \mathcal{Y}$ as

$$Y = f(\eta, \phi(X_1) \oplus \cdots \oplus \phi(X_n)) \, , \tag{25}$$

Example 7 (Pooling using U-statistics). Given a permutation invariant function of $k \leq n$ elements, $\phi_k : \mathcal{X}^k \to \mathcal{S}$, a permutation invariant conditional distribu-

fixed size when $\mathcal{X}$ is uncountable. We note that the same is true for Theorem 13, for measure-theoretic reasons: in the countable $\mathcal{X}$ case, the power sets of $\mathbb{N}$ form a valid Borel σ-algebra; for uncountable $\mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}$, there may be non-Borel sets and therefore the power sets do not form a Borel σ-algebra on which to define a probability distribution using standard techniques.

tion can be constructed with $f : [0,1] \times \mathcal{S} \to \mathcal{Y}$ as

$$f\left(\eta, \binom{n}{k}^{-1} \sum_{\{i_1,\dots,i_k\}\in[n]} \phi_k(X_{i_1},\dots,X_{i_k})\right) \tag{26}$$

where the pooling involves averaging over all k-element subsets of $[n]$. The average is a U-statistic (e.g., Cox and Hinkley, 1974), and examples include the sample mean ($k = 1$ and $\phi_k(x) = x$), the sample variance ($k = 2$ and $\phi_k(x,y) = \frac{1}{2}(x-y)^2$), and estimators of higher-order moments, mixed moments, and cumulants.

Murphy et al. (2019) develop a host of generalizations to the basic first-order pooling functions from Example 1 (Deep Sets: $\mathbb{S}_n$-invariant functions of sequences), many of them corresponding to k-order U-statistics, and develop tractable computational techniques that approximate the average over k-element subsets by random sampling.

The final example, from the machine learning literature, uses the sufficiency of the empirical measure to characterize the complexity of inference algorithms for exchangeable sequences.

Example 8 (Lifted inference). Niepert and Van den Broeck (2014) studied the tractability of exact inference procedures for exchangeable models, through so-called lifted inference. One of their main results shows that if $\mathbf{X}_n$ is a finitely exchangeable sequence of $\mathcal{X}^d$-valued random variables on a discrete domain (i.e., each element of $\mathbf{X}_n$ is a d-dimensional vector of discrete random variables) then there is a sufficient statistic S, and probabilistic inference (defined as computing marginal and conditional distributions) based on S has computational complexity that is polynomial in $d \times n$. In the simplest case, where $\mathcal{X} = \{0,1\}$, S is constructed as follows: encode all possible d-length binary vectors with unique bit strings $b_k \in \{0,1\}^d$, $k \in [2^d]$, and let $S(\mathbf{X}_n) = (c_1,\dots,c_{2^d})$ where $c_k = \sum_{i=1}^n \delta_{X_i}(b_k)$. Although not called the empirical measure by the authors, S is precisely that.

6.2. *$\mathbb{S}_n$-equivariant conditional distributions.* Let $\mathbf{X}_n$ be an input sequence of length n, and $\mathbf{Y}_n$ an output sequence. Theorem 8 shows that if the conditional distribution of $\mathbf{Y}_n$ given $\mathbf{X}_n$ is $\mathbb{S}_n$-equivariant, then $\mathbf{Y}_n$ can be expressed in terms of a noisy $\mathbb{S}_n$-equivariant function of $\mathbf{X}_n$. If the elements of $\mathbf{Y}_n$ are assumed to be conditionally independent given $\mathbf{X}_n$, then by using properties of the finite symmetric group, we obtain a more detailed representation of $\mathbf{Y}_n$ conditioned on $\mathbf{X}_n$. (The implications of the conditional independence assumption are discussed below.)

The resulting theorem is a one-dimensional special case of a more general representation theorem for exchangeable d-dimensional arrays (Theorem 22 in Appendix B). The following simplified proof for sequences shows how the necessary conditional independence relationships are established and provides a template for proving the more general result. The $d = 2$ case, which corresponds to graphs and networks, is taken up in Section 7.

THEOREM 14. *Let $\mathbf{X}_n \in \mathcal{X}^n$ be an exchangeable sequence and $\mathbf{Y}_n \in \mathcal{Y}^n$ another random sequence, and assume that $Y_i \perp\!\!\!\perp_{\mathbf{X}_n} (\mathbf{Y}_n \setminus Y_i)$. Then $P_{\mathbf{Y}_n|\mathbf{X}_n}$ is $\mathbb{S}_n$-equivariant given $\mathbf{X}_n$ if and only if there is a measurable function $f : [0,1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \to \mathcal{Y}$ such that*

(27)
$$\big(\mathbf{X}_n, \mathbf{Y}_n\big) \overset{\text{a.s.}}{=} \Big(\mathbf{X}_n, \big(f(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_n})\big)_{i\in[n]}\Big) \quad \textit{where} \quad \eta_i \overset{iid}{\sim} \mathrm{Unif}[0,1] \textit{ and } \eta_i \perp\!\!\!\perp X \ .$$

PROOF. For the forward direction, suppose $\mathbf{Y}_n$ is conditionally $\mathbb{S}_n$-equivariant given $\mathbf{X}_n$. For a fixed $i \in [n]$, let $\mathbf{X}_{n\setminus i} := \mathbf{X}_n \setminus X_i$ be $\mathbf{X}_n$ with its ith element removed, and likewise for $\mathbf{Y}_{n\setminus i}$. The proof in this direction requires that we establish the conditional independence relationship

$$Y_i \perp\!\!\!\perp_{X_i, \mathbb{M}_{\mathbf{X}_n}} \mathbf{X}_n, \mathbf{Y}_{n\setminus i} \ , \tag{28}$$

and then apply Lemma 3 ().

To that end, let $\mathbb{S}_{n\setminus i}$ be the stabilizer of i, i.e., the subgroup of $\mathbb{S}_n$ that fixes element i. $\mathbb{S}_{n\setminus i}$ consists of permutations $\pi_{\setminus i} \in \mathbb{S}_n$ for which $\pi_{\setminus i}(i) = i$. The action of $\pi_{\setminus i}$ on $\mathbf{X}_n$ fixes X_i; likewise it fixes Y_i in $\mathbf{Y}_n$. By Proposition 1, $(\mathbf{X}_n, \mathbf{Y}_n) \overset{\text{d}}{=} (\pi_{\setminus i} \cdot \mathbf{X}_n, \pi_{\setminus i} \cdot \mathbf{Y}_n)$, so that, marginalizing out $\mathbf{Y}_{n\setminus i}$ yields $(\mathbf{X}_n, Y_i) \overset{\text{d}}{=} (\pi_{\setminus i} \cdot \mathbf{X}_n, Y_i)$ for each $\pi_{\setminus i} \in \mathbb{S}_{n\setminus i}$. Moreover, $\mathbb{S}_{n\setminus i}$ forms a subgroup and is homomorphic to $\mathbb{S}_{n-1}$, so that the previous distributional equality is equivalent to

$$\big(\mathbf{X}_{n\setminus i}, (X_i, Y_i)\big) \overset{\text{d}}{=} \big(\pi' \cdot \mathbf{X}_{n\setminus i}, (X_i, Y_i)\big) \quad \text{for each} \quad \pi' \in \mathbb{S}_{n-1} \ .$$

Theorem 13, with input $\mathbf{X}_{n\setminus i}$ and output (X_i, Y_i) then implies $(X_i, Y_i) \perp\!\!\!\perp_{\mathbb{M}_{\mathbf{X}_{n\setminus i}}} \mathbf{X}_{n\setminus i}$. Conditioning on X_i as well gives $Y_i \perp\!\!\!\perp_{(X_i, \mathbb{M}_{\mathbf{X}_{n\setminus i}})} \mathbf{X}_n$, marginally for each Y_i. With the assumption of mutual conditional independence among $\mathbf{Y}_n$ conditioned on $\mathbf{X}_n$, the marginal conditional independence also holds jointly, and by the chain rule for conditional independence (Kallenberg, 2002, Prop. 6.8),

$$Y_i \perp\!\!\!\perp_{(X_i, \mathbb{M}_{\mathbf{X}_{n\setminus i}})} (\mathbf{X}_n, \mathbf{Y}_{n\setminus i}) \ . \tag{29}$$

Because conditioning on $(X_i, \mathbb{M}_{\mathbf{X}_{n\setminus i}})$ is the same as conditioning on $(X_i, \mathbb{M}_{\mathbf{X}_n})$, (29) is equivalent to the key conditional independence relationship (28).

By Lemma 3, there exists a measurable $f_i : [0,1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \to \mathcal{X}$ such that

$$(\mathbf{X}_n, \mathbf{Y}_{n\setminus i}, Y_i) \overset{\text{a.s.}}{=} \big(\mathbf{X}_n, \mathbf{Y}_{n\setminus i}, f_i(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_n})\big) \ ,$$

for $\eta_i \sim \mathrm{Unif}[0,1]$ and $\eta_i \perp\!\!\!\perp (\mathbf{X}_n, \mathbf{Y}_{n\setminus i})$. This is true for each $i \in [n]$, and $\mathbb{S}_n$-equivariance implies that $(\mathbf{X}_n, \mathbf{Y}_{n\setminus i}, Y_i) \overset{\text{d}}{=} (\mathbf{X}_n, Y_{[n]\setminus j}, Y_j)$ for all $i, j \in [n]$. Thus it is possible to choose the same function $f_i = f$ for all i. This yields (27).

The reverse direction is easy to verify, since the noise variables are i.i.d., $\mathbf{X}_n$ is exchangeable, and $\mathbb{M}_{\mathbf{X}_n}$ is permutation-invariant.

□

The impact of the conditional independence assumption $Y_i \perp\!\!\!\perp_{\mathbf{X}_n} (\mathbf{Y}_n \setminus Y_i)$. In the deterministic case, the assumed conditional independence relationships among the outputs $\mathbf{Y}_n$ are trivially satisfied, so that (27) (without outsourced noise) is the most general form for a permutation-equivariant function. However, in the stochastic case, the assumed conditional independence significantly simplifies the structure of the conditional distribution and the corresponding functional representation. While the assumed conditional independence is key in the simplicity of the representation (27), it may limit the expressiveness: there are permutation-equivariant conditional distributions which do not satisfy the conditional independence assumption (examples are given below). On the other hand, without conditional independence between the elements of $\mathbf{Y}_n$, it is possible to show that $Y_i = f(\eta_i, X_i, \mathbb{M}_{(X_j,Y_j)_{j\in[n]\setminus i}})$. Such dependence between elements of $\mathbf{Y}_n$, although more expressive than (27), induces cycles in the computation graph, similar to Restricted Boltzmann Machines (Smolensky, 1987) and other Exponential Family Harmoniums (Welling, Rosen-zvi and Hinton, 2005). Furthermore, they may be limited in practice by the computational requirements of approximate inference algorithms. Striking a balance between flexibility and tractability via some simplifying assumption seems desirable.

Two examples illustrate the existence of permutation-equivariant conditional distributions which do not satisfy the conditional independence assumption made in Theorem 14, and suggest another assumption. Both examples have a similar structure: there exists some random variable, say W, such that the conditional independence $Y_i \perp\!\!\!\perp_{(\mathbf{X}_n, W)} (\mathbf{Y}_n \setminus Y_i)$ holds. Assuming the existence of such a W would lead to the representation $Y_i = f(\eta_i, W, X_i, \mathbb{M}_{\mathbf{X}_n})$, and potentially allow for more expressive models, as W could be included in the neural network architecture and learned.

For the first example, let $\mathbf{Y}_n$ be given as in (27), but with a vector of finitely exchangeable (but not i.i.d.) noise $\boldsymbol{\eta}_n \perp\!\!\!\perp \mathbf{X}_n$. Then $\mathbf{Y}_n$ would still be conditionally $\mathbb{S}_n$-equivariant, but it would not satisfy the conditional independence assumption $Y_i \perp\!\!\!\perp_{\mathbf{X}_n} (\mathbf{Y}_n \setminus Y_i)$. However, it would satisfy $Y_i \perp\!\!\!\perp_{(\mathbf{X}_n, \boldsymbol{\eta}_n)} (\mathbf{Y}_n \setminus Y_i)$, which by similar arguments as in the proof of Theorem 14, implies the existence of a representation

$$Y_i = f'(\eta'_i, \eta_i, X_i, \mathbb{M}_{(X_i,\eta_i)_{i\in[n]}}) \,, \tag{30}$$

for some other function $f' : [0,1]^2 \times \mathcal{X} \times \mathcal{M}(\mathcal{X} \times [0,1])$ and i.i.d. noise $\eta'_i \perp\!\!\!\perp (\mathbf{X}_n, \boldsymbol{\eta}_n)$, in which case (27) would be a special case.

As a second example, in practice it is possible to construct more elaborate conditionally $\mathbb{S}_n$-equivariant distributions by composing multiple ones as in Proposition 2(ii). Suppose $\mathbf{Y}_n$ is conditionally $\mathbb{S}_n$-equivariant and mutually independent given $\mathbf{X}_n$, and $\mathbf{Z}_n$ is similarly conditionally $\mathbb{S}_n$-equivariant and mutually independent given $\mathbf{Y}_n$. Proposition 2 shows that with $\mathbf{Y}_n$ marginalized out, $\mathbf{Z}_n$ is conditionally $\mathbb{S}_n$-equivariant given $\mathbf{X}_n$, while Theorem 14 guarantees the existence of the functional representations for each $i \in [n]$:

$$(Y_i)_{i\in[n]} = (f(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_n}))_{i\in[n]} \quad \text{and} \quad (Z_i)_{i\in[n]} = (f'(\eta'_i, Y_i, \mathbb{M}_{\mathbf{Y}_n}))_{i\in[n]} \,.$$

Plugging the functional representation for $\mathbf{Y}_n$ into that for $\mathbf{Z}_n$ yields, for each $i \in [n]$,

$$Z_i = f'(\eta'_i, f(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_n}), \mathbb{M}_{(f(\eta_i,X_i,\mathbb{M}_{\mathbf{X}_n}))_{i\in[n]}}) = f''(\eta'_i, \eta_i, X_i, \mathbb{M}_{(X_i,\eta_i)_{i\in[n]}}) ,$$

for some other function f'', which is a special case of (30) and which implies $Z_i \perp\!\!\!\perp_{(\mathbf{X}_n,\boldsymbol{\eta}_n)} (\mathbf{Z}_n \setminus Z_i)$.

Modeling $\mathbb{S}_n$-equivariance with neural networks. One choice for $\mathbb{S}_n$-equivariant neural networks is

$$Y_i = f(\eta_i, X_i, g(\mathbf{X}_n)) \tag{31}$$

where f is an arbitrary (stochastic) neural network module, and g an arbitrary permutation-invariant module (say one of the examples in Section 6.1). An example of a permutation-equivariant module using of a permutation-invariant submodule is shown in the second panel of Fig. 2.

Proposition 2 enables the use of an architectural algebra for constructing complex permutation-invariant and -equivariant stochastic neural network modules. Specifically, permutation-equivariant modules may be constructed by composing simpler permutation-equivariant modules (see the third panel of Fig. 2), while permutation-invariant modules can be constructed by composing permutation-equivariant modules with a final permutation-invariant module (fourth panel of Fig. 2). Some examples from the literature illustrate.

EXAMPLE 9 ($\mathbb{S}_n$-equivariant neural network layers, revisited). It is straightforward to see that Example 2 is a deterministic special case of Theorem 14:

$$Y_i = \sigma(\theta_0 X_i + \theta_1 \textstyle\sum_{j=1}^n X_j)$$

where $\sum_{j=1}^n X_j = \int_{\mathcal{X}} \mathbb{M}_{\mathbf{X}_n}(dx)$ is a function of the empirical measure. While this example's nonlinear activation of linear combinations encodes a typical feed-forward neural network structure, Theorem 14 shows that more general functional relationships are allowed, for example (31).

