Invariant neural networks and probabilistic symmetry

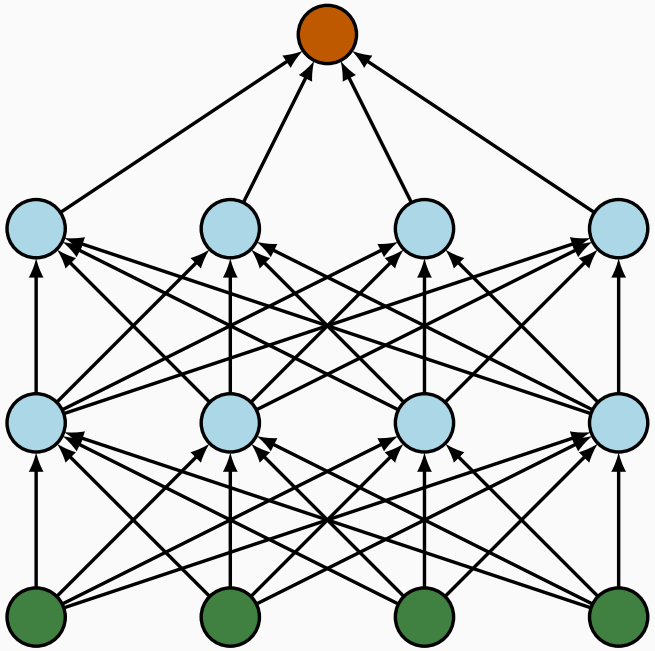
Benjamin Bloem-Reddy, University of Oxford
Work with Yee Whye Teh

5 October 2018, OxWaSP Workshop

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?

Symmetric neural networks

Convolutional neural networks encode translation invariance:

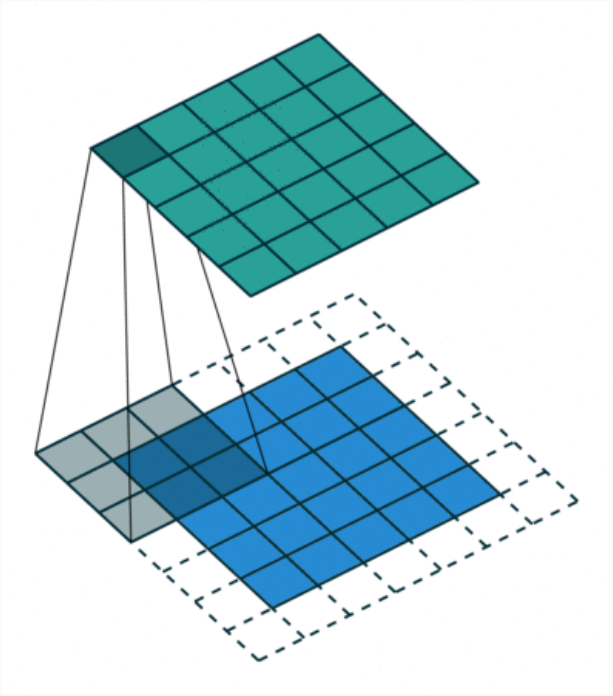


Illustration from medium.freecodecamp.org

Why symmetry?

Encoding symmetry in network architecture is a Good Thing*, i.e., it results in **stabler training** and **better generalization** through

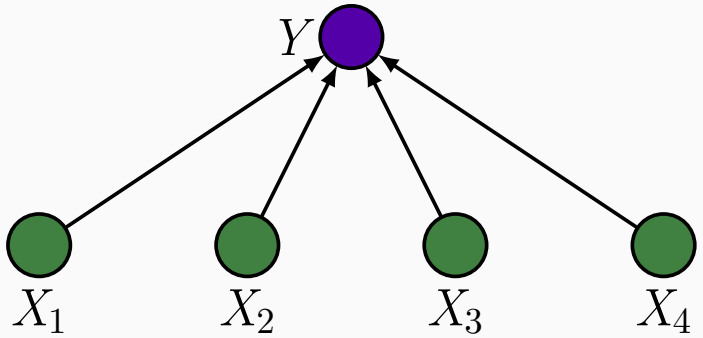
- reduction in dimension of parameter space through weight-tying; and
- capturing structure at multiple scales via pooling.

