Probabilistic symmetry and invariant neural networks

Benjamin Bloem-Reddy, University of Oxford
Work with Yee Whye Teh

14 January 2019, UBC Computer Science

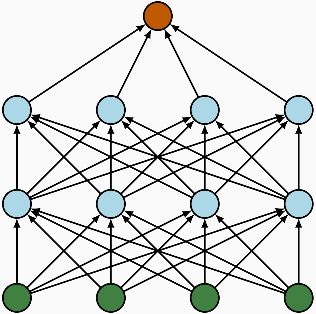
Outline

- Symmetry in neural networks
  - Permutation-invariant neural networks
- Symmetry in probability and statistics
  - Exchangeable sequences
- Permutation-invariant neural networks as exchangeable probability models
- Symmetry in neural networks as probabilistic symmetry

Deep learning and statistics

- Deep neural networks have been applied successfully in a range of settings.
- Effort under way to improve performance in *data poor* and *semi-/unsupervised* domains.
- Focus on **symmetry**.
- The study of symmetry in probability and statistics has a **long** history.

Symmetric neural networks



$$f_{\ell,i} = \sigma\left(\sum_{j=1}^{n} w_{i,j}^{(\ell)} f_{\ell-1,j}\right)$$

For input X and output Y, model $Y = h(X)$, where $h \in \mathcal{H}$ is a neural network.

If X and Y are assumed to satisfy a symmetry property, how is $\mathcal{H}$ restricted?

Convolutional neural networks encode translation invariance:

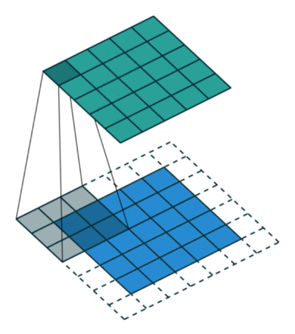


Illustration from medium.freecodecamp.org

Encoding symmetry in network architecture is a Good Thing*.

Stabler training and *better generalization* through

- reduction in dimension of parameter space through weight-tying; and
- capturing structure at multiple scales via pooling.

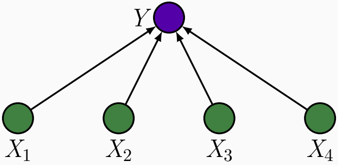
Historical note:

Interest in invariant neural networks goes back at least to Minsky and Papert [MP88]; extended by Shawe-Taylor and Wood [Sha89; WS96]. More recent work by a host of others.

Consider a sequence $\mathbf{X}_n := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

Permutation invariance:

$$Y = h(\mathbf{X}_n) = h(\pi \cdot \mathbf{X}_n) \text{ for all } \pi \in \mathbb{S}_n.$$

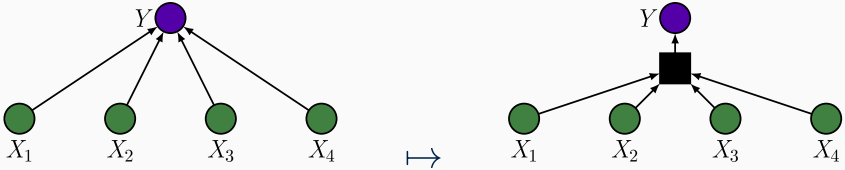


Neural networks for permutation-invariant data [Zah+17]

Consider a sequence $\mathbf{X}_n := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

Permutation invariance:

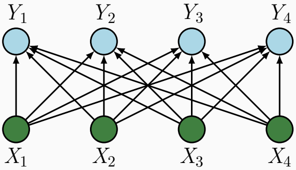
$$Y = h(\mathbf{X}_n) = h(\pi \cdot \mathbf{X}_n) \text{ for all } \pi \in \mathbb{S}_n.$$



$$Y = h(\mathbf{X}_n) \quad \mapsto \quad Y = \tilde{h}\Big(\sum_{i=1}^{n} \phi(X_i) \Big)$$

Equivariance:

