Statistical Foundations of Invariance and Equivariance in Deep Artificial Neural Network Learning

Yueyang Shen [1] Yupeng Zhang [2] Ivo Dinov [1]

[1] Statistical Online Computational Resource (SOCR), Computational Medicine and Bioinformatics, University of Michigan

[2]University of Wisconsin
petersyy@umich.edu
Slides can be found on SOCR news

March 18, 2025

www.SOCR.umich.edu

Overview

Review: Group representation theory

- Invariance and Equivariance: $\rho : G \to GL(V)$ is a group homomorphism $\rho(g_1 g_2) = \rho(g_1)\rho(g_2)$

  f is G- Invariant if $\quad f(\rho(\mathfrak{g})x) = f(x)$, $\quad f$ is G- Equivariant if $\quad f(\rho(\mathfrak{g})x) = \rho(\mathfrak{g})f(x) \quad \forall \mathfrak{g} \in G$

  ▶ Invariance requires information *compression* quotienting out symmetries, equivariance means information is *transformed* consistently.
- Group actions (in statistical contexts):
  1. Acting on group elements, G on G
  2. Acting on (statistical) parameters in $\mathbb{R}^d$, i.e., $T_g x$ (V finite dimensional, $\rho(g)$ invertible matrices)
     - ▶ Example: 1D Location Scale family, $T_g : \theta \to \theta_g = a\theta + b$
  3. Acting on functions f (Left regular representations), i.e., $L_g f(g') = f(g^{-1}g'), f \in \mathbb{L}_2(G)$[1] ($V$ infinite)
     - ▶ Example: Acting on Statistical estimators, $L_g : \hat{\Theta} \to \hat{\Theta}$, $L_g \hat{\theta}(\theta \mid x) = \hat{\theta}(T_g^{-1}\theta \mid x), \theta \in \mathbb{R}^d$
- Example - Spatial Rotational symmetry: $SO(3) = \{R^T R = I, det(R) = 1, R \in \mathbb{R}^{3\times 3}\}$
  1. Matrix composing with matrix (matrix product) defines G acting on G
  2. Matrix ($R = T_g = \rho(g)$) acting on $\mathbb{R}^3$ is trivial. $GL(V) = GL(3, \mathbb{R}) \equiv \mathbb{R}^{3\times 3}$
  3. $SO(3)$ acting on estimators acts on the parameters inversely.

[1] Can also be defined for other $\mathbb{L}_2$ spaces $L_g f(x) = f(T_g^{-1}x), f \in \mathbb{L}_2(\mathbb{R}^d), x \in \mathbb{R}^d$

Review: Deep Network Architecture

Two approaches to make deep network invariant/equivariant:

- Data Augmentation Limitation
- Architectural Design
  - $G-$invariant inference framework: several equivariant functions followed by a invariant layer.

Common architectural designs:

1. **MLP**: Universal function approximators, no symmetry built in, generalization contingent on training data distribution.
2. **CNN**: MLP with translational equivariance (segmentation)/invariance (classification). Equivariance realized via *translational weight sharing*.
3. **Discrete GCNN**: Data augmentation made implicit in the architectural design, discrete indexing g of the group G needed, *weight sharing across* G

$$f *_G K(g) = \sum_{h \in \mathbb{R}^n} f(h) K(T_g^{-1} h). \quad \text{Example: Scaling} : (f *_{\mathbb{R}_{>0}} K)(p, \lambda) = \sum_{q \in \mathbb{R}^2} f(p-q) K\left(\frac{1}{\lambda} q\right)$$

4. **Steerable CNN**: Does not need group sampling (discrete indexing) schemes, Information stored as Fourier coefficients (Peter-Weyl Theorem for compact group G) [11]

$$\text{Forward}: \hat{f}(\rho_\ell) = [\mathcal{F}_G f]_\ell = \int_G f(g)\rho_\ell(g)dg, \text{Backward}: \left[\mathcal{F}_G^{-1}\hat{f}\right]_\ell = \sum_\ell d_{\rho_\ell} tr\left[\hat{f}(\rho_\ell)\rho_\ell(g^{-1})\right], \tag{1}$$