* Oft-stated "fact". Mostly supported by heuristics and intuition, some empirical evidence, loose connections to learning theory and what we "know" about high-dimensional data analysis. Some PAC theory to this end [Sha91; Sha95]; I haven't found anything else.

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

Consider a sequence $X_{[n]} := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

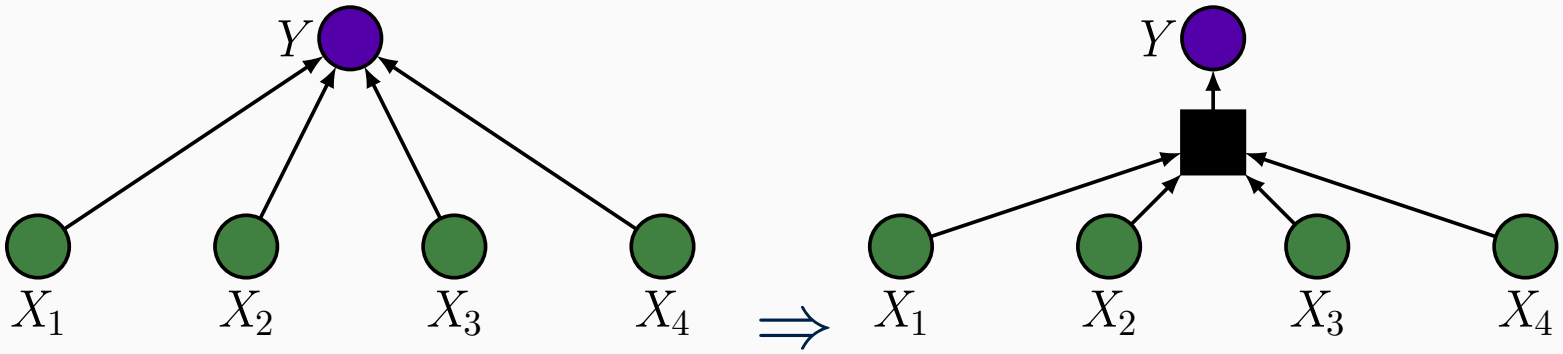
Invariance: $Y = h(X_{[n]}) = h(\pi \cdot X_{[n]})$ **for all** $\pi \in \mathbb{S}_n$**.**



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

Consider a sequence $X_{[n]} := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

Invariance: $Y = h(X_{[n]}) = h(\pi \cdot X_{[n]})$ **for all** $\pi \in \mathbb{S}_n$**.**

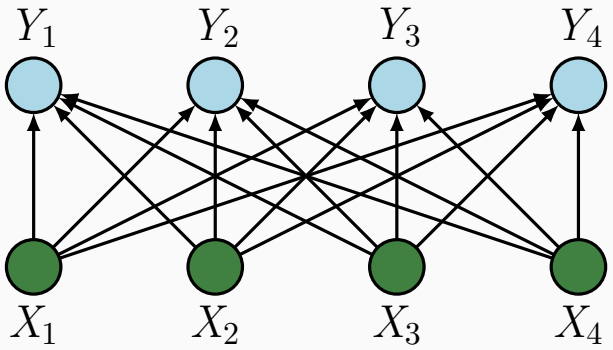


$$Y = h(X_{[n]}) \quad \mapsto \quad Y = \tilde{h}\left(\sum_{i=1}^{n} \phi(X_i)\right)$$

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

Consider a sequence $X_{[n]} := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