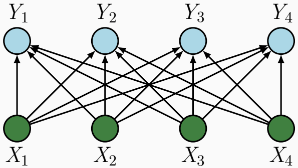
$$\mathbf{Y}_n = h(\mathbf{X}_n) \text{ such that } h(\pi \cdot \mathbf{X}_n) = \pi \cdot h(\mathbf{X}_n) \text{ for all } \pi \in \mathbb{S}_n.$$

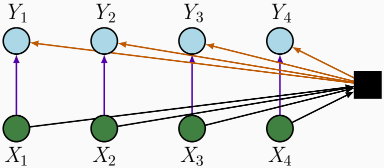


Neural networks for permutation-invariant data [Zah+17]

Equivariance:

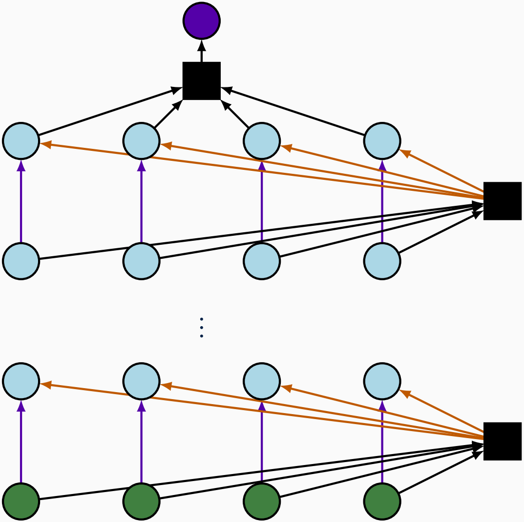
$$\mathbf{Y}_n = h(\mathbf{X}_n) \text{ such that } h(\pi \cdot \mathbf{X}_n) = \pi \cdot h(\mathbf{X}_n) \text{ for all } \pi \in \mathbb{S}_n.$$





$$[h(\mathbf{X}_n)]_i = \sigma\left(\sum_{j=1}^{n} w_{i,j} X_j\right) \quad \mapsto \quad [h(\mathbf{X}_n)]_i = \sigma\left(w_0 X_i + w_1 \sum_{j=1}^{n} X_j\right)$$

Neural networks for permutation-invariant data



⟨⟨Deep learning hat, off; statistics hat, on⟩⟩

Note to students: These were the first Google Image results for ”deep learning hat” and ”statistics hat”.

You could probably make some money making decent hats.

Consider a sequence $\mathbf{X}_n := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

A *statistical model* of $\mathbf{X}_n$ is a family of probability distributions on $\mathcal{X}^n$:

$$\mathcal{P} = \{P_\theta : \theta \in \Omega\} \; .$$

If X is assumed to satisfy a symmetry property,
how is $\mathcal{P}$ restricted?

A distribution P on $\mathcal{X}^n$ is *exchangeable* if

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

$\mathbf{X}_{\mathbb{N}}$ is infinitely exchangeable if this is true for all prefixes $\mathbf{X}_n \subset \mathbf{X}_{\mathbb{N}}$, $n \in \mathbb{N}$.

de Finetti's theorem:

$$\mathbf{X}_{\mathbb{N}} \text{ exchangeable} \iff X_i \mid Q \overset{\text{iid}}{\sim} Q \text{ for some random } Q.$$

Implication for Bayesian inference:

Our models for $\mathbf{X}_{\mathbb{N}}$ need only consist of i.i.d. distributions on $\mathcal{X}$.

Analogous theorems for other symmetries. The book by Kallenberg [Kal05] collects many of them. Some other accessible references: [Dia88; OR15].

de Finetti's theorem may fail for finite exchangeable sequences.

What else can we say?

The *empirical measure* of $\mathbf{X}_n$ is

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

The empirical measure is a *sufficient statistic*: P is exchangeable iff

$$P(\mathbf{X}_n \in \bullet \mid \mathbb{M}_{\mathbf{X}_n} = m) = \mathbb{U}_m(\bullet)\ ,$$

where $\mathbb{U}_m$ is the uniform distribution on all sequences $(x_1, \ldots, x_n)$ with empirical measure m.

The empirical measure is a *sufficient statistic*: P is exchangeable iff

$$P(\mathbf{X}_n \in \bullet \mid \mathbb{M}_{\mathbf{X}_n} = m) = \mathbb{U}_m(\bullet)\ ,$$

where $\mathbb{U}_m$ is the uniform distribution on all sequences $(x_1, \ldots, x_n)$ with empirical measure m.

