Neural network models of exchangeable sequences

Benjamin Bloem-Reddy   Yee Whye Teh
Department of Statistics
University of Oxford
Oxford, UK
{benjamin.bloem-reddy,y.w.teh}@stats.ox.ac.uk

Abstract

Much recent work has been devoted to determining neural network architectures that are invariant (and equivariant) under the action of a group. We treat the neural network input and output as random variables, and consider permutation-invariance from the probabilistic perspective of exchangeability. What neural network architectures are possible when the input is an exchangeable sequence and the distribution of the output is assumed to be invariant under permutations of the input? We obtain two general representation results that together determine the architecture of an invariant stochastic feed-forward network. We also recover some recent results as special deterministic cases.

We consider the problem of using a neural network to model the distribution of a random variable $Y$ conditioned on an exchangeable sequence $X_{[n]} := (X_1, \ldots, X_n)$. $Y$ might represent a latent encoding of $X_{[n]}$, a prediction $X_{n+1}$, or a label relevant to $X_{[n]}$. We do not consider any specific problem; rather, we focus on how distributional symmetry affects possible neural network architectures. In particular, we require that the conditional distribution $P_{Y|X_{[n]}}$ of $Y$ given $X_{[n]}$ be invariant to permutations of $X_{[n]}$ and ask,

How does requiring permutation invariance constrain the neural network architecture?

A similar problem was considered by Zaheer et al. [10]; those authors posited a deterministic model $Y = f(X_{[n]})$ for some function $f$, and considered the class of permutation-invariant functions. The approach here is probabilistic, and draws upon tools from probability and statistics to answer the question above in a way that is both theoretically general and practically useful. In particular, we establish the functional representation of $Y$ when its conditional distribution $P_{Y|X_{[n]}}$ is invariant to permutations of $X_{[n]}$. In doing so, we characterize the functional structure of stochastic neural networks such that the distribution of the output is invariant under permutations of the input: all such networks must be composed of stochastic equivariant modules and stochastic invariant modules. This echoes the results for deterministic networks in [10] but is substantially more general: the requirement of distributionally invariant output is weaker than deterministic invariance, and our results recover those of [10] in the special case that the conditional distribution of $Y$ given $X_{[n]}$ is a point mass at $f(X_{[n]})$. Our results also allow for more flexible network architectures than the standard multi-layer perceptron’s activation of linear combinations.

The rest of the paper is structured as follows. In Section 1, we define the basic properties of interest. The main results are stated in Section 2, along with an illustration of their practical implications. All technical arguments and proofs are given in Appendix A.
1 Preliminaries

We begin with the necessary definitions. Let $X_{[n]} := (X_1, \ldots, X_n)$ be a random $\mathcal{X}$-valued sequence.\footnote{We assume that all random variables take values in standard Borel spaces. A measurable space $\mathcal{X}$ is a standard Borel space if there is a measurable bijection between $\mathcal{X}$ and a Borel subset of $[0, 1]$. See [3, Ch. 1].} Denote the Borel $\sigma$-algebra of $\mathcal{X}^n$ by $\mathcal{B}(\mathcal{X}^n)$ and the set of all permutations of $n$ elements by $\mathcal{S}_n$. A probability distribution $P$ on $\mathcal{X}^n$ is \textit{finitely exchangeable} if for every $B_1, \ldots, B_n \in \mathcal{B}(\mathcal{X}^n)$,

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = P[X_{\pi(1)} \in B_1, \ldots, X_{\pi(n)} \in B_n] \quad \text{for each } \pi \in \mathcal{S}_n. \quad (1)$$

That is, $P$ is invariant to all permutations of the elements of $X_{[n]}$. We write $X_{[n]} \overset{d}= \pi \cdot X_{[n]} = X_{\pi:[n]}$ to denote this equality in distribution. More commonly encountered is \textit{infinite exchangeability}, which requires probabilistic coherence between the finite-dimensional distributions of an infinite sequence $X_{[\infty]}$. Only finitely exchangeable sequences of fixed length are considered here; for convenience we say that a sequence $X_{[n]}$ is exchangeable if its distribution is finitely exchangeable.

Let $Y \in \mathcal{Y}$ be some random variable of interest, to be predicted from $X_{[n]}$. Permutation invariance of the conditional distribution $P_{Y \mid X_{[n]}}$ of $Y$ given $X_{[n]}$ is defined as follows.