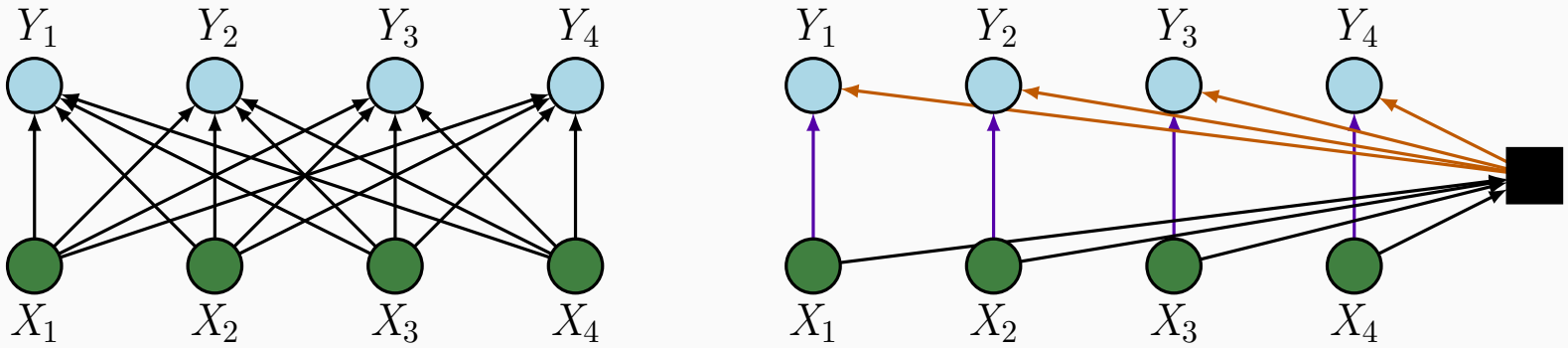
Equivariance: $Y_{[n]} = h(X_{[n]})$ **such that** $h(\pi \cdot X_{[n]}) = \pi \cdot h(X_{[n]})$ **for all** $\pi \in \mathbb{S}_n$**.**



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

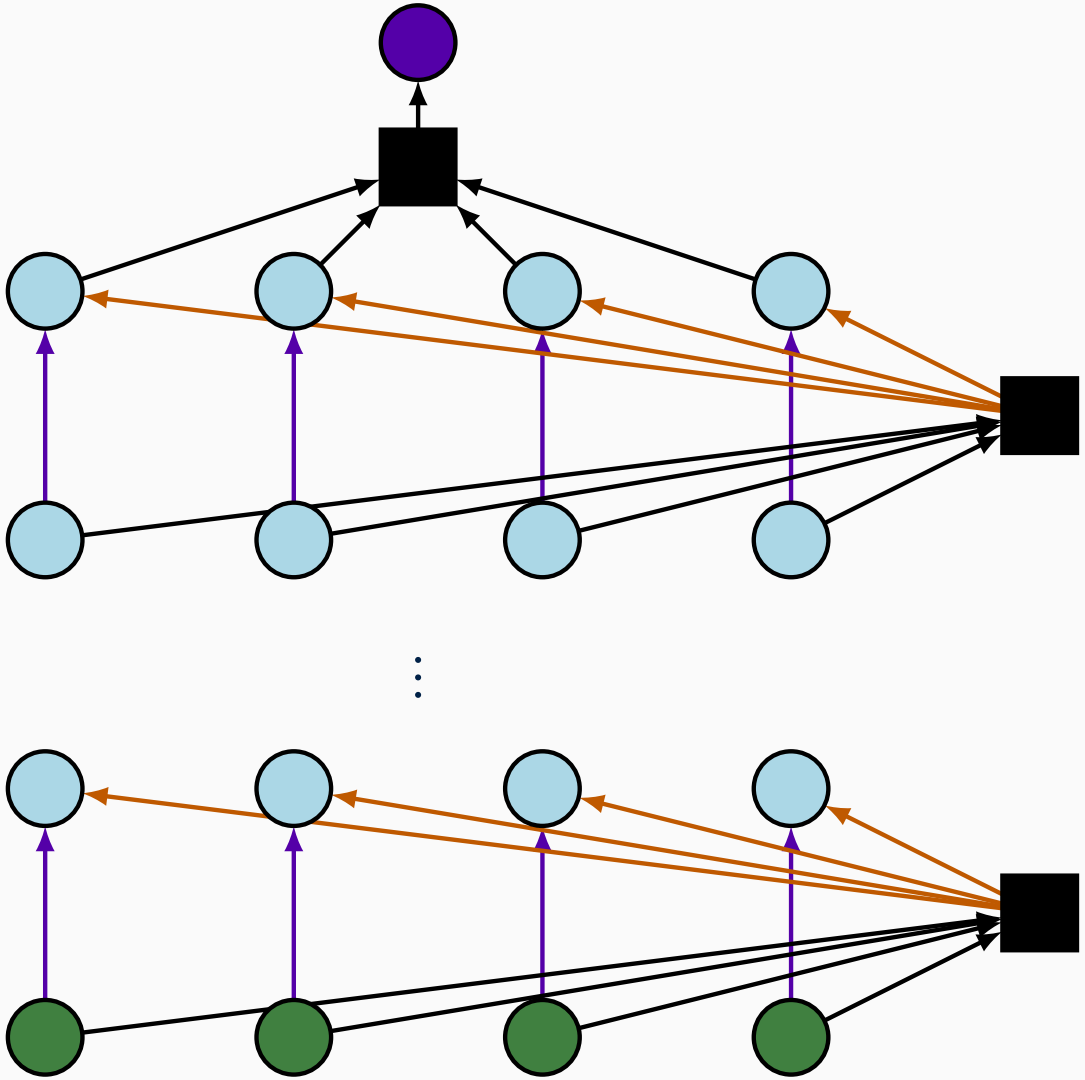
Consider a sequence $X_{[n]} := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