EXAMPLE 10 (Equivariance and convolution). Kondor and Trivedi (2018) characterized the properties of deterministic feed-forward neural networks that are equivariant under the action of a compact group, $\mathcal{G}$. Roughly speaking, their results show that each layer ℓ of the network must be a convolution of the output of the previous layer with some filter χ_ℓ. The general form of the convolution is defined in group theoretic terms that are beyond the scope of this paper. However, in the case of an exchangeable sequence, the situation is particularly simple. As an alternative to (27), Y_i may be represented by a function $f'(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_{n\setminus i}})$ (see (29) in the proof of Theorem 14), which makes clear the structure of the relationship between $\mathbf{X}_n$ and $\mathbf{Y}_n$: element Y_i has a "receptive field" that focuses on X_i and treats the elements $\mathbf{X}_{n\setminus i}$ as a background field via $\mathbb{M}_{\mathbf{X}_{n\setminus i}}$. The dependence on the latter is invariant

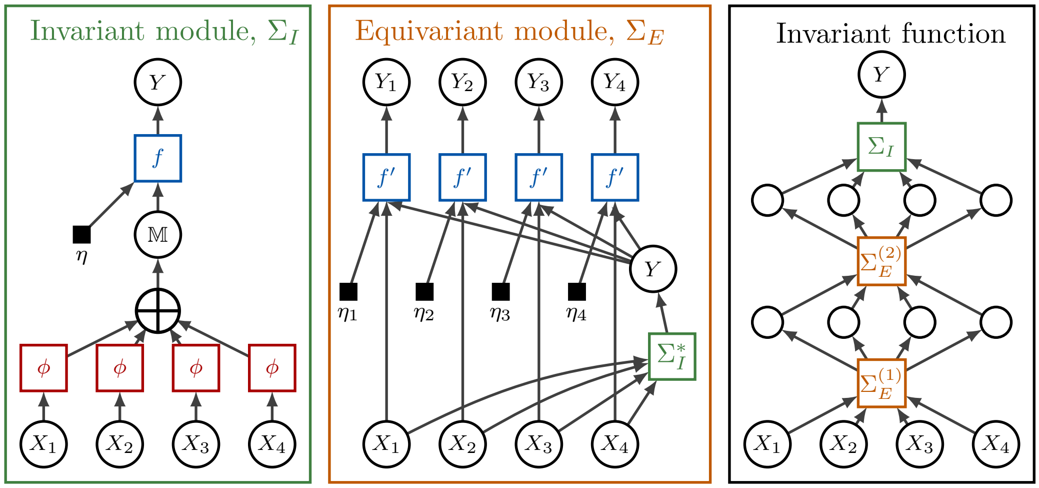


FIG 2. Left*: An invariant module depicting* (25). Middle*: An equivariant module depicting* (31)*; note that the invariant sub-module,* Σ_I^**, must be deterministic unless there are alternative conditional independence assumptions, such as* (30). Right*: An invariant stochastic function composed of equivariant modules. Functional representations of* $\mathbb{S}_n$*-invariant and -equivariant conditional distributions. Circles denote random variables, with a row denoting an exchangeable sequence. The blue squares denote arbitrary functions, possibly with outsourced noise* η *which are mutually independent and independent of everything else. Same labels mean that the functions are the same. Red squares denote arbitrary embedding functions, possibly parameterized by a neural network, and* $\oplus$ *denotes a symmetric pooling operation. Orange rectangles denote a module which gives a functional representation of a* $\mathbb{S}_n$*-equivariant conditional distribution. Likewise green rectangles for permutation-invariant conditional distributions.*

under permutations $\pi' \in \mathbb{S}_{n-1}$; in group theoretic language, $\mathbb{S}_{n-1}$ stabilizes i in $[n]$. That is, all permutations that fix i form an equivalence class. As such, for each i the index set $[n]$ is in one-to-one correspondence with the set of equivalent permutations that either move the ith element to some other element (there are $n-1$ of these), or fix i. This is the quotient space $\mathbb{S}_n/\mathbb{S}_{n-1}$; by the main result of Kondor and Trivedi (2018), any $\mathbb{S}_n$-equivariant feed-forward network with hidden layers all of size n must be composed of connections between layers $X \mapsto Y$ defined by the convolution

$$Y_i = \sigma((X * \chi)_i) = \sigma\left(\sum_{j=1}^{n} X_{(i+j) \bmod n} \chi(j) \right) , \; \chi(j) = \delta_n(j)(\theta_0 + \theta_1) + \sum_{k=1}^{n-1} \delta_k(j)\theta_1 \; .$$

The representation may be interpreted as a convolution of the previous layer's output with a filter "centered" on element X_i, and is equivalent to that of Example 9.

6.3. *Input sets of variable size.* In applications, the dataset may consist of finitely exchangeable sequences of varying length. Theorems 13 and 14 are statements about input sequences of fixed length n. In general, they suggest that a separate function

f_n needs to be learned for each n for which there is a corresponding observation $\mathbf{X}_n$ in the dataset. As a practical matter, this is clearly undesirable. In practice, the most common approach is to compute an independent embedding $\phi : \mathcal{X} \to \mathbb{R}$ of each element of an input set, and then combine the embeddings with a symmetric pooling function like sum or max. For example, $f(\mathbf{X}_n) = \max\{\phi(X_1), \ldots, \phi(X_n)\}$, or (4) from Examples 1 and 2. Clearly, such a function is invariant under permutations. However, recent work has explored the use of pairwise and higher-order interactions in the pooling function (the work by Murphy et al., 2019, mentioned in Example 7 is an example); empirical evidence indicates that the increased functional complexity results in higher model capacity and better performance for tasks that rely on modeling the dependence between elements in the input set. The following example illustrates.

EXAMPLE 11 (Self-attention). Lee et al. (2018) proposed a $\mathbb{S}_n$-invariant architecture based on *self-attention* (see also, Vaswani et al., 2017). For an input set of n d-dimensional observations, the so-called *Set Transformer* combines attention over the input dimensions with nonlinear functions of pairwise interactions between the inputs. In the simplest implementation, with no attention components, the Set Transformer computes in each network layer a nonlinear activation of the Gramian matrix $\mathbf{X}_n\mathbf{X}_n^T$; the full architecture with attention is somewhat complicated, and we refer the reader to Lee et al. (2018). Furthermore, to combat prohibitive computational cost, a method inspired by inducing point techniques from the Gaussian Process literature (Snelson and Ghahramani, 2006) was introduced. In a range of experiments focusing on tasks that benefit from modeling dependence between elements of the set, such as clustering, the Set Transformers architecture out-performed architectures that did not include pairwise or higher-order interactions, like Deep Sets (Example 1).

Intuitively, more complex functions, such as those composed of pairwise (or higher-order) functions of an input sequence, give rise to higher-capacity models able to model more complicated forms of dependence between elements of an input sequence. In the context of exchangeability, this can be made more precise, as a difference between finitely and infinitely exchangeable sequences. In particular, let $\mathbf{X}_n$ be the length n prefix of an infinitely exchangeable sequence $\mathbf{X}_\mathbb{N}$. If a sequence of sufficient statistics $S_n : \mathcal{X}^n \to \mathcal{S}_n$ exists, the conditionally i.i.d. representation (1) of the distribution of $\mathbf{X}_\mathbb{N}$ requires that they have the following properties (Freedman, 1962; Lauritzen, 1984, 1988):

(i) *symmetry under permutation*:

$$S_n(\pi \cdot \mathbf{X}_n) = S_n(\mathbf{X}_n) \quad \text{for all } \pi \in \mathbb{S}_n, \quad n \in \mathbb{N}\,; \tag{32}$$

(ii) *recursive computability*: for all $n, m \in \mathbb{N}$ there are functions $\psi_{n,m} : \mathcal{S}_n \times \mathcal{S}_m \to$

$\mathcal{S}_{n+m}$ such that

$$S_{n+m}(\mathbf{X}_{n+m}) = \psi_{n,m}(S_n(\mathbf{X}_n), S_m(\mathbf{X}_{n+m} \setminus \mathbf{X}_n)) \ . \tag{33}$$

A statistic that satisfies these properties must be of the form (Lauritzen, 1988)

$$S_n(\mathbf{X}_n) = S_1(X_1) \oplus \cdots \oplus S_1(X_n) \ , \tag{34}$$

where $(\mathcal{S}_1, \oplus)$ is an Abelian group or semigroup. Equivalently, we write $S_n(\mathbb{M}_{\mathbf{X}_n})$. Examples with $\mathcal{X} = \mathbb{R}_+$ include:

(i) $S_1(X_i) = \log X_i$ with $S_1(X_i) \oplus S_1(X_j) = \log X_i + \log X_j$;
(ii) $S_1(X_i) = X_i$ with $S_1(X_i) \oplus S_1(X_j) = X_i \vee X_j$;
(iii) $S_1(X_i) = \delta_{X_i}(\bullet)$ with $S_1(X_i) \oplus S_1(X_j) = \delta_{X_i}(\bullet) + \delta_{X_j}(\bullet)$.

Observe that pairwise (i.e., second-order) functions and higher-order statistics that do not decompose into first-order functions are precluded by the recursive computability property; an example is $S_n(\mathbf{X}_n) = \sum_{i,j\in[n]} X_i X_j$. Infinite exchangeability limits the types of dependence between elements in $\mathbf{X}_n$.

Conversely, using a model based on only first-order functions, so that properties (i) and (ii) are satisfied, limits the types of dependence that the model can capture. In practice, this can lead to shortcomings. For example, an infinitely exchangeable sequence cannot have negative correlation $\rho = \mathrm{Corr}(X_i, X_j)$, but a finitely exchangeable sequence that is not extendible to a longer exchangeable sequence can have $\rho < 0$ (e.g., Aldous, 1985, pp. 7-8). One way to interpret this fact is that the type of dependence that gives rise to negative covariance cannot be captured by first-order functions. When using $\mathbf{X}_n$ to predict another random variable, the situation becomes more complex, but to the extent that adequate (sufficient and d-separating) statistics are used, the same concepts are relevant. Ultimately, a balance between flexibility and computational efficiency must be found; the exact point of balance will depend on the details of the problem and may require novel computational methods (i.e., the inducing point-like methods in Example 11) so that more flexible function classes can be used.

6.4. *Partially exchangeable sequences and layers of different sizes.* Shawe-Taylor (1989) and Ravanbakhsh, Schneider and Póczos (2017) consider the problem of input-output invariance under a general discrete group $\mathcal{G}$ acting on a standard feed-forward network, which consists of layers, potentially with different numbers of nodes, connected by weights. Those papers each found that $\mathcal{G}$ and the neural network architecture must form a compatible pair: a pair of layers, treated as a weighted bipartite graph, forms a $\mathcal{G}$-equivariant function if the action of $\mathcal{G}$ on that graph is an automorphism. Essentially, $\mathcal{G}$ must partition the nodes of each layer into weight-preserving orbits; Ravanbakhsh, Schneider and Póczos (2017) provide some illuminating examples.

In the language of exchangeability, such a partition corresponds to *partial exchangeability*: the distributional invariance of $\mathbf{X}_{n_x}$ under the action of a subgroup of the symmetric group, $\mathcal{G} \subset \mathbb{S}_n$. Partially exchangeable analogues of Theorems 13 and 14 are possible using the same types of conditional independence arguments. The resulting functional representations would express elements of $\mathbf{Y}_{n_y}$ in terms of the empirical measures of blocks in the partition of $\mathbf{X}_{n_x}$; the basic structure is already present in Theorem 8, but the details would depend on conditional independence assumptions among the elements of $\mathbf{Y}_{n_y}$. We omit a specific statement of the result, which would require substantial notational development, for brevity.

7. Learning from finitely exchangeable matrices and graphs. Neural networks that operate on graph-valued input data have been useful for a range of tasks, from molecular design (Duvenaud et al., 2015) and quantum chemistry (Gilmer et al., 2017), to knowledge-base completion (Hamaguchi et al., 2017). See Zhou et al. (2018) for a thorough review.

In this section, we consider random matrices[10] whose distribution is invariant to permutations applied to the index set. In particular, let $\mathbf{X}_{\mathbf{n}_2}$ be a two-dimensional $\mathcal{X}$-valued array with index set $[\mathbf{n}_2] := [n_1] \times [n_2]$, such that $X_{i,j}$ is the element of $\mathbf{X}_{\mathbf{n}_2}$ at position (i, j). Let $\pi_k \in \mathbb{S}_{n_k}$ be a permutation of the set $[n_k]$ for $k \in \{1, 2\}$. Denote by $\mathbb{S}_{\mathbf{n}_2}$ the direct product $\mathbb{S}_{n_1} \times \mathbb{S}_{n_2}$. A *collection* of permutations $\boldsymbol{\pi}_2 := (\pi_1, \pi_2) \in \mathbb{S}_{\mathbf{n}_2}$ acts on $\mathbf{X}_{\mathbf{n}_2}$ in the natural way, separately on the corresponding dimension:

$$[\boldsymbol{\pi}_2 \cdot \mathbf{X}_{\mathbf{n}_2}]_{i,j} = X_{\pi_1(i),\pi_2(j)} \ . \tag{35}$$

The distribution of $\mathbf{X}_{\mathbf{n}_2}$ is *separately exchangeable* if

$$\boldsymbol{\pi}_2 \cdot \mathbf{X}_{\mathbf{n}_2} = (X_{\pi_1(i),\pi_2(j)})_{i\in[n_1],j\in[n_2]} \stackrel{\mathrm{d}}{=} (X_{i,j})_{i\in[n_1],j\in[n_2]} = \mathbf{X}_{\mathbf{n}_2} \ , \tag{36}$$

for every collection of permutations $\boldsymbol{\pi}_2 \in \mathbb{S}_{\mathbf{n}_2}$. We say that $\mathbf{X}_{\mathbf{n}_2}$ is separately exchangeable if its distribution is.

For symmetric arrays, such that $n_1 = n_2 = n$ and $X_{i,j} = X_{j,i}$, a different notion of exchangeability is needed. The distribution of a symmetric $\mathcal{X}$-valued array $\widetilde{\mathbf{X}}_n$ is *jointly exchangeable* if, for all $\pi \in \mathbb{S}_n$,

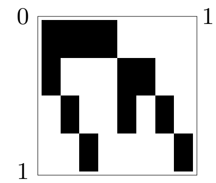
$$\pi \cdot \widetilde{\mathbf{X}}_n = (\widetilde{X}_{\pi(i),\pi(j)})_{i,j\in[n]} \stackrel{\mathrm{d}}{=} (\widetilde{X}_{i,j})_{i,j\in[n]} = \widetilde{\mathbf{X}}_n \ . \tag{37}$$

7.1. *Sufficient representations of exchangeable matrices.* In order to obtain functional representation results for matrices, a suitable analogue to the empirical measure is required. In contrast to the completely unstructured empirical measure of a sequence defined in (17), a sufficient representation of a matrix must retain the structural information encoded by the rows and columns of the matrix, but discard

[10]The results here are special cases of general results for d-dimensional arrays. For simplicity, we present the two-dimensional case and consider the general case in Appendix B.

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$\mathbf{X}_{\mathbf{n}_2}$



$\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$

FIG 3. *An example checkerboard function for a binary matrix* $\mathbf{X}_{\mathbf{n}_2} \in \{0,1\}^{4\times 8}$. *To construct* $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$, *the rows of the matrix are sorted in descending order of row sum, and then the columns are left-ordered as in Ghahramani and Griffiths (2006).*

any ordering information. For matrices, such an object corresponds to a step function (Lovász, 2012), also called the empirical graphon or a *checkerboard function* (Orbanz and Roy, 2015; Borgs and Chayes, 2017).

