
Neural network models of exchangeable sequences

Benjamin Bloem-Reddy

Department of Statistics
University of Oxford
Oxford, UK

benjamin.bloem-reddy@stats.ox.ac.uk

Yee Whye Teh

Department of Statistics
University of Oxford
Oxford, UK

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 conditional 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 $\mathbf{X}_n := (X_1, \dots, X_n)$. Y might represent a latent encoding of \mathbf{X}_n , a prediction X_{n+1} , or a label relevant to \mathbf{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 $P_{Y|\mathbf{X}_n}$ of Y given \mathbf{X}_n be invariant to permutations of \mathbf{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 $Y = f(\mathbf{X}_n)$ 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 $P_{Y|\mathbf{X}_n}$ is invariant to permutations of \mathbf{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 \mathbf{X}_n is a point mass at $f(\mathbf{X}_n)$. 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 $\mathbf{X}_n := (X_1, \dots, X_n)$ be a random \mathcal{X} -valued sequence.¹ Denote the Borel σ -algebra of \mathcal{X}^n by $\mathcal{B}(\mathcal{X}^n)$ and the set of all permutations of n elements by \mathbb{S}_n . A probability distribution P on \mathcal{X}^n is *finitely exchangeable* if for every $B_1, \dots, B_n \in \mathcal{B}(\mathcal{X}^n)$,

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

That is, P is invariant to all permutations of the elements of \mathbf{X}_n . We write $X_{[n]} \stackrel{d}{=} \pi \cdot X_{[n]} = X_{\pi \cdot [n]}$ to denote this equality in distribution. More commonly encountered is *infinite exchangeability*, which requires probabilistic coherence between the finite-dimensional distributions of an infinite sequence $X_{\mathbb{N}}$. Only finitely exchangeable sequences of fixed length are considered here; for convenience we say that a sequence \mathbf{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 \mathbf{X}_n . Permutation invariance of the conditional distribution $P_{Y|\mathbf{X}_n}$ of Y given \mathbf{X}_n is defined as follows.

Definition 1. The conditional distribution $P_{Y|\mathbf{X}_n}$ of Y given \mathbf{X}_n is \mathbb{S}_n -**invariant** if for all $B \in \mathcal{B}(\mathcal{Y})$ and $\pi \in \mathbb{S}_n$,

$$P_{Y|\mathbf{X}_n}[Y \in B \mid \pi \cdot \mathbf{X}_n] = P_{Y|\mathbf{X}_n}[Y \in B \mid \mathbf{X}_n] \quad \text{a.s.-}P_{\mathbf{X}_n}. \quad (2)$$

Together with exchangeability of $P_{\mathbf{X}_n}$, conditional invariance implies the joint equality $(\pi \cdot \mathbf{X}_n, Y) \stackrel{d}{=} (\mathbf{X}_n, Y)$ for all $\pi \in \mathbb{S}_n$.

In the neural networks literature, *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 variables plays the same role in \mathbb{S}_n -invariant stochastic neural networks.

Definition 2. Let $\mathbf{Y}_n \in \mathcal{Y}^n$ be a random \mathcal{Y} -valued sequence. The conditional distribution $P_{\mathbf{Y}_n|\mathbf{X}_n}$ of \mathbf{Y}_n given \mathbf{X}_n is \mathbb{S}_n -**equivariant** if for all $B \in \mathcal{B}(\mathcal{Y}^n)$ and $\pi \in \mathbb{S}_n$,

$$P_{\mathbf{Y}_n|\mathbf{X}_n}[\pi \cdot \mathbf{Y}_n \in B \mid \pi \cdot \mathbf{X}_n] = P_{\mathbf{Y}_n|\mathbf{X}_n}[\mathbf{Y}_n \in B \mid \mathbf{X}_n] \quad \text{a.s.-}P_{\mathbf{X}_n}. \quad (3)$$

Together with exchangeability of $P_{\mathbf{X}_n}$, conditional invariance implies the joint equality $(\pi \cdot \mathbf{X}_n, \pi \cdot \mathbf{Y}_n) \stackrel{d}{=} (\mathbf{X}_n, \mathbf{Y}_n)$ for all $\pi \in \mathbb{S}_n$.