Equivariance: $Y_{[n]} = h(X_{[n]})$ **such that** $h(\pi \cdot X_{[n]}) = \pi \cdot h(X_{[n]})$ **for all** $\pi \in \mathbb{S}_n$**.**



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

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.

Statistical models and symmetry

Consider a sequence $X_{[n]} := (X_1, \ldots, X_n)$, $X_i \in \mathcal{X}$.

A *statistical model* of $X_{[n]}$ 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?

Exchangeable sequences

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

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

$X_{\mathbb{N}}$ is infinitely exchangeable if this is true for all prefixes $X_{[n]} \subset X_{\mathbb{N}}$, $n \in \mathbb{N}$.

de Finetti's theorem: $X_{\mathbb{N}} \iff X_i \mid Q \overset{\text{iid}}{\sim} Q$ **for some random** Q

Our models for $X_{\mathbb{N}}$ 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].

Finite exchangeable sequences

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

What else can we say?

The *empirical measure* of $X_{[n]}$ is

$$\mathbb{M}_{X_{[n]}}(\bullet) := \sum_{i=1}^{n} \delta_{X_i}(\bullet) \, .$$

Finite exchangeable sequences

The empirical measure is *sufficient*:

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

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

Finite exchangeable sequences

The empirical measure is *sufficient*:

$$P(X_{[n]} \in \bullet \mid \mathbb{M}_{X_{[n]}} = m) = \mathbb{U}_m(\bullet)\,,$$

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

The empirical measure is *adequate* for any Y such that $(\pi \cdot X_{[n]},\, Y) \stackrel{\mathrm{d}}{=} (X_{[n]},\, Y)$:

$$P(\,Y \in \bullet \mid X_{[n]} = x_{[n]}) = P(\,Y \in \bullet \mid \mathbb{M}_{X_{[n]}} = \mathbb{M}_{x_{[n]}}).$$

$\mathbb{M}_{X_{[n]}}$ contains all information in $X_{[n]}$ that is relevant for predicting Y.

A useful theorem

Suppose $X_{[n]}$ is an exchangeable sequence.