The checkerboard function is defined on the unit square and constructed from a matrix $\mathbf{X}_{\mathbf{n}_2}$ as follows: partition $[0,1]$ into n_1 equal-length intervals $I_j^{(1)} := [(j-1)/n_1, j/n_1)$, and separately into n_2 equal intervals $I_j^{(2)} := [(j-1)/n_2, j/n_2)$; take the product of the two partitions to divide $[0,1]^2$ into rectangular patches of area $(n_1 n_2)^{-1}$, and set

$$\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}(u,v) = X_{i,j} \quad \text{for } u \in I_i^{(1)}, v \in I_j^{(2)} \ . \tag{38}$$

Denote the space of all $\mathcal{X}$-valued checkerboard functions with the partition structure induced by $\mathbf{n}_2$ as $\mathcal{F}_{\mathbf{n}_2}(\mathcal{X})$.

This construction defines an equivalence class of checkerboard functions; in order for $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$ to be invariant under permutations of the rows and columns of $\mathbf{X}_{\mathbf{n}_2}$, $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$ should be mapped to a canonical representative of the equivalence class. Alternatively, $\mathbb{F}$ can be defined as first mapping $\mathbf{X}_{\mathbf{n}_2}$ to an orbit representative and then constructing the function on the unit square. An example of such a mapping for $\mathcal{X} = \{0,1\}$ is as follows: first sort $\mathbf{X}_{\mathbf{n}_2}$ via its row sums (in descending order), and then put the matrix in left-order (Ghahramani and Griffiths, 2006); finally, apply (38). An example is shown in Fig. 3. Observe that sorting and left-ordering $\mathbf{X}_{\mathbf{n}_2}$ is equivalent to applying some $\boldsymbol{\pi}_2 \in \mathbb{S}_{\mathbf{n}_2}$, and therefore the process can be used to define a maximal equivariant $\tau : \{0,1\}^{n_1 \times n_2} \to \mathbb{S}_{\mathbf{n}_2}$. Other domains $\mathcal{X}$ will admit different canonicalization procedures; for example, a similar process can be used for $\mathbb{R}$-valued matrices, but ties in the row sums must be broken in some consistent way, whereas left-ordering is insensitive to ties in the row sums for binary matrices.

Defining the checkerboard function on the unit square is not strictly necessary—we only require that the input set retains the grouping structure of the index set $[n_1] \times [n_2]$—but it is useful for visualization. For our purposes, the input set (the unit square) has no meaning other than as a place to embed the indices of the array.

For a finitely exchangeable array, the input set equally well could be the index set $[n_1] \times [n_2]$.[11]

The corresponding orbit law (equivalently, urn law) on $\mathcal{X}^{n_1 \times n_2}$ is constructed by symmetrizing $\mathbf{X}_{\mathbf{n}_2}$ over all collections of permutations in $\mathbb{S}_{\mathbf{n}_2}$,

$$\mathbb{U}^{\mathbb{S}_{\mathbf{n}_2}}_{\mathbf{X}_{\mathbf{n}_2}}(\,\bullet\,) = \frac{1}{n_1! n_2!} \sum_{\boldsymbol{\pi}_2 \in \mathbb{S}_{\mathbf{n}_2}} \delta_{\boldsymbol{\pi}_2 \cdot \mathbf{X}_{\mathbf{n}_2}}(\,\bullet\,) \; . \tag{39}$$

Note that $\mathbb{U}^{\mathbb{S}_{\mathbf{n}_2}}_{\mathbf{X}_{\mathbf{n}_2}}$ is a measurable function of $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$, so we may also write $\mathbb{U}^{\mathbb{S}_{\mathbf{n}_2}}_{\mathbb{F}}$ for any checkerboard function $\mathbb{F}$.

7.2. $\mathbb{S}_{\mathbf{n}_2}$*-invariant conditional distributions.* As in Section 6.1, consider modeling Y as the output of a function whose input is a separately exchangeable matrix $\mathbf{X}_{\mathbf{n}_2}$. In analogy to sequences in Theorem 13, sufficiency of the checkerboard function defined in (39) characterizes the class of all finitely exchangeable distributions on $\mathcal{X}^{n_1 \times n_2}$. The proof simply requires showing that the checkerboard function is a maximal invariant of $\mathbb{S}_{\mathbf{n}_2}$ acting on $\mathcal{X}^{n_1 \times n_2}$.

THEOREM 15. *Let $\mathbf{X}_{\mathbf{n}_2}$ be a $\mathcal{X}$-valued matrix indexed by $[n_1] \times [n_2]$. Then $\mathbf{X}_{\mathbf{n}_2}$ is separately exchangeable if and only if*

$$P(\mathbf{X}_{\mathbf{n}_2} \in \bullet \mid \mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}} = c) = \mathbb{U}^{\mathbb{S}_{\mathbf{n}_2}}_{c}(\,\bullet\,) \; .$$

If $\mathbf{X}_{\mathbf{n}_2}$ is separately exchangeable and $Y \in \mathcal{Y}$ is any other random variable such that $Y|\mathbf{X}_{\mathbf{n}_2} \overset{\mathrm{d}}{=} \boldsymbol{\pi}_2 \cdot \mathbf{X}_{\mathbf{n}_2}$ for each $\boldsymbol{\pi}_2 \in \mathbb{S}_{\mathbf{n}_2}$, then $Y \perp\!\!\!\perp_{\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}} \mathbf{X}_{\mathbf{n}_2}$, and therefore $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$ is sufficient for the family $\mathcal{P}^{\mathbb{S}_{\mathbf{n}_2}}_{\mathbf{X}_{\mathbf{n}_2}}$ and adequate for $\mathcal{P}^{\mathbb{S}_{\mathbf{n}_2}}_{\mathbf{X}_{\mathbf{n}_2},Y}$. In particular, Y is conditionally $\mathbb{S}_{\mathbf{n}_2}$ given $\mathbf{X}_{\mathbf{n}_2}$ if and only if there is a measurable function $f : [0,1] \times \mathcal{F}_{\mathbf{n}_2}(\mathcal{X}) \to \mathcal{Y}$ such that

$$(\mathbf{X}_{\mathbf{n}_2}, Y) \overset{\text{a.s.}}{=} (\mathbf{X}_{\mathbf{n}_2}, f(\eta, \mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}})) \quad \textit{where} \quad \eta \sim \mathrm{Unif}[0,1] \quad \textit{and} \quad \eta \perp\!\!\!\perp \mathbf{X}_{\mathbf{n}_2} \; . \tag{40}$$

PROOF. $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$ is a maximal invariant by construction, and therefore Theorem 12 yields the result. □

An identical result holds for jointly exchangeable matrices $\widetilde{\mathbf{X}}_n$, with symmetric checkerboard functions $\mathbb{F}_{\widetilde{\mathbf{X}}_n}$.

As was the case for sequences, any function of the checkerboard function is invariant under permutations of the index set of $\mathbf{X}_{\mathbf{n}_2}$. Most of the recent examples from the deep learning literature incorporate vertex features; that composite case is addressed in Section 7.4. A simple example without vertex features is the read-out layer of a neural network that operates on undirected, symmetric graphs.

[11]The boundedness of the unit interval does play a role in the convergence behavior of infinite sequences of growing exchangeable arrays; see Borgs and Chayes (2017).

EXAMPLE 12 (Read-out for message-passing neural networks). Gilmer et al. (2017) reviewed recent work on neural networks whose input is an undirected graph (i.e., a symmetric matrix) on vertex set $[n]$, and whose hidden layers act as message-passing operations between vertices of the graph. These message-passing hidden layers are equivariant (see Sections 7.3 and 7.4); adding a final invariant layer makes the whole network invariant. A particularly simple architecture involves a single message-passing layer and no input features on the vertices, and a typical permutation-invariant read-out layer of the form

$$R = \sum_{i \in [n]} f\bigg(\sum_{j \in [n]} h(X_{i,j}) \bigg) = f'(\mathbb{F}_{\widetilde{\mathbf{X}}_n}) \ .$$

7.3. $\mathbb{S}_{\mathbf{n}_2}$*-equivariant conditional distributions.* In analogy to $\mathbf{X}_n$ and $\mathbf{Y}_n$ in Section 6.2, $\mathbf{X}_{\mathbf{n}_2}$ and $\mathbf{Y}_{\mathbf{n}_2}$ might represent adjacent neural network layers; in such cases the goal is to transfer the symmetry of $\mathbf{X}_{\mathbf{n}_2}$ to $\mathbf{Y}_{\mathbf{n}_2}$, and permutation-equivariance is the property of interest. With a collection of permutations acting on $\mathbf{X}_{\mathbf{n}_2}$ and $\mathbf{Y}_{\mathbf{n}_2}$ as in (35), permutation-equivariance is defined in the same way as for sequences. In particular, if $\mathbf{X}_{\mathbf{n}_2}$ is exchangeable then $\mathbf{Y}_{\mathbf{n}_2}$ is conditionally $\mathbb{S}_{\mathbf{n}_2}$-equivariant if and only if

$$(\boldsymbol{\pi}_2 \cdot \mathbf{X}_{\mathbf{n}_2}, \boldsymbol{\pi}_2 \cdot \mathbf{Y}_{\mathbf{n}_2}) \overset{\text{d}}{=} (\mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2}) \quad \text{for all } \boldsymbol{\pi}_2 \in \mathbb{S}_{\mathbf{n}_2} \ . \tag{41}$$

The main result in this section is a functional representation of the conditional distribution of a conditionally $\mathbb{S}_{\mathbf{n}_2}$-equivariant array $\mathbf{Y}_{\mathbf{n}_2}$ in terms of a separately exchangeable array $\mathbf{X}_{\mathbf{n}_2}$, when the elements of $\mathbf{Y}_{\mathbf{n}_2}$ are also conditionally independent given $\mathbf{X}_{\mathbf{n}_2}$.[12] In particular, each element $Y_{i,j}$ is expressed in terms of outsourced noise $\eta_{i,j}$ and $X_{i,j}$, and an augmented checkerboard function.

Let $\mathbf{X}_{i,:}$ denote the ith row of $\mathbf{X}_{\mathbf{n}_2}$, and $\mathbf{X}_{:,j}$ the jth column; define the *separately augmented checkerboard function* as

$$\overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n}_2}, \mathbf{X}_{i,:}, \mathbf{X}_{:,j}}(u,v) = (X_{k,\ell}, X_{i,\ell}, X_{k,j}) \ , \quad u \in I_k^{(1)}, \ v \in I_\ell^{(2)} \ . \tag{42}$$

$\overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n}_2}, \mathbf{X}_{i,:}, \mathbf{X}_{:,j}}$ augments the checkerboard function $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$ with two separate partitions of the unit interval that are broadcast over the appropriate dimensions of the checkerboard function; one encodes $\mathbf{X}_{i,:}$ and one encodes $\mathbf{X}_{:,j}$. These are analogues of the empirical measures $\mathbb{M}_{\mathbf{X}_{i,:}}$ and $\mathbb{M}_{\mathbf{X}_{:,j}}$, but with their structure coupled to that of $\mathbf{X}_{\mathbf{n}_2}$. An illustration is shown in Fig. 4. Denote by $\overline{\mathcal{F}}_{\mathbf{n}_2}(\mathcal{X})$ the space of all such functions with partition structure $\mathbf{n}_2$. The augmented checkerboard function plays a role similar to the one played by the empirical measure for equivariant representations of exchangeable sequences.

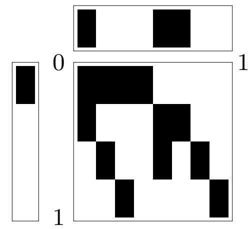
A general version of the following theorem, for d-dimensional arrays, is given in Appendix B. The proof has the same basic structure as the one for sequences in

[12] This is satisfied by neural networks without intra-layer or skip connections. Weaker conditional independence assumptions may be considered, as in Section 6.2.

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{X}_{i,:}$$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 1 & 0 & \mathbf{0} & 1 & 0 & 0 \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ 0 & 0 & 0 & 1 & \mathbf{0} & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & \mathbf{1} & 0 & 0 & 0 \end{pmatrix}$$

$\mathbf{X}_{:,j}$ $\qquad$ $\mathbf{X}_{\mathbf{n}_2}$



$\overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n}_2},\mathbf{X}_{i,:},\mathbf{X}_{:,j}}$

FIG 4. *The separately augmented checkerboard function for the binary matrix $\mathbf{X}_{\mathbf{n}_2} \in \{0,1\}^{4\times 8}$ shown on the left, with $i = 2$, $j = 5$. $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$ is augmented by the two one-dimensional checkerboard functions shown above and to the left of the two-dimensional $\mathbb{F}_{\mathbf{X}_{\mathbf{n}_2}}$. Note that sorting $\mathbf{X}_{\mathbf{n}_2}$ by row-sum and left-ordering also affects $\mathbf{X}_{i,:}$ and $\mathbf{X}_{:,j}$.*

Theorem 14, but with substantially more notation. Below, the proof for $d = 2$ is given in order to highlight the important structure.

THEOREM 16. *Suppose $\mathbf{X}_{\mathbf{n}_2}$ and $\mathbf{Y}_{\mathbf{n}_2}$ are $\mathcal{X}$- and $\mathcal{Y}$-valued arrays, respectively, each indexed by $[n_1] \times [n_2]$, and that $\mathbf{X}_{\mathbf{n}_2}$ is separately exchangeable. Assume that the elements of $\mathbf{Y}_{\mathbf{n}_2}$ are mutually conditionally independent given $\mathbf{X}_{\mathbf{n}_2}$. Then $\mathbf{Y}_{\mathbf{n}_2}$ is conditionally $\mathbb{S}_{\mathbf{n}_2}$-equivariant given $\mathbf{X}_{\mathbf{n}_2}$ if and only if there is a measurable function $f : [0,1] \times \mathcal{X} \times \overline{\mathcal{F}}_{\mathbf{n}_2}(\mathcal{X}) \to \mathcal{Y}$ such that*

$$\left(\mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2}\right) \overset{\text{a.s.}}{=} \left(\mathbf{X}_{\mathbf{n}_2}, \left(f(\eta_{i,j}, X_{i,j}, \overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n}_2},\mathbf{X}_{i,:},\mathbf{X}_{:,j}})\right)_{i\in[n_1], j\in[n_2]}\right), \tag{43}$$

for i.i.d. uniform random variables $(\eta_{i,j})_{i\in[n_1],j\in[n_2]} \perp\!\!\!\perp \mathbf{X}_{\mathbf{n}_2}$.

PROOF. First, assume that $\mathbf{Y}_{\mathbf{n}_2}$ is conditionally $\mathbb{S}_{\mathbf{n}_2}$-equivariant given $\mathbf{X}_{\mathbf{n}_2}$. Then $\boldsymbol{\pi}_2 \cdot (\mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2}) \overset{\text{d}}{=} (\mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2})$ for all $\boldsymbol{\pi}_2 \in \mathbb{S}_{\mathbf{n}_2}$. Let $\mathbb{S}_{\mathbf{n}_2}^{(i,j)} \subset \mathbb{S}_{\mathbf{n}_2}$ be the stabilizer subgroup of (i,j), i.e., the set of permutations that fixes element (i,j) in $\mathbf{X}_{\mathbf{n}_2}$. Note that each $\boldsymbol{\pi}_2^{(i,j)} \in \mathbb{S}_{\mathbf{n}_2}^{(i,j)}$ fixes both $X_{i,j}$ and $Y_{i,j}$, and that $\mathbb{S}_{\mathbf{n}_2}^{(i,j)}$ is homomorphic to $\mathbb{S}_{\mathbf{n}_2-\mathbf{1}}$. Observe that any $\boldsymbol{\pi}_2^{(i,j)} \in \mathbb{S}_{\mathbf{n}_2}^{(i,j)}$ may rearrange the elements within the ith row of $\mathbf{X}_{\mathbf{n}_2}$, but it remains the ith row in $\boldsymbol{\pi}_2^{(i,j)} \cdot \mathbf{X}_{\mathbf{n}_2}$. Similarly, the elements in the jth column, $\mathbf{X}_{:,j}$ may be rearranged but remains the jth column. As a result, the jth element of every row is fixed (though it moves with its row), as is the ith element of every column.