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 \mathbf{X}_n , which is defined as

$$\mathbb{M}_{\mathbf{X}_n}(\cdot) = \sum_{i=1}^n \delta_{X_i}(\cdot), \quad (4)$$

where δ_{X_i} denotes an atom of unit mass at X_i ;² it can be viewed as a generalization of a one-hot encoding to arbitrary spaces. As such, $\mathbb{M}_{\mathbf{X}_n}$ can be thought of as a pooling function that is invariant under permutations; it discards the order of the elements of \mathbf{X}_n , but retains all other information. Furthermore, any function of the empirical measure is invariant to permutations of \mathbf{X}_n .

The first main result characterizes the functional relationship between \mathbf{X}_n and Y , for any Y such that $P_{Y|\mathbf{X}_n}$ is \mathbb{S}_n -invariant. We use the notation $X \perp\!\!\!\perp Y$ to indicate that X and Y are stochastically independent, and $X \perp\!\!\!\perp_Z Y$ to indicate that X and Y are conditionally independent, given Z .

Theorem 3 (Invariant representation). *For fixed $n \in \mathbb{N}$, suppose $\mathbf{X}_n \in \mathcal{X}^n$, is an exchangeable sequence and $Y \in \mathcal{Y}$ is another random variable. Then $P_{Y|\mathbf{X}_n}$ is \mathbb{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*

$$(\mathbf{X}_n, Y) \stackrel{\text{a.s.}}{=} (\mathbf{X}_n, f(\eta, \mathbb{M}_{\mathbf{X}_n})) \quad \text{where } \eta \sim \text{Unif}[0, 1] \quad \text{and } \eta \perp\!\!\!\perp \mathbf{X}_n. \quad (5)$$

¹We 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].

² $\delta_{X_i}(A) = 1$ if $X_i \in A$ and 0 otherwise, for any measurable set A .

Note that in the deterministic case, the noise variable η is ignored, in which case the representation is equivalent to that in Theorem 2 of [10].

The second main result characterizes the functional relationship between \mathbf{X}_n and \mathbf{Y}_n , for any \mathbf{Y}_n such that $P_{\mathbf{Y}_n|\mathbf{X}_n}$ is \mathbb{S}_n -equivariant, and whose elements are conditionally independent given \mathbf{X}_n . The latter condition is satisfied by any architecture that does not have within-layer connections.

Theorem 4 (Equivariant representation). *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 $Y_i \perp\!\!\!\perp_{\mathbf{X}_n} (\mathbf{Y}_n \setminus Y_i)$ for all $i \in [n]$. Then $P_{\mathbf{Y}_n|\mathbf{X}_n}$ is \mathbb{S}_n -equivariant if and only if there is a measurable function $g : [0, 1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{Y}$ such that*

$$(\mathbf{X}_n, \mathbf{Y}_n) \stackrel{\text{a.s.}}{=} (\mathbf{X}_n, (g(\eta_i, X_i, \mathbb{M}_X))_{i \in [n]}) \quad \text{where } \eta_i \stackrel{\text{iid}}{\sim} \text{Unif}[0, 1] \text{ and } \eta_i \perp\!\!\!\perp \mathbf{X}_n. \quad (6)$$

As before, in the deterministic case the noise variables η_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.