Invariance theorem:

$(\pi \cdot X_{[n]}, Y) \overset{\mathrm{d}}{=} (X_{[n]}, Y)$ for all $\pi \in \mathbb{S}_n$ if and only if

$$(X_{[n]}, Y) = (X_{[n]}, \tilde{h}(\eta, \mathbb{M}_{X_{[n]}})) \quad \text{a.s.},$$

with $\tilde{h}$ a measurable function and $\eta \sim \mathrm{Unif}[0,1]$, $\eta \perp\!\!\!\perp X_{[n]}$.

A useful theorem

Suppose $X_{[n]}$ is an exchangeable sequence.

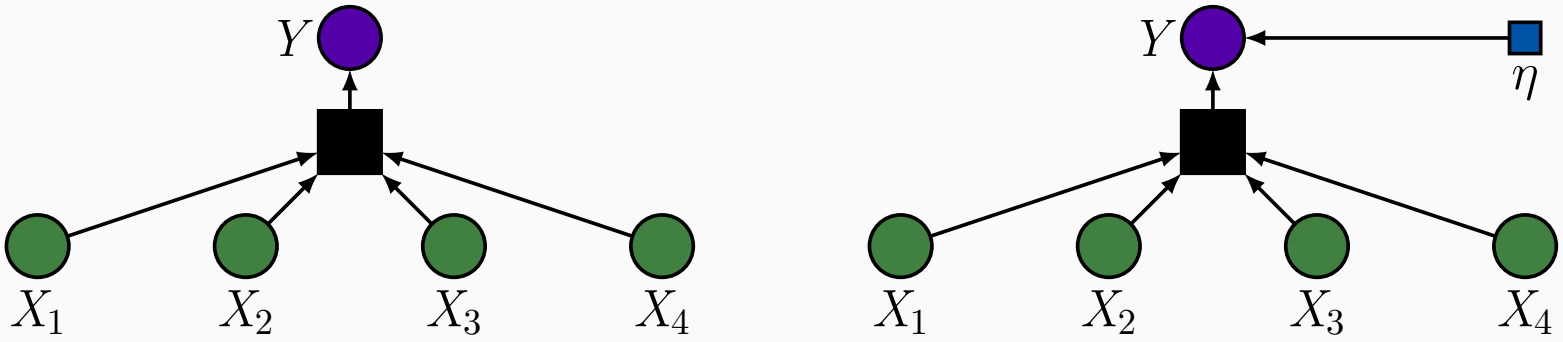
Invariance theorem:

$(\pi \cdot X_{[n]}, Y) \overset{\mathrm{d}}{=} (X_{[n]}, Y)$ for all $\pi \in \mathbb{S}_n$ if and only if

$$(X_{[n]}, Y) = (X_{[n]}, \tilde{h}(\eta, \mathbb{M}_{X_{[n]}})) \quad \text{a.s.},$$

with $\tilde{h}$ a measurable function and $\eta \sim \mathrm{Unif}[0,1]$, $\eta \perp\!\!\!\perp X_{[n]}$.

Deterministic invariance [Zah+17] $\mapsto$ stochastic invariance [this work]



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

Another useful theorem

Equivariance theorem:

$(\pi \cdot X_{[n]}, \pi \cdot Y_{[n]}) \overset{\mathrm{d}}{=} (X_{[n]}, Y_{[n]})$ for all $\pi \in \mathbb{S}_n$ if and only if

$$(X_{[n]}, Y_{[n]}) = \left(X_{[n]}, (\tilde{h}(\eta_i, X_i, \mathbb{M}_{X_{[n]}}))_{i \in [n]}\right) \quad \text{a.s.},$$

with $\tilde{h}$ a measurable function and i.i.d. $\eta_i \sim \mathrm{Unif}[0, 1]$, $\eta_i \perp\!\!\!\perp X_{[n]}$.