**Definition 1.** The conditional distribution $P_{Y \mid X_{[n]}}$ of $Y$ given $X_{[n]}$ is \textit{$\mathcal{S}_n$-invariant} if for all $B \in \mathcal{B}(\mathcal{Y})$ and $\pi \in \mathcal{S}_n$,

$$P_{Y \mid X_{[n]}}[Y \in B \mid \pi \cdot X_{[n]}] = P_{Y \mid X_{[n]}}[Y \in B \mid X_{[n]}] \quad \text{a.s.-} P_{X_{[n]}}. \quad (2)$$

Together with exchangeability of $P_{X_{[n]}}$, conditional invariance implies the joint equality

$$(\pi \cdot X_{[n]}, Y) \overset{d}= (X_{[n]}, Y) \quad \text{for all } \pi \in \mathcal{S}_n. \quad (3)$$

In the neural networks literature, \textit{equivariance} plays a key role in the hidden layers of invariant deterministic neural networks $[2, 8, 6]$. As we show in Section 2, the following adaptation to random equivariance plays the same role in $\mathcal{S}_n$-invariant stochastic neural networks.

**Definition 2.** Let $Y_{[n]} \in \mathcal{Y}^n$ be a random $\mathcal{Y}$-valued sequence. The conditional distribution $P_{Y_{[n]} \mid X_{[n]}}$ of $Y_{[n]}$ given $X_{[n]}$ is \textit{$\mathcal{S}_n$-equivariant} if for all $B \in \mathcal{B}(\mathcal{Y}^n)$ and $\pi \in \mathcal{S}_n$,

$$P_{Y_{[n]} \mid X_{[n]}}[\pi \cdot Y_{[n]} \in B \mid \pi \cdot X_{[n]}] = P_{Y_{[n]} \mid X_{[n]}}[Y_{[n]} \in B \mid X_{[n]}] \quad \text{a.s.-} P_{X_{[n]}}. \quad (4)$$

Together with exchangeability of $P_{X_{[n]}}$, conditional invariance implies the joint equality

$$(\pi \cdot X_{[n]}, \pi \cdot Y_{[n]}) \overset{d}= (X_{[n]}, Y_{[n]}) \quad \text{for all } \pi \in \mathcal{S}_n. \quad (5)$$

2 Main results

In this section, we state our main results. Proofs are given in Appendix A. Both of the main results express $Y$ in terms of a stochastic function of the empirical measure of $X_{[n]}$, which is defined as

$$\mathbb{M}_{X_{[n]}}(\cdot) = \sum_{i=1}^{n} \delta_{X_i}(\cdot), \quad (6)$$

(see also [3, Ch. 1]).\footnote{$\delta_{X_i}(A) = 1$ if $X_i \in A$ and 0 otherwise, for any measurable set $A$.} We denote $X_{[n]}$ by $\delta_{X_{[n]}}$.

The first main result characterizes the functional relationship between $X_{[n]}$ and $Y$, for any $Y$ such that $P_{Y \mid X_{[n]}}$ is $\mathcal{S}_n$-invariant.

**Theorem 3** (Invariant representation). \textit{For fixed $n \in \mathbb{N}$, suppose $X_{[n]} \in \mathcal{X}^n$, is an exchangeable sequence and $Y \in \mathcal{Y}$ is another random variable. Then $P_{Y \mid X_{[n]}}$ is $\mathcal{S}_n$-invariant if and only if there is a measurable function $f : [0, 1] \times \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{Y}$ such that

$$f(X_{[n]}, \mathbb{M}_{X_{[n]}}) \overset{d}= (X_{[n]}, \mathbb{M}_{X_{[n]}}(\eta, \mathbb{M}_{X_{[n]}}(\eta))) \quad \text{where } \eta \sim \text{Unif}([0, 1]) \quad \text{and } \eta \perp X_{[n]}. \quad (7)$$}
Theorem 4 (Equivariant representation). Let \( X_{[n]} \in \mathcal{X}^n \) be an exchangeable sequence and \( Y_{[n]} \in \mathcal{Y}^n \) another random sequence, and assume that \( Y_i \perp \!\!\!\!\perp X_{[n]} \setminus \{Y_i\} \) for all \( i \in [n] \). Then \( Y_{[n]} \) is \( S_n \)-equivariant if and only if there is a measurable function \( g : [0, 1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \to \mathcal{Y} \) such that

\[
(X_{[n]}, Y_{[n]}) \overset{\text{d}}{=} (X_{[n]}, (g(\eta_i, X_i, M_X))_{i \in [n]}) \quad \text{where} \quad \eta_i \overset{\text{ind}}{\sim} \text{Unif}[0, 1] \quad \text{and} \quad \eta_i \perp \!\!\!\!\perp X_{[n]}.
\]

As before, in the deterministic case the noise variables \( \eta_i \) are ignored. For example, Lemma 3 in Zaheer et al. [10] is precisely that, with a particular functional form of \( g \) (activation of a linear combination).