That fixed-element structure will be used to establish the necessary conditional independence relationships. To that end, let $r_i : [n_1] \setminus i \to [n_1 - 1]$ map the row indices of $\mathbf{X}_{\mathbf{n}_2}$ to the row indices of the matrix obtained by removing the ith row from $\mathbf{X}_{\mathbf{n}_2}$:

$$r_i(k) = \begin{cases} k & k < i \\ k-1 & k > i\,. \end{cases}$$

Analogously, let $c_j : [n_2] \setminus j \to [n_2 - 1]$ map the column indices of $\mathbf{X}_{\mathbf{n_2}}$ to those of $\mathbf{X}_{\mathbf{n_2}}$ with the jth column removed. Define the $\mathcal{X}^3$-valued array $\mathbf{Z}^{(i,j)}$ as

$$[\mathbf{Z}^{(i,j)}]_{k,\ell} = (X_{r_i(k),c_j(\ell)}, X_{i,c_j(\ell)}, X_{r_i(k),j}) \ , \quad k \in [n_1 - 1],\ \ell \in [n_2 - 1] \ . \tag{44}$$

That is, $\mathbf{Z}^{(i,j)}$ is formed by removing $\mathbf{X}_{i,:}$ and $\mathbf{X}_{:,j}$ from $\mathbf{X}_{\mathbf{n_2}}$ (and $X_{i,j}$ from $\mathbf{X}_{i,:}$ and $\mathbf{X}_{:,j}$), and broadcasting the removed row and column entries over the corresponding rows and columns of the matrix that remains. $\mathbf{Z}^{(i,j)}$ inherits the the exchangeability of $\mathbf{X}_{\mathbf{n_2}}$ in the first element of each entry, and the fixed-elements structure in the second two elements, and therefore overall it is separately exchangeable:

$$\boldsymbol{\pi}'_2 \cdot \mathbf{Z}^{(i,j)} \overset{\mathrm{d}}{=} \mathbf{Z}^{(i,j)} \ , \quad \text{for all} \quad \boldsymbol{\pi}'_2 \in \mathbb{S}_{\mathbf{n_2 - 1}} \ .$$

Now, marginally (for $Y_{i,j}$),

$$(\boldsymbol{\pi}'_2 \cdot \mathbf{Z}^{(i,j)}_{\mathbf{n_2}-1}, (X_{i,j}, Y_{i,j})) \overset{\mathrm{d}}{=} (\mathbf{Z}^{(i,j)}_{\mathbf{n_2}-1}, (X_{i,j}, Y_{i,j})) \ , \quad \text{for all} \quad \boldsymbol{\pi}'_2 \in \mathbb{S}_{\mathbf{n_2 - 1}} \ .$$

Therefore, by Theorem 15, $(X_{i,j}, Y_{i,j}) \perp\!\!\!\perp_{\mathbb{F}_{\mathbf{Z}^{(i,j)}}} \mathbf{Z}^{(i,j)}$. Conditioning on $X_{i,j}$ and $\mathbb{F}_{\mathbf{Z}^{(i,j)}}$ is equivalent to conditioning on $X_{i,j}$ and the augmented checkerboard function $\overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n_2}}, \mathbf{X}_{i,:}, \mathbf{X}_{:,j}}$, yielding

$$Y_{i,j} \perp\!\!\!\perp_{(X_{i,j}, \overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n_2}}, \mathbf{X}_{i,:}, \mathbf{X}_{:,j}})} \mathbf{X}_{\mathbf{n_2}} \ . \tag{45}$$

By Lemma 3, there is a measurable function $f_{i,j} : [0,1] \times \mathcal{X} \times \overline{\mathcal{F}}_{\mathbf{n_2}} \to \mathcal{Y}$ such that

$$Y_{i,j} = f_{i,j}(\eta_{i,j}, X_{i,j}, \overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n_2}}, \mathbf{X}_{i,:}, \mathbf{X}_{:,j}}) \ ,$$

for a uniform random variable $\eta_{i,j} \perp\!\!\!\perp \mathbf{X}_{\mathbf{n_2}}$. This is true for all $i \in [n_1]$ and $j \in [n_2]$; by equivariance the same $f_{i,j}$ must work for every (i,j). Furthermore, by assumption the elements of $\mathbf{Y}_{\mathbf{n_2}}$ are mutually conditionally independent given $\mathbf{X}_{\mathbf{n_2}}$, and therefore by the chain rule for conditional independence (Kallenberg, 2002, Prop. 6.8), the joint identity (43) holds.

The reverse direction is straightforward to verify. □

A particularly simple version of (43) is

$$Y_{i,j} = f\big(\eta_{i,j}, X_{i,j}, \sum_{k=1}^{n_2} h_1(X_{i,k}), \sum_{\ell=1}^{n_1} h_2(X_{\ell,j}), \sum_{k,\ell} h_3(X_{\ell,k})\big) \ , \tag{46}$$

for some functions $h_m : \mathcal{X} \to \mathbb{R}$, $m \in \{1,2,3\}$. Clearly, this is conditionally equivariant. The following example from the deep learning literature is an even simpler version.

EXAMPLE 13 (Array-based MLPs). Hartford et al. (2018) determined the parameter-sharing schemes that result from deterministic permutation-equivariant MLP layers for matrices. They call such layers "exchangeable matrix layers".[13] The equivariant weight-sharing scheme yields a simple expression:

$$Y_{i,j} = \sigma\left(\theta_0 + \theta_1 X_{i,j} + \theta_2 \sum_{k=1}^{n_2} X_{i,k} + \theta_3 \sum_{\ell=1}^{n_1} X_{\ell,j} + \theta_4 \sum_{k,\ell} X_{\ell,k}\right) .$$

That is, $Y_{i,j}$ is a nonlinear activation of a linear combination of $X_{i,j}$, the sums of the jth column and ith row of $\mathbf{X}_{\mathbf{n}_2}$, and the sum of the entire matrix $\mathbf{X}_{\mathbf{n}_2}$. It is straightforward to see that this is a special deterministic case of (46). Hartford et al. (2018) also derive analogous weight-sharing schemes for MLPs for d-dimensional arrays; those correspond with the d-dimensional version of Theorem 16 (see Appendix B).

Jointly exchangeable arrays. The case of jointly exchangeable symmetric arrays is of particular interest because it applies to graph-valued data. Importantly, the edge variables are not restricted to be $\{0,1\}$-valued; in practice they often take values in $\mathbb{R}_+$.

Let $\widetilde{\mathbf{X}}_n$ be a jointly exchangeable matrix in $\mathcal{X}^{n\times n}$, and let $(\widetilde{\mathbf{X}}_{i,:}, \widetilde{\mathbf{X}}_{j,:}) = ((\widetilde{X}_{i,k}, \widetilde{X}_{j,k}))_{k\in[n]}$ be the sequence of pairs of corresponding elements from $\widetilde{\mathbf{X}}_{i,:}$ and $\widetilde{\mathbf{X}}_{j,:}$. Similarly to (42), define the *jointly augmented checkerboard function* for a symmetric array as

$$\widetilde{\mathbb{F}}_{\widetilde{\mathbf{X}}_n,(\widetilde{\mathbf{X}}_{i,:},\widetilde{\mathbf{X}}_{j,:})}(u,v) = (\widetilde{X}_{k,\ell}, \{(\widetilde{X}_{i,k}, \widetilde{X}_{j,k}), (\widetilde{X}_{i,\ell}, \widetilde{X}_{j,\ell})\}) , \quad u \in I_k, v \in I_\ell . \tag{47}$$

The symmetry of $\widetilde{\mathbf{X}}_n$ requires that the row and column entries be paired, which results in the second, set-valued, element on the right-hand side of (47); Fig. 5 illustrates. (The curly braces indicate a set that is insensitive to the order of its elements, as opposed to parentheses, which indicate a sequence that is sensitive to order.) Denote by $\widetilde{\mathcal{F}}_n(\mathcal{X})$ the space of all such functions with partition structure $[n]\times[n]$. The following counterpart of Theorem 16 applies to jointly exchangeable arrays such as undirected graphs.

THEOREM 17. *Suppose $\widetilde{\mathbf{X}}_n$ and $\widetilde{\mathbf{Y}}_n$ are symmetric $\mathcal{X}$- and $\mathcal{Y}$-valued arrays, respectively, each indexed by $[n]\times[n]$, and that $\widetilde{\mathbf{X}}_n$ is jointly exchangeable. Assume that the elements of $\widetilde{\mathbf{Y}}_n$ are mutually conditionally independent given $\widetilde{\mathbf{X}}_n$. Then $\widetilde{\mathbf{Y}}_n$ is conditionally $\mathbb{S}_n$-equivariant given $\widetilde{\mathbf{X}}_n$ if and only if there is a measurable function $f : [0,1]\times\mathcal{X}\times\widetilde{\mathcal{F}}_n(\mathcal{X}) \to \mathcal{Y}$ such that*

$$(\widetilde{\mathbf{X}}_n, \widetilde{\mathbf{Y}}_n) \overset{\text{a.s.}}{=} \left(\widetilde{\mathbf{X}}_n, \left(f(\eta_{i,j}, \widetilde{X}_{i,j}, \widetilde{\mathbb{F}}_{\widetilde{\mathbf{X}}_n,(\widetilde{\mathbf{X}}_{i,:},\widetilde{\mathbf{X}}_{j,:})})\right)_{i\in[n],j\in[n]}\right) . \tag{48}$$

for i.i.d. uniform random variables $(\eta_{i,j})_{i\in[n],j\leq i} \perp\!\!\!\perp \widetilde{\mathbf{X}}_n$ with $\eta_{i,j} = \eta_{j,i}$.

[13] This is a misnomer: exchangeability is a distributional property and there is nothing random.

$$
\begin{array}{ccc}
 & \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \end{pmatrix} & \widetilde{\mathbf{X}}_{i,:} \\
\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ 1 & 0 & \mathbf{0} & 1 & 1 \\ 1 & 0 & \mathbf{0} & 0 & 0 \\ 0 & 1 & \mathbf{0} & 0 & 1 \\ 1 & 1 & \mathbf{0} & 1 & 0 \end{pmatrix} & \\
\widetilde{\mathbf{X}}_{j,:} & \widetilde{\mathbf{X}}_n &
\end{array}
$$

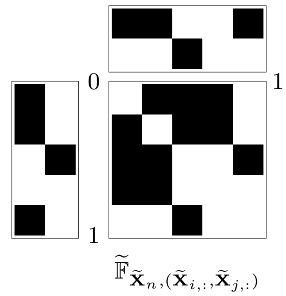


FIG 5. *The jointly augmented checkerboard function for the symmetric binary matrix $\widetilde{\mathbf{X}}_n \in \{0,1\}^{4\times 4}$ shown on the left, with $i = 1$, $j = 3$. $\mathbb{F}_{\widetilde{\mathbf{X}}_n}$ is augmented by the two $\mathcal{X}^2$-valued one-dimensional checkerboard functions shown above and to the left of the two-dimensional $\mathbb{F}_{\widetilde{\mathbf{X}}_n}$. Note that $\widetilde{\mathbf{X}}_n$ has been sorted by row-sum and left-ordered on the upper-diagonal to obtain $\widetilde{\mathbb{F}}_{\widetilde{\mathbf{X}}_n}$.*

PROOF. The proof is essentially the same as that of Theorem 16. The main difference stems from the symmetry of $\widetilde{\mathbf{X}}_n$: fixing row i also fixes column i, and fixing column j also fixes row j. Hence, the augmentations to the checkerboard function consist of the paired sequences $(\widetilde{\mathbf{X}}_{i,:}, \widetilde{\mathbf{X}}_{j,:}) = ((\widetilde{X}_{i,k}, \widetilde{X}_{j,k}))_{k\in[n]}$. The counterpart to $\mathbf{Z}^{(i,j)}$ in (44) is

$$[\widetilde{\mathbf{Z}}^{(i,j)}]_{k,\ell} = (\widetilde{X}_{k,\ell}, \{(\widetilde{X}_{i,\ell}, \widetilde{X}_{j,\ell}), (\widetilde{X}_{k,i}, \widetilde{X}_{k,j})\}) \ , \quad k, \ell \in [n-1] \ .$$

It is straightforward to show that $\widetilde{\mathbf{Z}}^{(i,j)}$ is jointly exchangeable; the rest of the argument is the same as in the proof of Theorem 16. □

7.4. *Vertex features: equivariant functions on the vertex set.* In many applications of neural networks to graph-structured data, the desired output is a function defined on the vertex set, which can be thought of as a collection of vertex features (Zhou et al., 2018). Additionally, the input graph may include vertex features. Formally, these features are encoded as a sequence $\mathbf{X}_n$ whose elements correspond to the rows and columns of a symmetric matrix $\widetilde{\mathbf{X}}_n$. We assume for simplicity that both $\mathbf{X}_n$ and $\widetilde{\mathbf{X}}_n$ are $\mathcal{X}$-valued, but it is not strictly necessary that they take values in the same domain. In the separately exchangeable case, two distinct feature sequences $\mathbf{X}_n^{(1)}$ and $\mathbf{X}_n^{(2)}$ correspond to rows and columns, respectively. For simplicity, we focus on the symmetric case. A permutation $\pi \in \mathbb{S}_n$ acts simultaneously on $\mathbf{X}_n$ and $\widetilde{\mathbf{X}}_n$, and the feature-augmented graph is jointly exchangeable if

$$(\pi \cdot \mathbf{X}_n, \pi \cdot \widetilde{\mathbf{X}}_n) \stackrel{\text{d}}{=} (\mathbf{X}_n, \widetilde{\mathbf{X}}_n) \ , \quad \text{for all } \pi \in \mathbb{S}_n \ . \tag{49}$$

Likewise, we may define a vertex feature sequence $\mathbf{Y}_n$, whose elements correspond to the rows and columns of $\widetilde{\mathbf{Y}}_n$.

Consider the array $\widetilde{\mathbf{X}}'_n$, defined as $\widetilde{\mathbf{X}}_n$ augmented with the vertex features broadcast over the appropriate dimensions: $\widetilde{X}'_{i,j} = (\widetilde{X}_{i,j}, \{X_i, X_j\})$. Observe that $\widetilde{\mathbf{X}}'_n$

is a jointly exchangeable array because $(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$ is exchangeable as in (49). Furthermore, if $\widetilde{\mathbf{Y}}_n$ is conditionally $\mathbb{S}_n$-equivariant given $\widetilde{\mathbf{X}}_n$, then $\widetilde{\mathbf{Y}}'_n$, with entries $\widetilde{Y}'_{i,j} = (\widetilde{Y}_{i,j}, \{Y_i, Y_j\})$, is conditionally $\mathbb{S}_n$-equivariant given $\widetilde{\mathbf{X}}'_n$. However, Theorem 17 does not apply: $\widetilde{\mathbf{Y}}'_n$ does not satisfy the mutually conditionally independent elements assumption because, for example, Y_i appears in every element in the ith row of $\widetilde{\mathbf{Y}}'_n$.