Another useful theorem

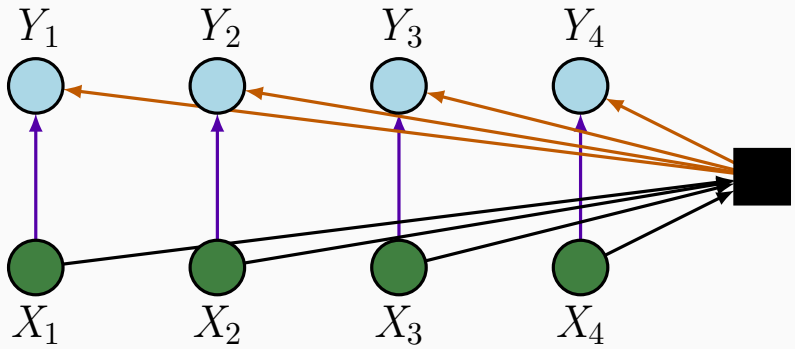
Equivariance theorem:

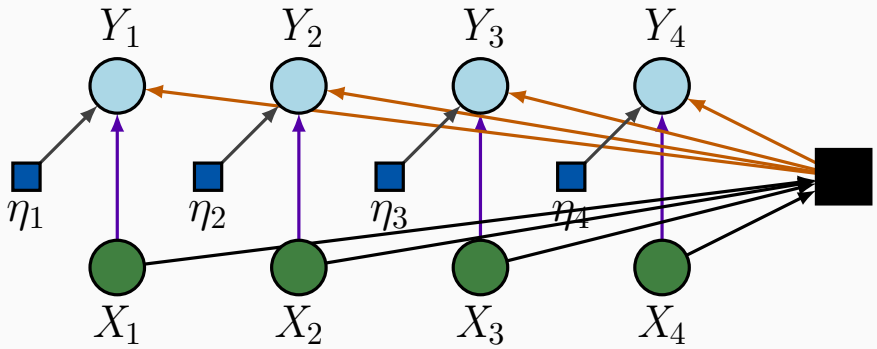
$(\pi \cdot X_{[n]}, \pi \cdot Y_{[n]}) \overset{\mathrm{d}}{=} (X_{[n]}, Y_{[n]})$ for all $\pi \in \mathbb{S}_n$ if and only if

$$(X_{[n]}, Y_{[n]}) = \left(X_{[n]}, (\tilde{h}(\eta_i, X_i, \mathbb{M}_{X_{[n]}}))_{i \in [n]}\right) \quad \text{a.s.},$$

with $\tilde{h}$ a measurable function and i.i.d. $\eta_i \sim \mathrm{Unif}[0,1]$, $\eta_i \perp\!\!\!\perp X_{[n]}$.

Deterministic equivariance [Zah+17] $\mapsto$ stochastic equivariance [this work]





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

$$= \sigma\left(w_0 X_i + w_1 \int_{\mathcal{X}} \sum_{j=1}^{n} \delta_{X_j}(dx)\right)$$

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

- For group symmetries that don't involve permutations—what are the analogous results? Equivariance is especially difficult.
- 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: *ICLR*. 2018. URL: https://openreview.net/pdf?id=Hkbd5xZRb.

[CW16] Taco 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: (Feb. 2018). eprint: 1802.05451. URL: https://arxiv.org/abs/1802.05451.

[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.

[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. URL: http://proceedings.mlr.press/v70/ravanbakhsh17a.html.

[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.

[Sha91] John Shawe-Taylor. “Threshold Network Learning in the Presence of Equivalences”. In: *Advances in Neural Information Processing Systems 4*. Ed. by J. E. Moody, S. J. Hanson, and R. P. Lippmann. Morgan-Kaufmann, 1991, pp. 879–886. URL: http://papers.nips.cc/paper/510-threshold-network-learning-in-the-presence-of-equivalences.pdf.

[Sha95] John Shawe-Taylor. “Sample Sizes for Threshold Networks with Equivalences”. In: *Information and Computation* 118.1 (1995), pp. 65–72. URL: http://www.sciencedirect.com/science/article/pii/S0890540185710528.

[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.

Symmetric neural networks

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

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