Finally, we establish that: (i) stochastic equivariance is transitive, which means that the repeated composition of equivariant functions is also equivariant; and (ii) stochastic equivariance preserves invariance.

Proposition 5 (Transitivity of equivariance and invariance).

(i) Let \( X_{[n]}, Y_{[n]}, Z_{[n]} \) be random sequences of length \( n \) such that \( \pi \cdot (X_{[n]}, Y_{[n]}) \overset{\text{d}}{=} (X_{[n]}, Y_{[n]}) \) and \( \pi \cdot (Y_{[n]}, Z_{[n]}) \overset{\text{d}}{=} (Y_{[n]}, Z_{[n]}) \) for all \( \pi \in S_n \), and \( X_{[n]} \perp \!\!\!\!\perp Y_{[n]} \perp \!\!\!\!\perp Z_{[n]} \). Then \( \pi \cdot (X_{[n]}, Y_{[n]}, Z_{[n]}) \overset{\text{d}}{=} (X_{[n]}, Y_{[n]}, Z_{[n]}) \) for all \( \pi \in S_n \).

(ii) Let \( X_{[n]}, Y_{[n]} \) be random sequences of length \( n \) and \( Z \) another random variable. Suppose that \( \pi \cdot (X_{[n]}, Y_{[n]}) \overset{\text{d}}{=} (X_{[n]}, Y_{[n]}) \) and \( (\pi \cdot Y_{[n]}, Z) \overset{\text{d}}{=} (Y_{[n]}, Z) \) for all \( \pi \in S_n \). Then \( (\pi \cdot X_{[n]}, \pi \cdot Y_{[n]}, Z) \overset{\text{d}}{=} (X_{[n]}, Y_{[n]}, Z) \) for all \( \pi \in S_n \).

Together, (i) and (ii) suggest how to construct a stochastic or deterministic neural network that is \( S_n \)-invariant.

2.1 How to build a permutation-invariant neural network

In practice, Proposition 5 indicates that an invariant neural network can be constructed from a sequence of equivariant function modules feeding into an invariant module. A computational diagram of such an architecture is displayed below. The invariant module with structure determined by Theorem 3 is shown on the left. In practice, one might learn a function \( \Phi : \mathcal{X} \to \mathbb{R} \) parameterized by a neural network, applied to each element \( X_i \), then pooled into the empirical measure. \( f \), the function of noise \( \eta \) and of the empirical measure produces the invariant \( Y \). The equivariant architecture determined by Theorem 4 is displayed in the middle. Note that there is an invariant module within the equivariant module; however, it is purely deterministic (i.e., there is no \( \eta \) variable), as indicated by the "\( \ast \). The full invariant architecture (with two equivariant layers shown) is displayed on the right.
References


A Proofs of the main results

A.1 Noise outsourcing

The main technical tool that allows us to establish functional representations for the conditional distributions of interest is a standard result from measure-theoretic probability known as transfer [3], also called noise outsourcing [1]. For $P_{Y|X_{[n]}}$, it expresses $Y$ as a function of $X_{[n]}$ and “outsourced” independent noise. This type of functional representation can be viewed as a measure-theoretic version of the so-called reparameterization trick [5, 9] for random variables taking values in fairly general spaces (not just $\mathbb{R}$).

We state the result in terms of generic random variables $W$ and $Z$ because it holds for more general objects than $X_{[n]}$ and $Y$ as we have defined them above.

**Proposition 6** (Noise outsourcing with a d-separating statistic). Let $W \in W$ and $Z \in Z$ be random variables with joint distribution $P_{W,Z}$. Let $S : W \to S$ be a measurable map. Then $S(W)$ d-separates $W$ and $Z$ if and only if there is a measurable function $f : [0,1] \times S \to Z$ such that

$$
(W, Z) \overset{d}{=} (W, f(\eta, S(W))) \quad \text{where} \quad \eta \sim \text{Unif}[0,1] \quad \text{and} \quad \eta \perp \! \! \! \perp W.
$$

(7)

In particular, $Z \overset{d}{=} f(\eta, S(W))$. 


Note that in general \( f \) is measurable but need not be differentiable, although for modelling purposes one can limit oneself to a differentiable function belonging to a tractable class (e.g. parameterized by a neural network). Note also that the identity map \( S(\xi) = \xi \) \( d \)-separates \( \xi \) and \( \zeta \) trivially, so that we have \( \zeta = f(\eta, \xi) \). This is a standard fact from measure-theoretic probability [e.g., 3, Thm. 6.10].