Further assumptions are required to obtain a useful functional representation of $\widetilde{\mathbf{Y}}_n$. In addition to the mutual conditional independence of the elements of $\widetilde{\mathbf{Y}}_n$ given $(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$, we assume further that

$$Y_i \perp\!\!\!\perp_{(\mathbf{X}_n, \widetilde{\mathbf{X}}_n, \widetilde{\mathbf{Y}}_n)} (\mathbf{Y}_n \setminus Y_i) \;, \quad i \in [n] \;. \tag{50}$$

In words, the elements of the output vertex feature sequence are conditionally independent, given the input data $\mathbf{X}_n, \widetilde{\mathbf{X}}_n$ and the output edge features $\widetilde{\mathbf{Y}}_n$. This is consistent with the implicit assumptions used in practice in the deep learning literature, and leads to a representation with simple structure.

To state the result, let $\widetilde{\mathbb{F}}_{\widetilde{\mathbf{X}}_n, \mathbf{X}_n, (\widetilde{\mathbf{X}}_{i,:}, \widetilde{\mathbf{X}}_{j,:})}$ be the augmented checkerboard function that includes the vertex features:

$$\widetilde{\mathbb{F}}_{\widetilde{\mathbf{X}}_n, \mathbf{X}_n, (\widetilde{\mathbf{X}}_{i,:}, \widetilde{\mathbf{X}}_{j,:})}(u,v) = (\widetilde{X}_{k,\ell}, \{X_k, X_\ell\}, \{(\widetilde{X}_{i,\ell}, \widetilde{X}_{j,\ell}), (\widetilde{X}_{k,i}, \widetilde{X}_{k,j})\}) \;, \quad u \in I_k, v \in I_\ell \;,$$

and let $(\widetilde{\mathbf{X}}_n, \widetilde{\mathbf{Y}}_n)$ be the symmetric array with entries $(\widetilde{X}_{i,j}, \widetilde{Y}_{i,j})$.

THEOREM 18. *Suppose $(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$ and $(\mathbf{Y}_n, \widetilde{\mathbf{Y}}_n)$ are $\mathcal{X}$- and $\mathcal{Y}$-valued vertex feature-augmented arrays, and that $(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$ is jointly exchangeable as in (49). Assume that the elements of $\widetilde{\mathbf{Y}}_n$ are mutually conditionally independent given $(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$, and that $\mathbf{Y}_n$ satisfies (50). Then $(\mathbf{Y}_n, \widetilde{\mathbf{Y}}_n)$ is conditionally $\mathbb{S}_n$-equivariant given $(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$ if and only if there are measurable functions $f_e : [0,1] \times \mathcal{X} \times \widetilde{\mathcal{X}}^2 \times \widetilde{\mathcal{F}}_n(\mathcal{X}) \to \mathcal{Y}$ and $f_v : [0,1] \times \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \times \widetilde{\mathcal{F}}_n(\mathcal{X} \times \mathcal{Y}) \to \mathcal{Y}$ such that*

$$\widetilde{Y}_{i,j} \overset{\text{a.s.}}{=} f_e\big(\eta_{i,j}, \widetilde{X}_{i,j}, \{X_i, X_j\}, \widetilde{\mathbb{F}}_{\widetilde{\mathbf{X}}_n, \mathbf{X}_n, (\widetilde{\mathbf{X}}_{i,:}, \widetilde{\mathbf{X}}_{j,:})}\big) \;, \quad i,j \in [n] \tag{51}$$

$$Y_i \overset{\text{a.s.}}{=} f_v\big(\eta_i, X_i, \widetilde{X}_{i,i}, \widetilde{Y}_{i,i}, \widetilde{\mathbb{F}}_{(\widetilde{\mathbf{X}}_n, \widetilde{\mathbf{Y}}_n), \mathbf{X}_n, (\widetilde{\mathbf{X}}_{i,:}, \widetilde{\mathbf{X}}_{j,:})}\big) \;, \quad i \in [n] \;, \tag{52}$$

for i.i.d. uniform random variables $(\eta_{i,j})_{i\in[n], j\le i} \perp\!\!\!\perp (\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$ with $\eta_{i,j} = \eta_{j,i}$, and $(\eta_i)_{i\in[n]} \perp\!\!\!\perp (\mathbf{X}_n, \widetilde{\mathbf{X}}_n, \widetilde{\mathbf{Y}}_n)$.

PROOF. The proof, like that of Theorem 17, is essentially the same as for Theorem 16. Incorporating vertex features require that for any permutation $\pi \in \mathbb{S}_n$, the fixed elements of $\mathbf{X}_n$ be collected along with the fixed rows and columns of $\widetilde{\mathbf{X}}_n$; the structure of the argument is identical. $\square$

Equations (51) and (52) indicate that given an input $(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$ and functional forms for f_e and f_v, computation of $\widetilde{\mathbf{Y}}_n$ and $\mathbf{Y}_n$ proceeds in two steps: first, compute the elements $\widetilde{Y}_{i,j}$ of $\widetilde{\mathbf{Y}}_n$; second, compute the vertex features $\mathbf{Y}_n$ from $\mathbf{X}_n$, $\widetilde{\mathbf{X}}_n$, and $\widetilde{\mathbf{Y}}_n$. Note that within each step, computations can be parallelized due to the conditional independence assumptions.

The following examples from the literature are special cases of Theorem 18.

EXAMPLE 14 (Graph-based structured prediction). Herzig et al. (2018) considered the problem of deterministic permutation-equivariant structured prediction in the context of mapping images to scene graphs. In particular, for a weighted graph with edge features $\widetilde{\mathbf{X}}_n$ and vertex features $\mathbf{X}_n$, those authors define a graph labeling function f to be "graph-permutation invariant" (GPI) if $f(\pi \cdot (\mathbf{X}_n, \widetilde{\mathbf{X}}_n)) = \pi \cdot f(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)$ for all $\pi \in \mathbb{S}_n$.[14] Furthermore, they implicitly set $\widetilde{\mathbf{Y}}_n = \widetilde{\mathbf{X}}_n$, and assume that $\widetilde{X}_{i,i} = 0$ for all $i \in [n]$ and that $\mathbf{X}_n$ is included in $\widetilde{\mathbf{X}}_n$ (e.g., on the diagonal). The main theoretical result is that a graph labeling function is GPI if and only if

$$[f(\mathbf{X}_n, \widetilde{\mathbf{X}}_n)]_i = \rho(X_i, \textstyle\sum_j \alpha(X_j, \sum_k \phi(X_j, \widetilde{X}_{j,k}, X_k))) \ ,$$

for appropriately defined functions ρ, α, and ϕ. Inspection of the proof reveals that the second argument of ρ (the sum of α functions) is equivalent to the checkerboard function. In experiments, a particular GPI neural network architecture showed better sample efficiency for training, as compared with an LSTM with the inputs in random order and with a fully connected feed-forward network, each network with the same number of parameters.

EXAMPLE 15 (Message passing graph neural networks). Gilmer et al. (2017) reviewed the recent literature on graph-based neural network architectures, and found that many of them fit in the framework of so-called message passing neural networks (MPNNs). MPNNs take as input a graph with vertex features and edge features (the features may be vector-valued, but for simplicity of notation we assume they are scalar-valued real numbers). Each neural network layer ℓ acts as a round of message passing between adjacent vertices, with the typically edge-features held fixed from layer to layer. In particular, denote the (fixed) input edge features by $\widetilde{\mathbf{X}}_n$ and the computed vertex features of layer $\ell - 1$ by $\mathbf{X}_n$; then the vertex features for layer ℓ, $\mathbf{Y}_n$, are computed as

$$Y_i = U_\ell\big(X_i, \textstyle\sum_{j\in[n]} \mathbb{1}_{\{\widetilde{X}_{i,j}>0\}} M_\ell(X_i, X_j, \widetilde{X}_{i,j})\big) \ , \quad i \in [n] \ ,$$

for some functions $M_\ell : \mathbb{R}^3 \to \mathbb{R}$ and $U_\ell : \mathbb{R}^2 \to \mathbb{R}$. This is a deterministic special case of (52). Gilmer et al. (2017) note that Kearnes et al. (2016) also compute different edge features, $\widetilde{Y}_{i,j}$, in each layer. The updated edge features are used in place of $\widetilde{X}_{i,j}$ in the third argument of M_ℓ in the equation above, and are an example of the two-step implementation of (51)-(52).

[14] Although Herzig et al. (2018) call this invariance, we note that it is actually equivariance.

Separate exchangeability with features. Separately exchangeable matrices with features, which may be regarded as representing a bipartite graph with vertex features, have a similar representation. Let $\mathbf{X}_n^{(1)}$ and $\mathbf{X}_n^{(2)}$ denote the row and column features, respectively. Separate exchangeability for such data structures is defined in the obvious way. Furthermore, a similar conditional independence assumption is made for the output vertex features: $\mathbf{Y}_n^{(1)} \perp\!\!\!\perp_{(\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2})} \mathbf{Y}_n^{(2)}$ and

$$Y_i^{(m)} \perp\!\!\!\perp_{(\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2})} (\mathbf{Y}_n^{(m)} \setminus Y_i^{(m)}) \;, \quad i \in [n_m], m \in \{1,2\} \;. \tag{53}$$

The representation result is stated here for completeness. The proof is similar to that of Theorem 18, and is omitted for brevity.

THEOREM 19. *Suppose $(\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2})$ and $(\mathbf{Y}_n^{(1)}, \mathbf{Y}_n^{(2)}, \mathbf{Y}_{\mathbf{n}_2})$ are $\mathcal{X}$- and $\mathcal{Y}$-valued vertex feature-augmented arrays, and that $(\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2})$ is separately exchangeable. Assume that the elements of $\mathbf{Y}_{\mathbf{n}_2}$ are mutually conditionally independent given $(\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2})$, and that $\mathbf{Y}_n^{(1)}$ and $\mathbf{Y}_n^{(2)}$ satisfy (53). Then $(\mathbf{Y}_n^{(1)}, \mathbf{Y}_n^{(2)}, \mathbf{Y}_{\mathbf{n}_2})$ is conditionally $\mathbb{S}_{\mathbf{n}_2}$-equivariant given $(\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2})$ if and only if there are measurable functions $f_e : [0,1] \times \mathcal{X}^3 \times \overline{\mathcal{F}}_n(\mathcal{X}) \to \mathcal{Y}$ and $f_v^{(m)} : [0,1] \times \mathcal{X} \times \mathcal{X}^{n_m} \times \mathcal{Y}^{n_m} \times \overline{\mathcal{F}}_n(\mathcal{X} \times \mathcal{Y}) \to \mathcal{Y}$, for $m \in \{1,2\}$, such that*

$$Y_{i,j} \overset{\text{a.s.}}{=} f_e\big(\eta_{i,j}, X_{i,j}, X_i^{(1)}, X_j^{(2)}, \overline{\mathbb{F}}_{\mathbf{X}_{\mathbf{n}_2}, (\mathbf{X}_n^{(1)}, \mathbf{X}_{:,j}), (\mathbf{X}_n^{(2)}, \mathbf{X}_{i,:})}\big) \;, \quad i \in [n_1], j \in [n_2] \tag{54}$$

$$Y_i^{(1)} \overset{\text{a.s.}}{=} f_v^{(1)}\big(\eta_i^{(1)}, X_i^{(1)}, \mathbf{X}_{i,:}, \mathbf{Y}_{i,:}, \overline{\mathbb{F}}_{(\mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2}), \mathbf{X}_n^{(1)}, (\mathbf{X}_n^{(2)}, \mathbf{X}_{i,:})}\big) \;, \quad i \in [n_1] \;, \tag{55}$$

$$Y_j^{(2)} \overset{\text{a.s.}}{=} f_v^{(2)}\big(\eta_j^{(2)}, X_j^{(2)}, \mathbf{X}_{:,j}, \mathbf{Y}_{:,j}, \overline{\mathbb{F}}_{(\mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2}), (\mathbf{X}_n^{(1)}, \mathbf{X}_{:,j}), \mathbf{X}_n^{(2)}}\big) \;, \quad j \in [n_2] \;, \tag{56}$$

for i.i.d. uniform random variables $((\eta_{i,j})_{i \in [n_1], j \in [n_2]}) \perp\!\!\!\perp (\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2})$ and $((\eta_i^{(1)})_{i \in [n_1]}, (\eta_j^{(2)})_{j \in [n_2]}) \perp\!\!\!\perp (\mathbf{X}_n^{(1)}, \mathbf{X}_n^{(2)}, \mathbf{X}_{\mathbf{n}_2}, \mathbf{Y}_{\mathbf{n}_2})$.

8. Discussion. The probabilistic approach to symmetry has allowed us to draw on tools from an area that is typically outside the purview of the deep learning community. Those tools shed light on the underlying structure of previous work on invariant neural networks, and expand the scope of what is possible with such networks. Moreover, those tools place invariant neural networks in a broader statistical context, making connections to the fundamental concepts of sufficiency and adequacy.

To conclude, we give some examples of other data structures to which the theory developed in the present work could be applied and describe some questions prompted by this work and by others.

8.1. *Other exchangeable structures.* The results in Section 6 can be adapted to other exchangeable structures in a straightforward manner. We briefly describe two settings; Orbanz and Roy (2015) survey a number of other exchangeable structures from statistics and machine learning to which this work could apply.

Edge-exchangeable graphs. Cai, Campbell and Broderick (2016); Williamson (2016); Crane and Dempsey (2018) specified generative models for network data as an exchangeable sequence of edges, $\mathbf{E}_n = ((u,v)_1, \dots, (u,v)_n)$. The sequence theory from Section 6 applies, rather than the array theory from Section 7, which applies to vertex-exchangeable graphs. However, incorporating vertex features into edge-exchangeable models would require some extra work, as a permutation acting on the edge sequence has a different (random) induced action on the vertices; we leave it as an interesting problem for future work.

The model of Caron and Fox (2017) is finitely edge-exchangeable when conditioned on the random number of edges in the network. Therefore, the sequence theory could be incorporated into inference procedures for that model, with the caveat that the neural network architecture would be required to accept inputs of variable size and the discussion from Section 6.3 would be relevant.

Markov chains and recurrent processes. A Markov chain $\mathbf{Z}_n := (Z_1, Z_2, \dots, Z_n)$ on state-space $\mathcal{Z}$ has exchangeable sub-structure. In particular, define a *z-block* as a sub-sequence of $\mathbf{Z}_n$ that starts and ends on some state $z \in \mathcal{Z}$. Clearly, the joint probability

$$P_{\mathbf{Z}_n}(\mathbf{Z}_n) = P_{Z_1}(Z_1) \prod_{i=2}^{n} Q(Z_i \mid Z_{i-1})$$

is invariant under all permutations of Z_1-blocks (and of any other z-blocks for $z \in \mathbf{Z}_n$). Denote these blocks by $(B_{Z_1,j})_{j\geq 1}$. Diaconis and Freedman (1980b) used this notion of *Markov exchangeability* to show that all recurrent processes on countable state-spaces are mixtures of Markov chains, and Bacallado, Favaro and Trippa (2013) used similar ideas to analyze reversible Markov chains on uncountable state-spaces.