Consider Y such that $(\pi \cdot \mathbf{X}_n, Y) \stackrel{\mathrm{d}}{=} (\mathbf{X}_n, Y)$.

The empirical measure is an *adequate statistic* for any such Y:

$$P(Y \in \bullet \mid \mathbf{X}_n = \mathbf{x}_n) = P(Y \in \bullet \mid \mathbb{M}_{\mathbf{X}_n} = \mathbb{M}_{\mathbf{x}_n}).$$

$\mathbb{M}_{\mathbf{X}_n}$ contains all information in $\mathbf{X}_n$ that is relevant for predicting Y.

A useful theorem

Theorem (Invariant representation; B-R, Teh)

Suppose $\mathbf{X}_n$ is an exchangeable sequence.

Then $(\pi \cdot \mathbf{X}_n, Y) \stackrel{\mathrm{d}}{=} (\mathbf{X}_n, Y)$ for all $\pi \in \mathbb{S}_n$ if and only if there is a measurable function $\tilde{h} : [0,1] \times \mathcal{M}(\mathcal{X}) \to \mathcal{Y}$ such that

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

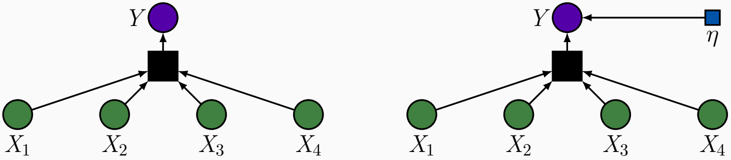
Theorem (Invariant representation; B-R, Teh)

Suppose $\mathbf{X}_n$ is an exchangeable sequence.

Then $(\pi \cdot \mathbf{X}_n, Y) \overset{\text{d}}{=} (\mathbf{X}_n, Y)$ for all $\pi \in \mathbb{S}_n$ if and only if there is a measurable function $\tilde{h} : [0,1] \times \mathcal{M}(\mathcal{X}) \to \mathcal{Y}$ such that

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

Deterministic invariance [Zah+17] $\mapsto$ stochastic invariance [B-R, Teh]



$$Y = \tilde{h}\left(\sum_{i=1}^{n} \phi(X_i)\right) \quad \mapsto \quad Y = \tilde{h}\left(\eta, \sum_{i=1}^{n} \delta_{X_i}\right)$$

Another useful theorem

Theorem (Equivariant representation; B-R, Teh)

Suppose $\mathbf{X}_n$ is an exchangeable sequence and $Y_i \perp\!\!\!\perp_{\mathbf{X}_n} (\mathbf{Y}_n \setminus Y_i)$.

Then $(\pi \cdot \mathbf{X}_n, \pi \cdot \mathbf{Y}_n) \overset{\mathrm{d}}{=} (\mathbf{X}_n, \mathbf{Y}_n)$ for all $\pi \in \mathbb{S}_n$ if and only if there is a measurable function $\tilde{h} : [0,1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \to \mathcal{Y}$ such that

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

Another useful theorem

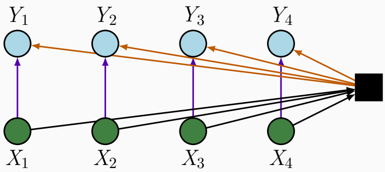
Theorem (Equivariant representation; B-R, Teh)

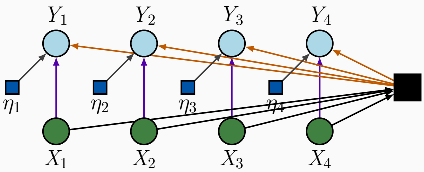
Suppose $\mathbf{X}_n$ is an exchangeable sequence and $Y_i \perp\!\!\!\perp_{\mathbf{X}_n} (\mathbf{Y}_n \setminus Y_i)$.