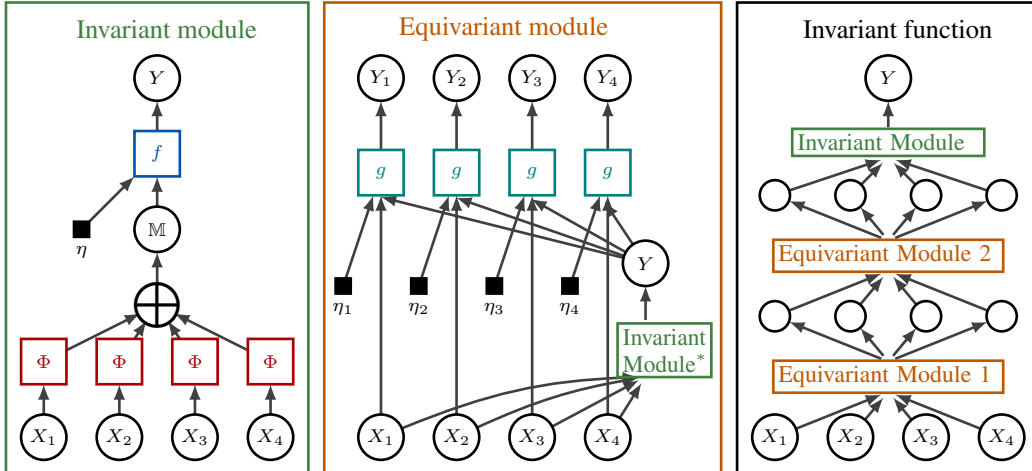
Proposition 5 (Transitivity of equivariance and invariance).

- (i) *Let $\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_{[n]}$ be random sequences of length n such that $\pi \cdot (\mathbf{X}_n, \mathbf{Y}_n) \stackrel{d}{=} (\mathbf{X}_n, \mathbf{Y}_n)$ and $\pi \cdot (\mathbf{Y}_n, \mathbf{Z}_{[n]}) \stackrel{d}{=} (\mathbf{Y}_n, \mathbf{Z}_{[n]})$ for all $\pi \in \mathbb{S}_n$, and $\mathbf{X}_n \perp\!\!\!\perp_{\mathbf{Y}_n} \mathbf{Z}_{[n]}$. Then $\pi \cdot (\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_{[n]}) \stackrel{d}{=} (\mathbf{X}_n, \mathbf{Y}_n, \mathbf{Z}_{[n]})$ for all $\pi \in \mathbb{S}_n$.*
- (ii) *Let $\mathbf{X}_n, \mathbf{Y}_n$ be random sequences of length n and Z another random variable. Suppose that $\pi \cdot (\mathbf{X}_n, \mathbf{Y}_n) \stackrel{d}{=} (\mathbf{X}_n, \mathbf{Y}_n)$ and $(\pi \cdot \mathbf{Y}_n, Z) \stackrel{d}{=} (\mathbf{Y}_n, Z)$ for all $\pi \in \mathbb{S}_n$. Then $(\pi \cdot \mathbf{X}_n, \pi \cdot \mathbf{Y}_n, Z) \stackrel{d}{=} (\mathbf{X}_n, \mathbf{Y}_n, Z)$ for all $\pi \in \mathbb{S}_n$.*

Together, (i) and (ii) suggest how to construct a stochastic or deterministic neural network that is \mathbb{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 $\Phi : \mathcal{X} \rightarrow \mathbb{R}$ parameterized by a neural network, applied to each element X_i , then pooled into the empirical measure. f , the function of noise η 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 η variable), as indicated by the *. The full invariant architecture (with two equivariant layers shown) is displayed on the right.



Acknowledgments

BBR’s and YWT’s research leading to these results received funding from the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071.

References

- [1] T. Austin. Exchangeable random measures. *Ann. Inst. H. Poincaré Probab. Statist.*, 51(3): 842–861, 08 2015.
- [2] T. Cohen and M. Welling. Group equivariant convolutional networks. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA, 2016. PMLR. URL <http://proceedings.mlr.press/v48/cohenc16.html>.
- [3] O. Kallenberg. *Foundations of Modern Probability*. Springer-Verlag New York, 2nd edition, 2002.
- [4] O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.
- [5] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.
- [6] R. Kondor and S. Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2747–2755, Stockholm, Sweden, 2018. PMLR.
- [7] S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
- [8] S. Ravanbakhsh, J. Schneider, and B. Póczos. Equivariance through parameter-sharing. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2892–2901. PMLR, 2017. URL <http://proceedings.mlr.press/v70/ravanbakhsh17a.html>.
- [9] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Beijing, China, 22–24 Jun 2014. PMLR. URL <http://proceedings.mlr.press/v32/rezende14.html>.
- [10] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3391–3401. Curran Associates, Inc., 2017.