For a finite Markov chain, the initial state, Z_1, and the Z_1-blocks are sufficient statistics. Equivalently, Z_1 and the empirical measure of the $m \leq n$ Z_1-blocks,

$$\mathbb{M}_{\mathbf{Z}_n}(\,\bullet\,) = \sum_{j=1}^{m} \delta_{B_{Z_1,j}}(\,\bullet\,) ,$$

plays the same role as $\mathbb{M}_{\mathbf{X}_n}$ for an exchangeable sequence. (If $\mathbf{Z}_n$ is the prefix of an infinite, recurrent process, then Z_1 and the empirical measure of transitions are sufficient.) It is clear that the theory from Section 6 can be adapted to accommodate Markov chains.

8.2. *Open questions.*

More flexible conditional independence assumptions. The simplicity of results on functional representations of equivariant conditional distributions relied on conditional independence assumptions like $Y_i \perp\!\!\!\perp_{\mathbf{X}_n} (\mathbf{Y}_n \setminus Y_i)$. As discussed in Section 6.2, additional flexibility could be obtained by assuming the existence of a random variable W such that $Y_i \perp\!\!\!\perp_{(\mathbf{X}_n, W)} (\mathbf{Y}_n \setminus Y_i)$ holds, and learning W for each layer. W

could be interpreted as providing additional shared "context" for the input-output relationship, a construct that has been useful in recent work (Edwards and Storkey, 2017; Kim et al., 2019).

Choice of pooling functions. The results here are quite general, and say nothing about what other properties the function classes under consideration should have. For example, different symmetric pooling functions may lead to very different performance. This seems to be well understood in practice; for example, the pooling operation of element-wise mean of the top decile in Chan et al. (2018) is somewhat unconventional, but appears to have been chosen based on performance. Recent theoretical work by Xu et al. (2019) studies different pooling operations in the context of graph discrimination tasks with graph neural networks, a problem also considered from a slightly different perspective by Shawe-Taylor (1993).

Learning and generalization. Our results leave open questions pertaining to learning and generalization. However, they point to some potential approaches, one of which we sketch here. Shawe-Taylor (1991, 1995) applied PAC theory to derive generalization bounds for feed-forward networks that employ a symmetry-induced weight-sharing scheme; these bounds are improvements on those for standard fully-connected multi-layer perceptrons. Weight-sharing might be viewed as a form of pre-training compression; recent work uses PAC-Bayes theory to demonstrate the benefits of compressibility of trained networks (Arora et al., 2018; Zhou et al., 2019).

APPENDIX A: PROOFS

A.1. Proof of Lemma 3. d-separation is a statement of conditional independence and therefore the proof of Lemma 3 is a straightforward application of the following basic result about conditional independence, which appears (with different notation) as Proposition 6.13 in Kallenberg (2002).

LEMMA 20 (Conditional independence and randomization). *Let X, Y, Z be random elements in some measurable spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, respectively, where $\mathcal{Y}$ is Borel. Then $Y \perp\!\!\!\perp_Z X$ if and only if $Y \overset{\text{a.s.}}{=} f(\eta, Z)$ for some measurable function $f : [0,1] \times \mathcal{Z} \to \mathcal{Y}$ and some uniform random variable $\eta \perp\!\!\!\perp (X, Z)$.*

A.2. Proof of Lemma 5.

PROOF OF LEMMA 5. For the first claim, a similar result appears in Eaton (1989, Ch. 4). For completeness, a slightly different proof is given here. Suppose that P_X is $\mathcal{G}$-invariant. Then $(g \cdot X, M) \overset{\text{d}}{=} (X, M)$ for all $g \in \mathcal{G}$. By Fubini's theorem, the statement is true even for a random group element $G \perp\!\!\!\perp X$. Let G be a random element of $\mathcal{G}$, sampled from $\lambda_{\mathcal{G}}$. Then for any measurable function $\varphi : \mathcal{X} \to \mathbb{R}_+$, and for any set $A \in \sigma(M)$,

$$\mathbb{E}[\varphi(X); A] = \mathbb{E}[\varphi(G \cdot X); A)] = \mathbb{E}[\mathbb{E}[\varphi(G \cdot X) \mid X]; A] = \mathbb{E}[\mathbb{U}^{\mathcal{G}}_{M(X)}[\varphi]; A] \; ,$$

which establishes the first equality of (11); the second equality follows because $\mathcal{S}$ is a Borel space, as is $\mathcal{X}$ (Kallenberg, 2002, Thm. 6.3). Conversely, suppose that (11) is true. $P_X(X \in \bullet) = \mathbb{E}[\mathbb{U}^{\mathcal{G}}_M(\bullet)]$, where the expectation is taken with respect to M. $\mathbb{U}^{\mathcal{G}}_M(\bullet)$ is $\mathcal{G}$-invariant for all $M \in \mathcal{S}$, and therefore so is the marginal distribution P_X.

For the second claim, suppose Y is such that $P_{Y|X}$ is $\mathcal{G}$-invariant, which by Proposition 1 is equivalent to $(g \cdot X, Y) \overset{\mathrm{d}}{=} (X, Y)$ for all $g \in \mathcal{G}$. The latter is equivalent to $g \cdot (X, Y) \overset{\mathrm{d}}{=} (X, Y)$, with $g \cdot Y = Y$ almost surely, for all $g \in \mathcal{G}$. Therefore, $\mathcal{G} \cdot Y = \{Y\}$ almost surely, and $\widetilde{M}(X, Y) := (M(X), Y)$ is a maximal invariant of $\mathcal{G}$ acting on $\mathcal{X} \times \mathcal{Y}$. Therefore, $\mathbb{U}^{\mathcal{G}}_{\widetilde{M}(X,Y)} = \mathbb{U}^{\mathcal{G}}_{M(X)} \otimes \delta_Y$, and

$$P_X(X \in \bullet \mid M(X) = m, Y) = \mathbb{U}^{\mathcal{G}}_m(\bullet) = P_X(X \in \bullet \mid M(X) = m) \ ,$$

which implies $Y \perp\!\!\!\perp_{M(X)} X$. The converse is straightforward to check. □

A.3. Proof of Lemma 7.

Proof of Lemma 7. Property (i) is proved in Eaton (1989), Ch. 2, as follows. Observe that for all $g \in \mathcal{G}$ and $x \in \mathcal{X}$,

$$M_\tau(g \cdot x) = \tau^{-1}_{g \cdot x} \cdot (g \cdot x) = (g \cdot \tau_x)^{-1} \cdot (g \cdot x) = \tau^{-1}_x \cdot g^{-1} \cdot g \cdot x = \tau^{-1}_x \cdot x = M_\tau(x) \ ,$$

so $M_\tau(x)$ is invariant. Now suppose $M_\tau(x_1) = M_\tau(x_2)$ for some x_1, x_2, and define $g = \tau_{x_1} \cdot \tau^{-1}_{x_2}$. Then

$$\begin{aligned} \tau^{-1}_{x_1} \cdot x_1 &= \tau^{-1}_{x_2} \cdot x_2 \\ x_1 &= (\tau_{x_1} \cdot \tau^{-1}_{x_2}) \cdot x_2 = g \cdot x_2 \ . \end{aligned}$$

Therefore, x_1 and x_2 are in the same orbit and M_τ is a maximal invariant.

Our proof of property (ii) is adapted from that of Kallenberg (2005), Lemma 7.10. Observe that due to the invariance of $M_\tau(x)$,

$$\begin{aligned} \tau^{-1}_{g \cdot x} \cdot (g \cdot x) &= \tau^{-1}_x \cdot x \\ \tau_x \cdot \tau^{-1}_{g \cdot x} \cdot (g \cdot x) &= x \ , \end{aligned} \tag{57}$$

so $\tau_x \cdot \tau^{-1}_{g \cdot x} \cdot g \in \mathcal{G}_x$, for all $g \in \mathcal{G}$, $x \in \mathcal{X}$. Similarly, $g^{-1} \cdot \tau_{g \cdot x} \cdot \tau^{-1}_x \in \mathcal{G}_x$. Now consider an arbitrary mapping $b : \mathcal{X} \to \mathcal{Y}$, and define the function f by (14). Then

$$\begin{aligned} f(g \cdot x) &= \tau_{g \cdot x} \cdot b(\tau^{-1}_{g \cdot x} \cdot g \cdot x) = \tau_{g \cdot x} \cdot b(\tau^{-1}_x \cdot x) = \tau_{g \cdot x} \cdot \tau^{-1}_x \cdot f(x) \\ &= g \cdot g^{-1} \cdot \tau_{g \cdot x} \cdot \tau^{-1}_x \cdot f(x) = g \cdot f(x) \ , \end{aligned}$$

where the last equality follows from the hypothesis $\mathcal{G}_x \subseteq \mathcal{G}_{f(x)}$. □

A.4. Proof of Theorem 8.

Proof of Theorem 8. The proof of sufficiency closely follows the proof of Lemma 7.11 in Kallenberg (2005). It relies on proving that $\tau_X^{-1} \cdot Y \perp\!\!\!\perp_{M_\tau} X$, and applying Lemmas 3 and 7.

Assume that $g \cdot (X, Y) \overset{\text{d}}{=} (X, Y)$ for all $g \in \mathcal{G}$. Let $M_\tau(x) = \tau_x^{-1} \cdot x$, and for any $x \in \mathcal{X}$, let $M'_{\tau,x} : \mathcal{Y} \to \mathcal{Y}$ be defined by $M'_{\tau_x}(y) = \tau_x^{-1} \cdot y$, for $y \in \mathcal{Y}$. As shown in (57), $\tau_x \cdot \tau_{g\cdot x}^{-1} \cdot g \in \mathcal{G}_x$, and therefore $\tau_X \cdot \tau_{g\cdot X}^{-1} \cdot g \in \mathcal{G}_X \subseteq \mathcal{G}_Y$, almost surely. Therefore, the random elements $M_\tau(X) \in \mathcal{X}$ and $M'_{\tau,X}(Y) \in \mathcal{Y}$ satisfy

$$(M_\tau(X), M'_{\tau,X}(Y)) = \tau_x^{-1} \cdot (X, Y) = \tau_{g\cdot X}^{-1} \cdot g \cdot (X, Y) \,, \quad g \in \mathcal{G} \,. \tag{58}$$

Now let $G \sim \lambda_{\mathcal{G}}$ be a random element of $\mathcal{G}$ such that $G \perp\!\!\!\perp (X, Y)$. Because $g\cdot(X, Y) \overset{\text{d}}{=} (X, Y)$ and $g \cdot G \overset{\text{d}}{=} G \cdot g \overset{\text{d}}{=} G$ for all $g \in \mathcal{G}$, and using Fubini's theorem,

$$G \cdot (X, Y) \overset{\text{d}}{=} (X, Y) \,, \quad \text{and} \quad (G \cdot \tau_X^{-1}, X, Y) \overset{\text{d}}{=} (G, X, Y) \,. \tag{59}$$

Using (58) and (59),

$$\begin{aligned}
(X, M_\tau(X), M'_{\tau,X}(Y)) &= (X, \tau_X^{-1} \cdot X, \tau_X^{-1} \cdot Y) \\
&\overset{\text{d}}{=} (G \cdot X, \tau_{G\cdot X}^{-1} \cdot G \cdot X, \tau_{G\cdot X}^{-1} \cdot G \cdot Y) \\
&= (G \cdot X, \tau_X^{-1} \cdot X, \tau_X^{-1} \cdot Y) \\
&\overset{\text{d}}{=} (G \cdot \tau_X^{-1} \cdot X, \tau_X^{-1} \cdot X, \tau_X^{-1} \cdot Y) \\
&= (G \cdot M_\tau(X), M_\tau(X), M'_{\tau,X}(Y))
\end{aligned} \tag{60}$$

That is, even jointly with the orbit representative $M_\tau(X)$ and the maximal equivariant applied to Y, $M'_{\tau,X}(Y)$, the distribution of X is the same as if applying a random group element $G \sim \lambda_{\mathcal{G}}$ to the orbit representative.

Now, $G \perp\!\!\!\perp (X, Y)$ implies that $G \cdot \tau_X^{-1} \cdot X \perp\!\!\!\perp_{M_\tau(X)} \tau_X^{-1} \cdot Y$, which by (60) implies

$$M'_{\tau,X}(Y) \perp\!\!\!\perp_{M_\tau(X)} X \,.$$

Therefore, by Lemma 3, there exists some measurable $b : [0, 1] \times \mathcal{X} \to \mathcal{Y}$ such that $M'_{\tau,X}(Y) = \tau_X^{-1} \cdot Y \overset{\text{a.s.}}{=} b(\eta, M_\tau(X))$ for $\eta \sim \text{Unif}[0, 1]$ and $\eta \perp\!\!\!\perp X$. By Lemma 7,

$$Y \overset{\text{a.s.}}{=} \tau_X \cdot b(\eta, M_\tau(X)) = \tau_X \cdot b(\eta, \tau_X^{-1} \cdot X) =: f(\eta, X)$$

is $\mathcal{G}$-equivariant in the second argument. This establishes sufficiency.

Conversely, for necessity, using (15) and the assumption that $g \cdot X \overset{\text{d}}{=} X$ for all $g \in \mathcal{G}$,

$$g \cdot (X, Y) = (g \cdot X, f(\eta, g \cdot X)) \overset{\text{d}}{=} (X, f(\eta, X)) = (X, Y) \,.$$

$\square$

APPENDIX B: REPRESENTATIONS OF D-DIMENSIONAL ARRAYS

The functional representations of conditionally $\mathbb{S}_{\mathbf{n}_2}$-invariant (Theorem 15) and -equivariant (Theorem 16) distributions in Section 7 are special cases with $d = 2$ of more general results for d-dimensional arrays.

Some notation is needed in order to state the results. For a fixed $d \in \mathbb{N}$, let $\mathbf{X}_{\mathbf{n}_d}$ be a d-dimensional $\mathcal{X}$-valued array (called a d-array) with index set $[\mathbf{n}_d] := [n_1] \times \cdots \times [n_d]$, and with $X_{i_1,\dots,i_d}$ the element of $\mathbf{X}_{\mathbf{n}_d}$ at position $(i_1, \dots, i_d)$. The size of each dimension is encoded in the vector of integers $\mathbf{n}_d = (n_1, \dots, n_d)$. In this section, we consider random d-arrays whose distribution is invariant to permutations applied to the index set. In particular let $\pi_k \in \mathbb{S}_{n_k}$ be a permutation of the set $[n_k]$. Denote by $\mathbb{S}_{\mathbf{n}_d} := \mathbb{S}_{n_1} \times \cdots \times \mathbb{S}_{n_d}$ the direct product of each group $\mathbb{S}_{n_k}$, $k \in [d]$.

A collection of permutations $\boldsymbol{\pi}_d := (\pi_1, \dots, \pi_d) \in \mathbb{S}_{\mathbf{n}_d}$ acts on $\mathbf{X}_{\mathbf{n}_d}$ in the natural way, separately on the corresponding input dimension of $\mathbf{X}_{\mathbf{n}_d}$:

$$[\boldsymbol{\pi}_d \cdot \mathbf{X}_{\mathbf{n}_d}]_{i_1,\dots,i_d} = X_{\pi_1(i_1),\dots,\pi_d(i_d)} \ .$$

The distribution of $\mathbf{X}_{\mathbf{n}_d}$ is separately exchangeable if

$$(X_{i_1,\dots,i_d})_{(i_1,\dots,i_d)\in[n_1,\dots,d]} \overset{\text{d}}{=} (X_{\pi_1(i_1),\dots,\pi_d(i_d)})_{(i_1,\dots,i_d)\in[n_1,\dots,d]} \ , \tag{61}$$

for every collection of permutations $\boldsymbol{\pi}_d \in \mathbb{S}_{\mathbf{n}_d}$. We say that $\mathbf{X}_{\mathbf{n}_d}$ is separately exchangeable if its distribution is.