Then $(\pi \cdot \mathbf{X}_n, \pi \cdot \mathbf{Y}_n) \overset{\text{d}}{=} (\mathbf{X}_n, \mathbf{Y}_n)$ for all $\pi \in \mathbb{S}_n$ if and only if there is a measurable function $\tilde{h} : [0,1] \times \mathcal{X} \times \mathcal{M}(\mathcal{X}) \to \mathcal{Y}$ such that

$$(\mathbf{X}_n, \mathbf{Y}_n) \overset{\text{a.s.}}{=} \left(\mathbf{X}_n, (\tilde{h}(\eta_i, X_i, \mathbb{M}_{\mathbf{X}_n}))_{i \in [n]}\right) \text{ and } \eta_i \overset{\text{iid}}{\sim} \text{Unif}[0,1],$$
$$(\eta_i)_{i\in[n]} \perp\!\!\!\perp \mathbf{X}_n.$$

Deterministic equivariance [Zah+17] $\mapsto$ stochastic equivariance [B-R, Teh]





$$Y_i = \sigma\left(w_0 X_i + w_1 \sum_{j=1}^{n} X_j\right) \quad \mapsto \quad Y_i = \tilde{h}\left(\eta_i, X_i, \sum_{j=1}^{n} \delta_{X_j}\right)$$

Outline

- Symmetry in neural networks
  - Permutation-invariant neural networks
- Symmetry in probability and statistics
  - Exchangeable sequences
- Permutation-invariant neural networks as exchangeable probability models
- **Symmetry in neural networks as probabilistic symmetry**

For a group $\mathcal{G}$ acting on a set $\mathcal{X}$:

- The *orbit* of any $x \in \mathcal{X}$ is the subset of $\mathcal{X}$ generated by applying $\mathcal{G}$ to x: $\mathcal{G} \cdot x = \{g \cdot x; g \in \mathcal{G}\}$.
- A *maximal invariant statistic* $M : \mathcal{X} \to \mathcal{S}$
  (i) is constant on an orbit, i.e., $M(g \cdot x) = M(x)$ for all $g \in \mathcal{G}$ and $x \in \mathcal{X}$; and
  (ii) takes a different value on each orbit, i.e., $M(x_1) = M(x_2)$ implies $x_1 = g \cdot x_2$ for some $g \in \mathcal{G}$.
- A *maximal equivariant* $\tau : \mathcal{X} \to \mathcal{G}$ satisfies

$$\tau(g \cdot X) = g \cdot \tau(x) \;, \quad g \in \mathcal{G} \;, \; x \in \mathcal{X} \;.$$

Theorem (B-R, Teh)

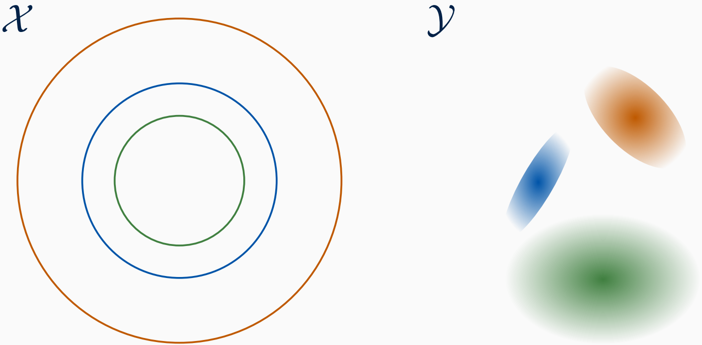
Let $\mathcal{G}$ be a compact group and assume that $g \cdot X \overset{\mathrm{d}}{=} X$ for all $g \in \mathcal{G}$.

Let $M : \mathcal{X} \to \mathcal{S}$ be a maximal invariant.