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

$$(W, Z) \stackrel{\text{a.s.}}{=} (W, f(\eta, S(W))) \quad \text{where } \eta \sim \text{Unif}[0, 1] \quad \text{and } \eta \perp\!\!\!\perp W. \quad (7)$$

In particular, $Z \stackrel{\text{a.s.}}{=} f(\eta, S(W))$.

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 ξ and ζ trivially, so that we have $\zeta \stackrel{\text{a.s.}}{=} f(\eta, \xi)$. 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 $S : \mathcal{X}^n \rightarrow \mathcal{Y}$ d-separates \mathbf{X}_n from Y if Y is conditionally independent from \mathbf{X}_n , given $S(\mathbf{X}_n)$. We denote this by $Y \perp\!\!\!\perp_{S(\mathbf{X}_n)} \mathbf{X}_n$.

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 $P(\mathbf{X}_n \mid \mathbb{M}_{\mathbf{X}_n} = m)$ is the uniform distribution on all sequences (x_1, \dots, 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 \mathbf{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 $(\pi \cdot \mathbf{X}_n, Y) \stackrel{\text{d}}{=} (\mathbf{X}_n, Y)$, for all $\pi \in \mathbb{S}_n$ (which amounts to saying that \mathbf{X}_n is exchangeable and the conditional distribution of Y given \mathbf{X}_n is invariant under permutations of \mathbf{X}_n). In order to state the result, define the *urn law* of \mathbf{X}_n as

$$\mathbb{U}_{\mathbf{X}_n}^\pi(\cdot) = \frac{1}{n!} \sum_{\pi \in \mathbb{S}_n} \delta_{\pi \cdot \mathbf{X}_n}(\cdot). \quad (8)$$

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 \mathbf{X}_n ; equivalently, by applying to \mathbf{X}_n a permutation sampled uniformly at random from \mathbb{S}_n .

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

$$P[\mathbf{X}_n \in \cdot \mid \mathbb{M}_{\mathbf{X}_n} = m] = \mathbb{U}_m^\pi(\cdot). \quad (9)$$

When \mathbf{X}_n is exchangeable, if Y is any other random variable such that $(\pi \cdot \mathbf{X}_n, Y) \stackrel{\text{d}}{=} (\mathbf{X}_n, Y)$ for each $\pi \in \mathbb{S}_n$, then $Y \perp\!\!\!\perp_{\mathbb{M}_{\mathbf{X}_n}} \mathbf{X}_n$.

A.3 Proof of Theorem 3

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

A.4 Proof of Theorem 4

In order to prove Theorem 4, we require a result about equivariance under a finite group \mathcal{G} due to Kallenberg [4] (Prop. 7.9), adapted slightly for our purposes. To state the result, let $\mathcal{G}_x = \{g \in \mathcal{G} : g \cdot x = x\}$ 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 \mathcal{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 $\mathcal{G}_X \subseteq \mathcal{G}_Y$ almost surely. Then*

$g \cdot (X, Y) \stackrel{d}{=} (X, Y)$ if and only if: (i) $g \cdot X \stackrel{d}{=} X$ for all $g \in \mathcal{G}$; and (ii) there exists a measurable function $f : [0, 1] \times \mathcal{X} \rightarrow \mathcal{Y}$ such that

$$g \cdot Y \stackrel{\text{a.s.}}{=} f(\eta, g \cdot X), \quad \text{for each } g \in \mathcal{G}, \quad (10)$$

for a uniform random variable $\eta \perp\!\!\!\perp X$.