### A.2 d-separation with the empirical measure

The proofs of the main results rely on establishing various conditional independence relationships. To this end, borrowing from the graphical models literature [7], we say that a statistic \( Y \) \( d \)-separates \( X[n] \) from \( Y \) if \( Y \) is conditionally independent from \( X[n] \), given \( S(X[n]) \). We denote this by \( Y \perp \perp S(X[n]), X[n] \).

A standard fact (see Proposition 7 below) is that a distribution \( P \) on \( \mathcal{X}^n \) is exchangeable if and only if the conditional distribution \( P(X[n] \mid \mathbb{M}_{X[n]} = m) \) is the uniform distribution on all sequences \((x_1, \ldots, x_n)\) that have empirical measure \( m \); that is, it is the uniform distribution on all sequences that can be obtained by applying a permutation to \( X[n] \). Furthermore, the empirical measure characterizes the structure of prediction from finite exchangeable sequences: it is \( d \)-separating for any random variable \( Y \) satisfying \((\pi \cdot X[n], Y) \overset{d}{=} (X[n], Y)\), for all \( \pi \in \mathbb{S}_n \) (which amounts to saying that \( X[n] \) is exchangeable and the conditional distribution of \( Y \) given \( X[n] \) is invariant under permutations of \( X[n] \)). In order to state the result, define the urn law of \( X[n] \) as

\[
\mathbb{U}_{X[n]}(\cdot) = \frac{1}{n!} \sum_{\pi \in \mathbb{S}_n} \delta_{\pi \cdot X[n]}(\cdot)
\]  

(8)

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 \( X[n] \); equivalently, by applying to \( X[n] \) a permutation sampled uniformly at random from \( \mathbb{S}_n \).

**Proposition 7** (d-separation with the empirical measure). Suppose \( X[n] \in \mathcal{X}^n \) for some \( n \in \mathbb{N} \). Then \( X[n] \) is exchangeable if and only if

\[
P[ X[n] \in \cdot \mid \mathbb{M}_{X[n]} = m ] = \mathbb{U}_m(\cdot)
\]

(9)

When \( X[n] \) is exchangeable, if \( Y \) is any other random variable such that \((\pi \cdot X[n], Y) \overset{d}{=} (X[n], Y)\) for each \( \pi \in \mathbb{S}_n \), then \( Y \perp \perp_{\mathbb{M}_{X[n]}} X[n] \).

### A.3 Proof of Theorem 3

Firstly, if (5) is true then \((\pi \cdot X[n], f(\eta, \mathbb{M}_{\pi \cdot X[n]})) \overset{d}{=} (X[n], f(\eta, \mathbb{M}_{X[n]}))\) for all \( \pi \in \mathbb{S}_n \) because \( X[n] \) is exchangeable and \( \mathbb{M}_{X[n]} \) is invariant. Conversely, assume that \((\pi \cdot X[n], Y) \overset{d}{=} (X[n], Y)\) for all \( \pi \in \mathbb{S}_n \). Then by Proposition 7, \( Y \perp \perp_{\mathbb{M}_{X[n]}} X[n] \), which implies that \( \mathbb{M}_{X[n]} \) is adequate for \( Y \) from \( X[n] \); by Proposition 6, (5) holds.

### A.4 Proof of Theorem 4

In order to prove Theorem 4, we require a result about equivariance under a finite group \( G \) due to Kallenberg [4] (Prop. 7.9), adapted slightly for our purposes. To state the result, let \( G_x = \{ g \in G : g \cdot x = x \} \) be the stabilizer of \( x \in \mathcal{X} \), and denote the distributive application of \( g \) as \( g \cdot (X, Y) = (g \cdot X, g \cdot Y) \).

**Proposition 8** (Kallenberg [4]). Let \( G \) be a finite group acting measurably on Borel spaces \( \mathcal{X} \) and \( \mathcal{Y} \), and consider random elements \( X \in \mathcal{X} \) and \( Y \in \mathcal{Y} \) such that \( G_X \subseteq G_Y \) almost surely. Then \( g \cdot (X, Y) \overset{d}{=} (X, Y) \) if and only if: (i) \( g \cdot X \overset{d}{=} X \) for all \( g \in G \); and (ii) there exists a measurable function \( f : [0, 1] \times \mathcal{X} \to \mathcal{Y} \) such that

\[
g \cdot Y \overset{d}{=} f(\eta, g \cdot X), \quad \text{for each} \quad g \in G
\]

(10)

for a uniform random variable \( \eta \perp \perp X \).