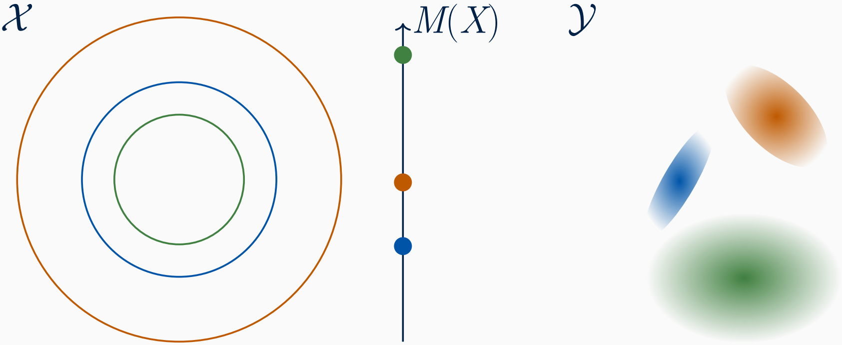
Then $(g \cdot X, Y) \overset{\mathrm{d}}{=} (X, Y)$ for all $g \in \mathcal{G}$ if and only if there exists a measurable function $\tilde{h} : [0,1] \times \mathcal{S} \to \mathcal{Y}$ such that

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

$$P(g \cdot X, Y) = P(X, Y) \text{ for all } g \in \mathcal{G}$$



$$P(g \cdot X, M(g \cdot X), Y) = P(X, M(X), Y) \text{ for all } g \in \mathcal{G}$$
$$\Rightarrow \; Y \perp\!\!\!\perp_{M(X)} X$$



Theorem (Kallenberg; B-R, Teh)

Let $\mathcal{G}$ be a compact group and assume that $g \cdot X \overset{\mathrm{d}}{=} X$ for all $g \in \mathcal{G}$.

Assume that a maximal equivariant $\tau : \mathcal{X} \to \mathcal{G}$ exists.

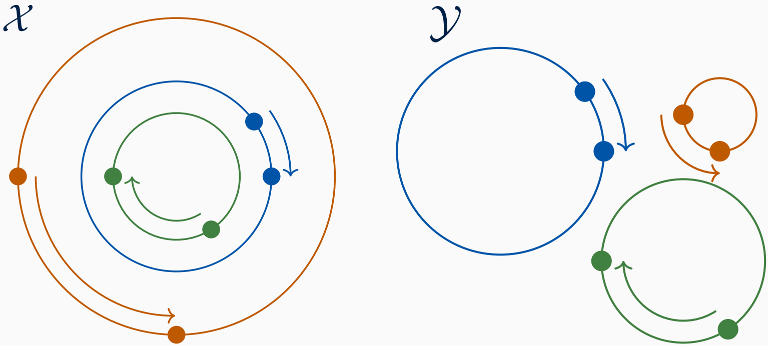
Then $(g \cdot X, g \cdot Y) \overset{\mathrm{d}}{=} (X, Y)$ for all $g \in \mathcal{G}$ if and only if there exists a measurable function $\tilde{h} : [0,1] \times \mathcal{X} \to \mathcal{Y}$ such that

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

where $\tilde{h}$ is equivariant:

$$\tilde{h}(\eta, g \cdot X) \overset{\text{a.s.}}{=} g \cdot \tilde{h}(\eta, X) \, , \quad g \in \mathcal{G} \, .$$

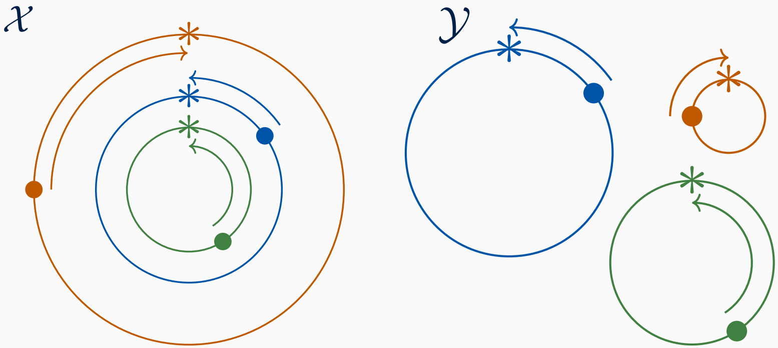
$$P(g \cdot X, g \cdot Y) = P(X, Y) \text{ for all } g \in \mathcal{G}$$



$$P(g \cdot X, \tau(g \cdot X)^{-1} \cdot g \cdot X, g \cdot Y) = P(X, \tau(X)^{-1} \cdot X, Y)$$

for all $g \in \mathcal{G}$

$$\Rightarrow \tau(X)^{-1} \cdot Y \perp\!\!\!\perp_{\tau(X)^{-1} \cdot X} X$$



Some answers

- Sufficiency/adequacy provides the magic.
- Similar results for exchangeable graphs/arrays/tensors and some other related structures.
- Framework is general enough that it catches a lot of existing work as special cases.
- Suggests some new (stochastic) network architectures.