For a symmetric array $\widetilde{\mathbf{X}}_{n^d}$, such that $n_1 = n_2 = \cdots = n_d = n$ and $\widetilde{X}_{i_1,\dots,i_d} = \widetilde{X}_{i_{\rho(1)},\dots,i_{\rho(d)}}$ for all $\rho \in \mathbb{S}_d$, the distribution of $\widetilde{\mathbf{X}}_{n^d}$ is *jointly exchangeable* if, for all $\pi \in \mathbb{S}_n$

$$(\widetilde{X}_{i_1,\dots,i_d})_{(i_1,\dots,i_d)\in[n]^d} \overset{\text{d}}{=} (\widetilde{X}_{\pi(i_1),\dots,\pi(i_d)})_{(i_1,\dots,i_d)\in[n]^d} \ . \tag{62}$$

The checkerboard function defined in Section 7.1 is extended to higher-dimensional arrays in the natural way, with input space $[0,1]^d$ partitioned into hyperrectangles of volume $\prod_{k=1}^d n_k^{-1}$. Denote the space of d-dimensional checkerboard functions by $\mathcal{F}_{\mathbf{n}_d}(\mathcal{X})$.

The first result concerns $\mathbb{S}_{\mathbf{n}_d}$-invariant conditional distributions. An equivalent result holds for jointly exchangeable d-arrays and $\mathbb{S}_n$-invariant conditional distributions, which we omit for brevity.

THEOREM 21. *Suppose $\mathbf{X}_{\mathbf{n}_d}$ is a separately exchangeable $\mathcal{X}$-valued array on index set $\mathbf{n}_d$, and $Y \in \mathcal{Y}$ is another random variable. Then Y is conditionally $\mathbb{S}_{\mathbf{n}_d}$-invariant given $\mathbf{X}_{\mathbf{n}_d}$ if and only if there is a measurable function $f : [0,1] \times \mathcal{F}_{\mathbf{n}_d}(\mathcal{X}) \to \mathcal{Y}$ such that*

$$(\mathbf{X}_{\mathbf{n}_d}, Y) \overset{\text{a.s.}}{=} (\mathbf{X}_{\mathbf{n}_d}, f(\eta, \mathbb{F}_{\mathbf{X}_{\mathbf{n}_d}})) \quad \textit{where} \quad \eta \sim \text{Unif}[0,1] \quad \textit{and} \quad \eta \perp\!\!\!\perp \mathbf{X}_{\mathbf{n}_d} \ . \tag{63}$$

PROOF. Clearly, the representation (63) satisfies $(\boldsymbol{\pi}_d \cdot \mathbf{X}_{\mathbf{n}_d}, Y) \overset{\text{d}}{=} (\mathbf{X}_{\mathbf{n}_d}, Y)$, for all $\boldsymbol{\pi}_d \in \mathbb{S}_{\mathbf{n}_d}$, which by Proposition 1 and the assumption that $\mathbf{X}_{\mathbf{n}_d}$ is separately exchangeable implies that Y is conditionally $\mathbb{S}_{\mathbf{n}_d}$-invariant given $\mathbf{X}_{\mathbf{n}_d}$. The converse is a easy consequence of the fact that the checkerboard function is a maximal invariant of $\mathbb{S}_{\mathbf{n}_d}$ acting on $\mathcal{X}^{n_1\times\cdots\times n_d}$ and Theorem 12. □

$\mathbb{S}_{\mathbf{n}_d}$*-equivariant conditional distributions..* Let $\mathbf{Y}_{\mathbf{n}_d}$ be a $\mathcal{Y}$-valued array indexed by $[\mathbf{n}_d]$. Theorem 22 below states that each element $Y_{i_1,\dots,i_d}$ can be represented as a function of a uniform random variable $\eta_{i_1,\dots,i_d} \perp\!\!\!\perp \mathbf{X}_{\mathbf{n}_d}$, and of a sequence of checkerboard functions: one for each sub-array of $\mathbf{X}_{\mathbf{n}_d}$ that contains $X_{i_1,\dots,i_d}$. As was the case for $d = 2$, we assume

$$Y_{i_1,\dots,i_d} \perp\!\!\!\perp_{\mathbf{X}_{\mathbf{n}_d}} (\mathbf{Y}_{\mathbf{n}_d} \setminus Y_{i_1,\dots,i_d}) \;, \quad \text{for each } (i_1,\dots,i_d) \in [\mathbf{n}_d] \;. \tag{64}$$

For convenience, denote by $\mathbf{i} := (i_1,\dots,i_d)$ an element of the index set $[\mathbf{n}_d]$. For each $k_1 \in [d]$, let $\mathbf{X}^{\downarrow}_{\mathbf{i}(\{k_1\})}$ be the $(d-1)$-dimensional sub-array of $\mathbf{X}_{\mathbf{n}_d}$ obtained by fixing the k_1th element of $\mathbf{i}$ and letting all other indices vary over their entire range. For example, for an array with $d = 3$, $\mathbf{X}^{\downarrow}_{\mathbf{i}(\{1\})}$ is the matrix extracted from $\mathbf{X}_{\mathbf{n}_d}$ by fixing i_1, $\mathbf{X}_{i_1,:,:}$.

Iterating, let $\mathbf{X}^{\downarrow}_{\mathbf{i}(\{k_1,\dots,k_p\})}$ be the $(d-p)$-dimensional sub-array of $\mathbf{X}_{\mathbf{n}_d}$ obtained by fixing elements $i_{k_1},\dots,i_{k_p}$ of $\mathbf{i}$ and letting all other indices vary. For each $p \in [d]$, denote by $[d]^p$ the collections of subsets of $[d]$ with exactly p elements; let $[d]^{(p)}$ be the collection of subsets of $[d]$ with p or fewer elements. Let the collection of $(d-p)$-dimensional sub-arrays containing $\mathbf{i}$

$$\mathbf{X}^{\downarrow(d-p)}_{\mathbf{i}} = \big(\mathbf{X}^{\downarrow}_{\mathbf{i}(s)}\big)_{s\in[d]^p} \;.$$

A d-dimensional version of the augmented checkerboard function (42) is needed. To that end, let $\mathbf{j} = (j_1,\dots,j_d) \in [\mathbf{n}_d]$ and define the index sequence $(\mathbf{i}(s),\mathbf{j}) \in [\mathbf{n}_d]$ as

$$(\mathbf{i}(s),\mathbf{j})_k = \begin{cases} i_k & \text{if } k \in s \\ j_k & \text{if } k \notin s \end{cases} \;, \quad \text{for } s \in [d]^{(p)} \text{ and } k \in [d] \;. \tag{65}$$

Let $\mathbf{u} = (u_1,\dots,u_d)$ be an element of $[0,1]^d$, and define the d-dimensional p-augmented checkerboard function,

$$\mathbb{F}^{(p)}_{\mathbf{i},\mathbf{X}_{\mathbf{n}_d}}(\mathbf{u}) = \big(X_{\mathbf{j}}, \big(X_{(\mathbf{i}(s),\mathbf{j})}\big)_{s\in[d]^{(p)}}\big) \;, \text{ for } u_1 \in I^{(1)}_{j_1},\dots,u_d \in I^{(d)}_{j_d} \;. \tag{66}$$

Denote by $\mathcal{F}^{(d)}_{\mathbf{n}_d}$ the space of all such functions with partition structure induced by $\mathbf{n}_d$.

The function returns the collection of elements from $\mathbf{X}_{\mathbf{n}_d}$ that correspond to $\mathbf{j}$ in each of the $(d-q)$-dimensional sub-arrays containing $\mathbf{i}$, $\mathbf{X}^{\downarrow(d-q)}_{\mathbf{i}}$, for $q = 0,\dots,p$.

Alternatively, observe that the function can be constructed by recursing through the $(d-q)$-dimensional 1-augmented checkerboard functions (there are $\binom{d}{q}$ of them for each q), for $q = 0, \dots, p-1$, and returning the first argument of each. This recursive structure captures the action of $\mathbb{S}_{\mathbf{n}_d}$ on $\mathbf{X}_{\mathbf{n}_d}$, and is at the heart of the following result, which gives a functional representation of $\mathbb{S}_{\mathbf{n}_d}$-equivariant conditional distributions.

THEOREM 22. *Suppose $\mathbf{X}_{\mathbf{n}_d}$ and $\mathbf{Y}_{\mathbf{n}_d}$ are $\mathcal{X}$-valued arrays indexed by $[\mathbf{n}_d]$, and that $\mathbf{X}_{\mathbf{n}_d}$ is separately exchangeable. Assume that the elements of $\mathbf{Y}_{\mathbf{n}_d}$ are conditionally independent given $\mathbf{X}_{\mathbf{n}_d}$. Then $\mathbf{Y}_{\mathbf{n}_d}$ is conditionally $\mathbb{S}_{\mathbf{n}_d}$-equivariant given $\mathbf{X}_{\mathbf{n}_d}$ if and only if*

$$(\mathbf{X}_{\mathbf{n}_d}, \mathbf{Y}_{\mathbf{n}_d}) \overset{\text{a.s.}}{=} \left(\mathbf{X}_{\mathbf{n}_d}, \big(f(\eta_{\mathbf{i}}, X_{\mathbf{i}}, \mathbb{F}^{(d)}_{\mathbf{i},\mathbf{X}_{\mathbf{n}_d}})_{\mathbf{i}\in[\mathbf{n}_d]} \big) \right), \tag{67}$$

for some measurable function $f : [0,1] \times \mathcal{X} \times \mathcal{F}^{(d)}_{\mathbf{n}_d} \to \mathcal{Y}$ and i.i.d. uniform random variables $(\eta_{\mathbf{i}})_{\mathbf{i}\in[\mathbf{n}_d]} \perp\!\!\!\perp \mathbf{X}_{\mathbf{n}_d}$.

PROOF. First, assume that $\mathbf{Y}_{\mathbf{n}_d}$ is conditionally $\mathbb{S}_{\mathbf{n}_d}$-equivariant given $\mathbf{X}_{\mathbf{n}_d}$. By assumption, $\mathbf{X}_{\mathbf{n}_d}$ is separately exchangeable, so by Proposition 1, $\boldsymbol{\pi}_d \cdot (\mathbf{X}_{\mathbf{n}_d}, \mathbf{Y}_{\mathbf{n}_d}) \overset{\text{d}}{=} (\mathbf{X}_{\mathbf{n}_d}, \mathbf{Y}_{\mathbf{n}_d})$ for all $\boldsymbol{\pi}_d \in \mathbb{S}_{\mathbf{n}_d}$. Fix $\mathbf{i} \in [\mathbf{n}_d]$, and let $\mathbb{S}^{(\mathbf{i})}_{\mathbf{n}_d} \subset \mathbb{S}_{\mathbf{n}_d}$ be the stabilizer of $\mathbf{i}$. Observe that each $\boldsymbol{\pi}^{(\mathbf{i})}_d \in \mathbb{S}^{(\mathbf{i})}_{\mathbf{n}_d}$ fixes $X_{\mathbf{i}}$ and $Y_{\mathbf{i}}$.

In analogy to the fixed -row and -column structure for the $d = 2$ case, any $\boldsymbol{\pi}^{(\mathbf{i})}_d \in \mathbb{S}^{\mathbf{i}}_{\mathbf{n}_d}$ results in sub-arrays of $\mathbf{X}_{\mathbf{n}_d}$ in which the elements my be rearranged, but the sub-array maintains its position within $\mathbf{X}_{\mathbf{n}_d}$. For example, consider $d = 3$. Each of the two-dimensional arrays $\mathbf{X}_{i_1,:,:}$, $\mathbf{X}_{:,i_2,:}$, and $\mathbf{X}_{:,:,i_3}$ may have their elements rearranged, but they will remain the two-dimensional sub-arrays that intersect at $\mathbf{i}$. Likewise for the one-dimensional sub-arrays $\mathbf{X}_{i_1,i_2,:}$, $\mathbf{X}_{i_1,:,i_3}$, and $\mathbf{X}_{:,i_2,i_3}$.

Recall that $\mathbf{X}^{\downarrow}_{\mathbf{i}(\{k_1,\dots,k_p\})}$ is the $(d-p)$-dimensional sub-array of $\mathbf{X}_{\mathbf{n}_d}$ obtained by fixing elements $i_{k_1}, \dots, i_{k_p}$ of $\mathbf{i}$ and letting the other indices vary, and $\mathbf{X}^{\downarrow(d-p)}_{\mathbf{i}}$ is the collection of these sub-arrays. For $p = 1$, $\mathbf{X}^{\downarrow(d-1)}_{\mathbf{i}} = (\mathbf{X}^{\downarrow}_{\mathbf{i}(\{k_1\})})_{k_1\in[d]}$ is the collection of $(d-1)$-dimensional sub-arrays containing the element $X_{\mathbf{i}}$. Denote by $\mathbf{X}^{\uparrow(d)}_{\mathbf{i}} := \mathbf{X}_{\mathbf{n}_d} \backslash \mathbf{X}^{\downarrow(d-1)}_{\mathbf{i}}$ the d-dimensional array that remains after extracting $\mathbf{X}^{\downarrow(d-1)}_{\mathbf{i}}$ from $\mathbf{X}_{\mathbf{n}_d}$. Call $\mathbf{X}^{\uparrow(d)}_{\mathbf{i}}$ a d-dimensional *remainder array.* For example, with $d = 3$, the remainder array $\mathbf{X}^{\uparrow(d)}_{\mathbf{i}}$ consists of $\mathbf{X}_{\mathbf{n}_d}$ with the three two-dimensional arrays (matrices) that contain $X_{\mathbf{i}}$ removed.