Observe that (10) implies that

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

Proof of Theorem 4. First, assume that $P_{\mathbf{Y}_n | \mathbf{X}_n}$ is \mathbb{S}_n -equivariant. The proof in this direction establishes the conditional independence relationship

$$Y_i \perp\!\!\!\perp_{X_i, \mathbb{M}_{\mathbf{X}_n}} (\mathbf{X}_n, \mathbf{Y}_{n \setminus i}), \quad (11)$$

and applies Proposition 6.

Now, by Proposition 8 there exists a measurable \mathbb{S}_n -equivariant function $f_y : [0, 1] \times \mathcal{X}^n \rightarrow \mathcal{X}^n$ such that $\mathbf{Y}_n \stackrel{\text{a.s.}}{=} f_y(\eta, \mathbf{X}_n)$ for $\eta \sim \text{Unif}[0, 1]$ and $\eta \perp\!\!\!\perp \mathbf{X}_n$. Fix $i \in [n]$ and let $\pi_{-i} \in \mathbb{S}_n$ be any permutation that fixes i . Then

$$[f_y(\eta, \pi_{-i} \cdot \mathbf{X}_n)]_i = [\pi_{-i} \cdot f_y(\eta, \mathbf{X}_n)]_i = [f_y(\eta, \mathbf{X}_n)]_{\pi_{-i}(i)} = [f_y(\eta, \mathbf{X}_n)]_i = Y_i,$$

so that any permutation that fixes X_i also fixes Y_i .

Let $\pi' \in \mathbb{S}_{n-1}$. Because $(\mathbf{X}_n, \mathbf{Y}_n)$ are \mathbb{S}_n -equivariant,

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

Proposition 7 and the assumption that $Y_i \perp\!\!\!\perp_{X_i, \mathbb{M}_{\mathbf{X}_{[n] \setminus i}}} \mathbf{Y}_{n \setminus i}$ therefore imply

$$Y_i \perp\!\!\!\perp_{X_i, \mathbb{M}_{\mathbf{X}_{[n] \setminus i}}} (\mathbf{X}_n, \mathbf{Y}_{n \setminus i}). \quad (12)$$

Now, the information contained in $(X_i, \mathbb{M}_{\mathbf{X}_{[n] \setminus i}})$ is the same as that in $(X_i, \mathbb{M}_{\mathbf{X}_n})$. Together with (12), this implies the key conditional independence relationship (11).

By Proposition 6, there exists a measurable $f_i : [0, 1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{X}$ such that

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

for $\eta_i \sim \text{Unif}[0, 1]$ and $\eta_i \perp\!\!\!\perp (\mathbf{X}_n, \mathbf{Y}_{n \setminus i})$. This is true for each $i \in [n]$, and equivariance requires that $f_i = f$ be the same for each i (up to measure-preserving transformations of η_i), which yields (6).

In the other direction, to show that the right-hand side of (6) implies that $P_{\mathbf{Y}_n | \mathbf{X}_n}$ is \mathbb{S}_n -equivariant, note that any permutation $\pi \in \mathbb{S}_n$ applied to the sequence $(f(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_n}))_{i \in [n]}$ is

$$\pi \cdot (f(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_n}))_{i \in [n]} = (f(\eta_{\pi^{-1}(i)}, X_{\pi^{-1}(i)}, \mathbb{M}_{\pi \cdot \mathbf{X}_n}))_{i \in [n]},$$

so that

$$\pi \cdot (\mathbf{X}_n, \mathbf{Y}_n) = \left((X_{\pi^{-1}(i)})_{i \in [n]}, (f(\eta_{\pi^{-1}(i)}, X_{\pi^{-1}(i)}, \mathbb{M}_{\pi \cdot \mathbf{X}_n}))_{i \in [n]} \right) \stackrel{d}{=} (\mathbf{X}_n, \mathbf{Y}_n),$$

where the equality in distribution is due to the fact that the η_i are i.i.d., \mathbf{X}_n is exchangeable, and $\mathbb{M}_{\mathbf{X}_n}$ is invariant under permutations. \square

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. \square