Many questions

- There are models with sufficient statistics that don't have **group** symmetry (though they typically have a set of symmetry transformations)—what are the analogous results? Are they useful?
- Evidence that adding noise during training has beneficial effects; in this context it amounts to the difference between deterministic invariance and distributional invariance—can we prove anything rigorous in these settings?
- Relatedly, can we put the "fact" (encoding symmetry in neural networks is a Good Thing) on rigorous footing?

Thank you.

[Aus13] Tim Austin. “Exchangeable random arrays”. Lecture notes for IIS. 2013. URL: http://www.math.ucla.edu/~tim/ExchnotesforIISc.pdf.

[Coh+18] Taco S. Cohen et al. “Spherical CNNs”. In: *International Conference on Learning Representations*. 2018.

[CW16] Taco S. Cohen and Max Welling. “Group Equivariant Convolutional Networks”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016, pp. 2990–2999. URL: http://proceedings.mlr.press/v48/cohenc16.html.

[Dia88] P. Diaconis. “Sufficiency as statistical symmetry”. In: *Proceedings of the AMS Centennial Symposium*. Ed. by F. Browder. American Mathematical Society, 1988, pp. 15–26.

[GD14] Robert Gens and Pedro M Domingos. “Deep Symmetry Networks”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2537–2545. URL: http://papers.nips.cc/paper/5424-deep-symmetry-networks.pdf.

[Har+18] Jason Hartford et al. “Deep Models of Interactions Across Sets”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1914–1923.

[Her+18] Roei Herzig et al. “Mapping Images to Scene Graphs with Permutation-Invariant Structured Prediction”. In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 7211–7221.

[Kal05] Olav Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.

[KT18] Risi Kondor and Shubhendu Trivedi. “On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, 2018, pp. 2747–2755.

[MP88] Marvin L. Minsky and Seymour A. Papert. *Perceptrons: Expanded Edition*. Cambridge, MA, USA: MIT Press, 1988.

[OR15] Peter Orbanz and Daniel M. Roy. “Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2 (Feb. 2015), pp. 437–461.

[RSP17] Siamak Ravanbakhsh, Jeff Schneider, and Barnabás Póczos. “Equivariance Through Parameter-Sharing”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 2892–2901.

[Sha89] John Shawe-Taylor. “Building symmetries into feedforward networks”. In: *1989 First IEE International Conference on Artificial Neural Networks, (Conf. Publ. No. 313)*. Oct. 1989, pp. 158–162.

[WS96] Jeffrey Wood and John Shawe-Taylor. “Representation theory and invariant neural networks”. In: *Discrete Applied Mathematics* 69.1 (1996), pp. 33–60.

[Zah+17] Manzil Zaheer et al. “Deep Sets”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 3391–3401.

Recent work generalizes the idea to other symmetries and data:

- Affine transformations (translation, rotation, scaling, shear) [GD14]
- Discrete translations, reflections, rotations [CW16]
- Continuous rotations in three dimensions [Coh+18]
- Permutations of sequences [Zah+17] and arrays [Har+18; Her+18]
- Fairly general permutation group symmetries [RSP17]
- Compact groups [KT18]
- Discrete groups, finite linear groups [Sha89; WS96]

A useful tool: noise outsourcing (e.g., [Aus13])

If X and Y are random variables in "nice" (e.g., Borel) spaces $\mathcal{X}$ and $\mathcal{Y}$, then there are a random variable $\eta \sim \mathrm{Unif}[0,1]$ and a measurable function $h : [0,1] \times \mathcal{X} \to \mathcal{Y}$ such that $\eta \perp\!\!\!\perp X$ and

$$(X, Y) = (X, h(\eta, X)) \quad \text{a.s.}$$

Can show that if $S(X)$ is adequate for Y, then

$$(X, Y) = (X, \tilde{h}(\eta, S(X))) \quad \text{a.s.}$$