Continue recursively and construct a remainder array $\mathbf{X}^{\uparrow(d-p)}_{\mathbf{i}(s)}$, $s \in [d]^p$, from each sub-array in the collection $\mathbf{X}^{\downarrow(d-p)}_{\mathbf{i}} = (\mathbf{X}^{\downarrow}_{\mathbf{i}(s)})_{s\in[d]^p}$. The recursive structure of the collection of remainder arrays $\big((\mathbf{X}^{\uparrow(d-p)}_{\mathbf{i}(s)})_{s\in[d]^p}\big)_{0\leq p\leq d-1}$ is captured by the array

$\mathbf{Z}^{(i)}$, with entries

$$[\mathbf{Z}^{(\mathbf{i})}]_{\mathbf{j}} = \big(X_{\mathbf{j}}, (X_{(\mathbf{i}(s),\mathbf{j})})_{s\in[d]^{(d)}}\big) \; , \tag{68}$$

where $(\mathbf{i}(s),\mathbf{j})$ is as in (65). Define the action of a collection of permutations $\boldsymbol{\pi}'_d \in \mathbb{S}_{\mathbf{n}_d-\mathbf{1}_d}$ on $\mathbf{Z}^{(\mathbf{i})}$ to be such that, with $\mathbf{j}_\pi = (\pi_1(j_1),\dots,\pi_d(j_d))$,

$$[\boldsymbol{\pi}'_d \cdot \mathbf{Z}^{(\mathbf{i})}]_{\mathbf{j}} = \big(X_{\mathbf{j}_\pi}, (X_{(\mathbf{i}(s),\mathbf{j}_\pi)})_{s\in[d]^{(d)}}\big) \; . \tag{69}$$

$\mathbf{Z}^{(\mathbf{i})}$ inherits the exchangeability of $\mathbf{X}_{\mathbf{n}_d}$, so that marginally for $Y_{\mathbf{i}}$,

$$(\boldsymbol{\pi}_d \cdot \mathbf{Z}^{(\mathbf{i})}, (X_{\mathbf{i}}, Y_{\mathbf{i}})) \overset{\mathrm{d}}{=} (\mathbf{Z}^{(\mathbf{i})}, (X_{\mathbf{i}}, Y_{\mathbf{i}})) \quad \text{for all } \boldsymbol{\pi}_d \in \mathbb{S}_{\mathbf{n}_d-\mathbf{1}_d} \; .$$

which implies $(X_{\mathbf{i}}, Y_{\mathbf{i}}) \perp\!\!\!\perp_{\mathbb{F}_{\mathbf{Z}^{(\mathbf{i})}}} \mathbf{Z}^{(i)}$. Conditioning on $X_{\mathbf{i}}$ and $\mathbb{F}_{\mathbf{Z}^{(\mathbf{i})}}$ is the same as conditioning on $X_{\mathbf{i}}$ and the d-dimensional d-augmented checkerboard function $\mathbb{F}^{(d)}_{\mathbf{i},\mathbf{X}_{\mathbf{n}_d}}$ defined in (66), implying that

$$Y_i \perp\!\!\!\perp_{(X_{\mathbf{i}}, \mathbb{F}^{(d)}_{\mathbf{i},\mathbf{X}_{\mathbf{n}_d}})} \mathbf{X}_{\mathbf{n}_d} \; .$$

By Lemma 3, there is a measurable function $f_{\mathbf{i}} : [0,1] \times \mathcal{X} \times \mathcal{F}^{(d)}_{\mathbf{n}_d} \to \mathcal{Y}$ such that

$$Y_{\mathbf{i}} = f_{\mathbf{i}}(\eta_{\mathbf{i}}, X_{\mathbf{i}}, \mathbb{F}^{(d)}_{\mathbf{i},\mathbf{X}_{\mathbf{n}_d}}) \; ,$$

for a uniform random variable $\eta_{\mathbf{i}} \perp\!\!\!\perp \mathbf{X}_{\mathbf{n}_d}$. This is true for all $\mathbf{i} \in [\mathbf{n}_d]$; by equivariance the same $f_{\mathbf{i}} = f$ must work for every $\mathbf{i}$. Furthermore, by assumption the elements of $\mathbf{Y}_{\mathbf{n}_d}$ are mutually conditionally independent given $\mathbf{X}_{\mathbf{n}_d}$, and therefore by the chain rule for conditional independence (Kallenberg, 2002, Prop. 6.8), the joint identity (67) holds.

The converse is straightforward to verify.

□

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Aldous, D. J. (1985). Exchangeability and Related Topics. In *École d'Été de Probabilités de Saint-Flour XIII - 1983*, (P. L. Hennequin, ed.). *Lecture Notes in Mathematics* **1117** 1–198. Springer.

Andersson, S. (1982). Distributions of Maximal Invariants Using Quotient Measures. *The Annals of Statistics* **10** 955–961.

Arora, S., Ge, R., Neyshabur, B. and Zhang, Y. (2018). Stronger Generalization Bounds for Deep Nets via a Compression Approach. In *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 254–263. PMLR, Stockholmsmässan, Stockholm Sweden.

Austin, T. (2015). Exchangeable random measures. *Ann. Inst. H. Poincaré Probab. Statist.* **51** 842–861.

Bacallado, S., Favaro, S. and Trippa, L. (2013). Bayesian nonparametric analysis of reversible Markov chains. *Ann. Statist.* **41** 870–896.

Bahadur, R. R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25** 423–462.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics* **1** 353—355.

Borgs, C. and Chayes, J. (2017). Graphons: A Nonparametric Method to Model, Estimate, and Design Algorithms for Massive Networks. *Proceedings of the 2017 ACM Conference on Economics and Computation - EC '17*.

Bruna, J., Zaremba, W., Szlam, A. and LeCun, Y. (2014). Spectral Networks and Locally Connected Networks on Graphs. In *Proceedings of the 2nd International Conference on Learning Representations*.

Cai, D., Campbell, T. and Broderick, T. (2016). Edge-exchangeable graphs and sparsity. In *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett, eds.) 4249–4257. Curran Associates, Inc.

Caron, F. and Fox, E. B. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 1–44.

Chan, J., Perrone, V., Spence, J. P., Jenkins, P. A., Mathieson, S. and Song, Y. S. (2018). A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks.

Cohen, T. S. and Welling, M. (2016). Group Equivariant Convolutional Networks. In *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 2990–2999. PMLR, New York, New York, USA.

Cohen, T. S. and Welling, M. (2017). Steerable CNNs. In *International Conference on Learning Representations*.

Cohen, T. S., Geiger, M., Köhler, J. and Welling, M. (2018). Spherical CNNs. In *International Conference on Learning Representations*.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Crane, H. and Dempsey, W. (2018). Edge Exchangeable Models for Interaction Networks. *Journal of the American Statistical Association* **113** 1311–1326.

de Finetti, B. (1930). Fuzione caratteristica di un fenomeno aleatorio. *Mem. R. Acc. Lincei* **6** 86–133.

Diaconis, P. (1977). Finite forms of de Finetti's theorem on exchangeability. *Synthese* **36** 271–281.

Diaconis, P. and Freedman, D. (1980a). Finite Exchangeable Sequences. *The Annals of Probability* **8** 745–764.

Diaconis, P. and Freedman, D. (1980b). De Finetti's Theorem for Markov Chains. *Ann. Probab.* **8** 115–130.

Diaconis, P. and Freedman, D. (1987). A dozen de Finetti-style results in search of a theory. *Ann. Inst. Henri Poincaré, Probab. Stat.* **23** 397–423.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A. and Adams, R. P. (2015). Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds.) 2224–2232. Curran Associates, Inc.

Eaton, M. L. (1989). *Group invariance in applications in statistics. Regional Conference Series in*

Probability and Statistics **Volume 1**. Institute of Mathematical Statistics and American Statistical Association, Haywood, CA and Alexandria, VA.

Edwards, H. and Storkey, A. (2017). Towards a Neural Statistician. In *International Conference on Learning Representations*.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **222** 309—368.

Fortini, S., Ladelli, L. and Regazzini, E. (2000). Exchangeability, Predictive Distributions and Parametric Models. *SankhyÄ: The Indian Journal of Statistics* **62** 86–109.

Freedman, D. A. (1962). Invariants Under Mixing which Generalize de Finetti's Theorem. *The Annals of Mathematical Statistics* **33** 916–923.

Freedman, D. A. (1963). Invariants Under Mixing Which Generalize de Finetti's Theorem: Continuous Time Parameter. *The Annals of Mathematical Statistics* **34** 1194–1216.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D. and Eslami, S. M. A. (2018). Conditional Neural Processes. In *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 1704–1713. PMLR, Stockholmsmässan, Stockholm Sweden.

Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18* (Y. Weiss, B. Schölkopf and J. C. Platt, eds.) 475–482. MIT Press.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. and Dahl, G. E. (2017). Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.). *Proceedings of Machine Learning Research* **70** 1263–1272. PMLR.

Giri, N. C. (1996). *Group Invariance in Statistical Inference*. World Scientific.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, eds.) 2672–2680. Curran Associates, Inc.

The GAP Group (2018). GAP – Groups, Algorithms, and Programming, Version 4.10.0.

Gutmann, M. U. and Corander, J. (2016). Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. *Journal of Machine Learning Research* **17** 1–47.

Hall, W. J., Wijsman, R. A. and Ghosh, J. K. (1965). The Relationship Between Sufficiency and Invariance with Applications in Sequential Analysis. *The Annals of Mathematical Statistics* **36** 575–614.

Halmos, P. R. and Savage, L. J. (1949). Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics. *The Annals of Mathematical Statistics* **20** 225—241.

Hamaguchi, T., Oiwa, H., Shimbo, M. and Matsumoto, Y. (2017). Knowledge Transfer for Out-of-knowledge-base Entities: A Graph Neural Network Approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence. IJCAI'17* 1802–1808. AAAI Press.

Hartford, J., Graham, D., Leyton-Brown, K. and Ravanbakhsh, S. (2018). Deep Models of Interactions Across Sets. In *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 1914–1923. PMLR.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778.

Herzig, R., Raboh, M., Chechik, G., Berant, J. and Globerson, A. (2018). Mapping Images to Scene Graphs with Permutation-Invariant Structured Prediction. In *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) 7211–7221. Curran Associates, Inc.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* **2** 359–366.

Kallenberg, O. (2002). *Foundations of Modern Probability*, 2nd ed. Springer-Verlag New York.

Kallenberg, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer.

Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Springer International Publishing.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V. and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **30** 595–608.

Kim, H., Mnih, A., Schwarz, J., Garnelo, M., Eslami, A., Rosenbaum, D., Vinyals, O. and Teh, Y. W. (2019). Attentive Neural Processes. In *International Conference on Learning Representations*. To appear.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Kondor, R. and Trivedi, S. (2018). 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* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 2747–2755. PMLR, Stockholmsmässan, Stockholm Sweden.

Korshunova, I., Degrave, J., Huszar, F., Gal, Y., Gretton, A. and Dambre, J. (2018). BRUNO: A Deep Recurrent Model for Exchangeable Data. In *Advances in Neural Information Processing Systems 31* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) 7190–7198. Curran Associates, Inc.

Lauritzen, S. L. (1974a). Sufficiency, Prediction and Extreme Models. *Scandinavian Journal of Statistics* **1** 128–134.

Lauritzen, S. L. (1974b). On the Interrelationships among Sufficiency, Total Sufficiency and Some Related Concepts Technical Report No. 8, Institute of Mathematical Statistics, University of Copenhagen.

Lauritzen, S. L. (1984). Extreme Point Models in Statistics [with Discussion and Reply]. *Scandinavian Journal of Statistics* **11** 65–91.

Lauritzen, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics*. *Lecture Notes in Statistics*. Springer.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. and Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* **1** 541–551.

Lee, J., Lee, Y., Kim, J., Kosiorek, A. R., Choi, S. and Teh, Y. W. (2018). Set Transformer. *ArXiv e-prints*.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. *Sprinter Texts in Statistics*. Springer-Verlag New York.

Lovász, L. (2012). *Large Networks and Graph Limits*. American Mathematical Society.

Mahmoud, H. (2008). *Pólya Urn Models*. *Chapman & Hall/CRC Texts in Statistical Science*. Chapman & Hall/CRC.

Minsky, M. L. and Papert, S. A. (1988). *Perceptrons: Expanded Edition*. MIT Press, Cambridge, MA, USA.

Murphy, R. L., Srinivasan, B., Rao, V. and Ribeiro, B. (2019). Janossy Pooling: Learning Deep Permutation-Invariant Functions for Variable-Size Inputs. In *International Conference on Learning Representations*. To appear.

Niepert, M., Ahmed, M. and Kutzkov, K. (2016). Learning Convolutional Neural Networks for Graphs. In *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.). *Proceedings of Machine Learning Research* **48** 2014–2023. PMLR, New York, New York, USA.

Niepert, M. and Van den Broeck, G. (2014). Tractability Through Exchangeability: A New Perspective on Efficient Probabilistic Inference. In *Proceedings of the Twenty-Eighth AAAI Con-*

ference on Artificial Intelligence. AAAI'14 2467–2475. AAAI Press.

Orbanz, P. and Roy, D. M. (2015). Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37** 437-461.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS 2017 Workshop on Automatic Differentiation*.

Ravanbakhsh, S., Schneider, J. and Póczos, B. (2017). Equivariance Through Parameter-Sharing. In *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.). *Proceedings of Machine Learning Research* **70** 2892–2901. PMLR.

Rezende, D. J., Mohamed, S. and Wierstra, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.). *Proceedings of Machine Learning Research* **32** 1278–1286. PMLR, Bejing, China.

Rotman, J. (1995). *An Introduction to the Theory of Groups*, 4 ed. *Graduate Texts in Mathematics* **148**. Springer-Verlag New York.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. and Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks* **20** 61–80.

Schervish, M. J. (1995). *Theory of Statistics*. Springer-Verlag New York.

Schindler, W. (2003). *Measures with Symmetry Properties*. *Lecture Notes in Mathematics* **1808**. Springer-Verlag Berlin Heidelberg, Bernlin.

Shawe-Taylor, J. (1989). Building symmetries into feedforward networks. In *1989 First IEE International Conference on Artificial Neural Networks, (Conf. Publ. No. 313)* 158–162.

Shawe-Taylor, J. (1991). Threshold Network Learning in the Presence of Equivalences. In *Advances in Neural Information Processing Systems 4* (J. E. Moody, S. J. Hanson and R. P. Lippmann, eds.) 879–886. Morgan-Kaufmann.

Shawe-Taylor, J. (1993). Symmetries and discriminability in feedforward network architectures. *IEEE Transactions on Neural Networks* **4** 816–826.

Shawe-Taylor, J. (1995). Sample Sizes for Threshold Networks with Equivalences. *Information and Computation* **118** 65 - 72.

Skibinsky, M. (1967). Adequate Subfields and Sufficiency. *The Annals of Mathematical Statistics* **38** 155–161.

Smolensky, P. (1987). *Information Processing in Dynamical Systems: Foundations of Harmony Theory* In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* 194–281. MIT Press.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian Processes using Pseudo-inputs. In *Advances in Neural Information Processing Systems 18* (Y. Weiss, B. Schölkopf and J. C. Platt, eds.) 1257–1264. MIT Press.

Speed, T. P. (1978). A Factorisation Theorem for Adequate Statistics. *Australian Journal of Statistics* **20** 240–249.

Sukhbaatar, S., szlam, a., Weston, J. and Fergus, R. (2015). End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds.) 2440–2448. Curran Associates, Inc.

Theano Development Team (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints* **abs/1605.02688**.

van de Meent, J.-W., Paige, B., Yang, H. and Wood, F. (2018). An Introduction to Probabilistic Programming.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 5998–6008. Curran Associates, Inc.

Welling, M., Rosen-zvi, M. and Hinton, G. E. (2005). Exponential Family Harmoniums with an Application to Information Retrieval. In *Advances in Neural Information Processing Systems 17* (L. K. Saul, Y. Weiss and L. Bottou, eds.) 1481–1488. MIT Press.

WIJSMAN, R. A. (1990). *Invariant measures on groups and their use in statistics*. *Lecture Notes–Monograph Series* **14**. Institute of Mathematical Statistics, Hayward, CA.

WILLIAMSON, S. A. (2016). Nonparametric Network Models for Link Prediction. *Journal of Machine Learning Research* **17** 1-21.

WOLFRAM RESEARCH, INC. (2018). Mathematica, Version 11.3. Champaign, IL.

WOOD, J. and SHAWE-TAYLOR, J. (1996). Representation theory and invariant neural networks. *Discrete Applied Mathematics* **69** 33–60.

XU, K., HU, W., LESKOVEC, J. and JEGELKA, S. (2019). How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*. To appear.

YU, F. and KOLTUN, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations*.

ZABELL, S. L. (2005). *The Rule of Succession*. In *Symmetry and its Discontents: Essays on the History of Inductive Probability*. *Cambridge Studies in Probability, Induction and Decision Theory* 38–73. Cambridge University Press.

ZAHEER, M., KOTTUR, S., RAVANBAKHSH, S., POCZOS, B., SALAKHUTDINOV, R. R. and SMOLA, A. J. (2017). Deep Sets. In *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 3391–3401. Curran Associates, Inc.

ZHOU, J., CUI, G., ZHANG, Z., YANG, C., LIU, Z. and SUN, M. (2018). Graph Neural Networks: A Review of Methods and Applications.

ZHOU, W., VEITCH, V., AUSTERN, M., ADAMS, R. P. and ORBANZ, P. (2019). Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach. In *International Conference on Learning Representations*. To appear.

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
OXFORD OX1 3LB
UK
E-MAIL: benjamin.bloem-reddy@stats.ox.ac.uk; y.w.teh@stats.ox.ac.uk