Observe that (10) implies that

\[
f(\eta, g \cdot X) = g \cdot Y = g \cdot f(\eta, X), \quad \text{for each} \quad g \in G
\]
Proof of Theorem 4. First, assume that $P_{Y_{[n]}|X_{[n]}}$ is $S_n$-equivariant. The proof in this direction establishes the conditional independence relationship
\begin{equation}
Y_i \perp X_i, M_{X_{[n]}}(X_{[n]}, Y_{[n] \setminus i}) ,
\end{equation}
and applies Proposition 6.

Now, by Proposition 8 there exists a measurable $S_n$-equivariant function $f_y : [0, 1] \times \mathcal{X}^n \rightarrow \mathcal{X}^n$ such that $Y_{[n]} \overset{d}{=} f_y(\eta, X_{[n]})$ for $\eta \sim \text{Unif}[0, 1]$ and $\eta \perp X_{[n]}$. Fix $i \in [n]$ and let $\pi_{-i} \in S_n$ be any permutation that fixes $i$. Then
\[ [f_y(\eta, \pi_{-i} \cdot X_{[n]}))]_i = [\pi_{-i} \cdot f_y(\eta, X_{[n]}))]_i = [f_y(\eta, X_{[n]}))]_{\pi_{-i}(i)} = [f_y(\eta, X_{[n]}))]_i = Y_i , \]
so that any permutation that fixes $X_i$ also fixes $Y_i$.

Let $\pi' \in S_{n-1}$. Because $(X_{[n]}, Y_{[n]})$ are $S_n$-equivariant,
\[ (\pi' \cdot (X_{[n]} \setminus i), Y_{[n]} \setminus i), (X_i, Y_i) \overset{d}{=} ((X_{[n]} \setminus i), Y_{[n]} \setminus i), (X_i, Y_i)) \text{ for each } \pi' \in S_{n-1} . \]
Proposition 7 and the assumption that $Y_i \perp X_i, M_{X_{[n]}}(X_{[n]} \setminus i)$ therefore imply
\begin{equation}
Y_i \perp X_i, M_{X_{[n]}}(X_{[n]} \setminus i) ,
\end{equation}
Now, the information contained in $(X_i, M_{X_{[n]}}(X_{[n]} \setminus i))$ is the same as that in $(X_i, M_{X_{[n]}}(X_{[n]} \setminus i))$. Together with
\begin{equation}
\end{equation}

By Proposition 6, there exists a measurable $f_i : [0, 1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{X}$ such that
\[ (X_{[n]}, Y_{[n] \setminus i}, Y_i) \overset{d}{=} (X_{[n]}, Y_{[n] \setminus i}, f_i(\eta_i, X_i, M_{X_{[n]}})) , \]
for $\eta_i \sim \text{Unif}[0, 1]$ and $\eta_i \perp (X_{[n]}, Y_{[n] \setminus i})$. This is true for each $i \in [n]$, and equivariance requires that $f_i = f$ be the same for each $i$ (up to measure-preserving transformations of $\eta_i$), which yields (6).

In the other direction, to show that the right-hand side of (6) implies that $P_{Y_{[n]}|X_{[n]}}$ is $S_n$-equivariant, note that any permutation $\pi \in S_n$ applied to the sequence $(f(\eta_i, X_i, M_{X_{[n]}}))_{i \in [n]}$ is
\[ \pi \cdot (f(\eta_i, X_i, M_{X_{[n]}}))_{i \in [n]} = (f(\eta_{\pi^{-1}(i)}, X_{\pi^{-1}(i)}, M_{\pi \cdot X_{[n]}}))_{i \in [n]} , \]
so that
\[ \pi \cdot (X_{[n]}, Y_{[n]}) = \left( (X_{\pi^{-1}(i)})_{i \in [n]}, (f(\eta_{\pi^{-1}(i)}, X_{\pi^{-1}(i)}, M_{\pi \cdot X_{[n]}}))_{i \in [n]} \right) \overset{d}{=} (X_{[n]}, Y_{[n]}) , \]
where the equality in distribution is due to the fact that the $\eta_i$ are i.i.d., $X_{[n]}$ is exchangeable, and $M_{X_{[n]}}$ is invariant under permutations.

Proof of Proposition 5. (i) is a special (finite) case of Lemma 7.2(i) in Kallenberg [4]. (ii) follows from nearly identical arguments used in the proof of (i), omitted here for brevity.