Steerable kernels satisfy kernel constraints: $K(hx) = \rho_{out}(h)K(x)\rho_{in}(h^{-1})$

- $SO(3)$ example: $K(x) = \sum_{\ell=0}^{L}\sum_{m=-\ell}^{\ell} c_m^\ell(\|x\|)Y_m^\ell\left(\frac{x}{\|x\|}\right)$,
- Equivariance: $Y_m^\ell(R(\theta,\phi)) = \rho^\ell(R)Y_m^\ell(\theta,\phi), (\theta,\phi) \in S^2, R \in SO(3), \rho^\ell \in \mathbb{R}^{(2\ell+1)\times(2\ell+1)}$ are the Wigner-D matrices. $\rho^\ell = [D_{-\ell}^\ell, ..., D_{-1}^\ell, D_0^\ell, ..., D_\ell^\ell]$, $[D_m^\ell(\cdot)]_{m'} : SO(3) \to \mathbb{R}$ is Wigner-D function.

5. **Seq to Seq Transformers**: Non-convolutional Approach, attention mechanism is permutation equivariant, unlike MLP the model weights are *feature dependent* $w(X)$
   - Properties: Scaling laws [8], In context learning functional regression problem [6].

Overview

Relaxing the equivariance constraint

Motivation: Material Impurity (Non-isotropicity for ∇^2), physical non-ideality factors

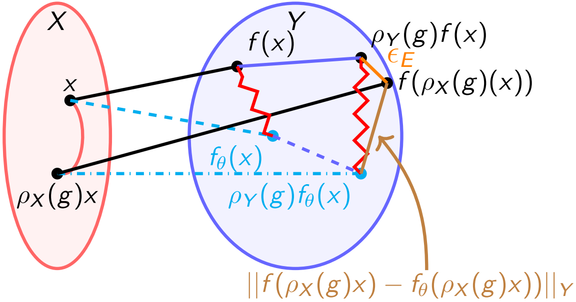


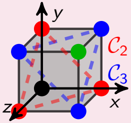
Figure: The problem with approximating an approximate G-equivariant function with G-equivariant function is that the two red zig zag lines cannot be simultaneously small. The solid lines stand for connections from G-equivariant (f_θ) inference. The dashed lines represent approximate G-equivariant (f) inferences.

- Approximate Equivariance[10]/Invariance : ϵ-approximate G-equivariant: $\forall\ g \in G$ and $\forall\ x \in X$, $\|f(\rho_X(g)(x)) - \rho_Y(g)f(x)\| \leq \epsilon_E$ ϵ-approximate G-invariant: $\forall\ g \in G$ and $\forall\ x \in X$, $\|f(\rho_X(g)(x)) - f(x)\| \leq \epsilon_I$
- Lower Bound Error for approximate equivariance inference with full equivariance parametrization[10]
  - f_θ denotes the NN based G-equivariant network and f be the approximate equivariant framework. Assuming the Lipschitz condition, $\|\rho_Y(g)f_\theta(x) -\rho_Y(g)f(x)\|_Y \leq \kappa\|f_\theta(x) - f(x)\|_Y$. Then, $\exists x$, $\|f_\theta(x) - f(x)\| \geq \frac{1}{1+\kappa}\epsilon_E$

Overview

Invariance, probabilistic symmetry and statistical inference

Discrete Case



$\mathcal{C}_1 = \{(0,0,0)\}, S_{\mathcal{C}_1}(x) = 0$

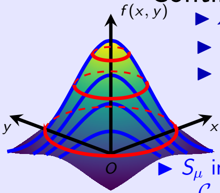
$\mathcal{C}_2 = \{(0,0,1),(0,1,0),(1,0,0)\}, S_{\mathcal{C}_2}(x) = 1$

$\mathcal{C}_3 = \{(0,1,1),(1,1,0),(1,0,1)\}, S_{\mathcal{C}_3}(x) = 2$

$\mathcal{C}_4 = \{(1,1,1)\}, S_{\mathcal{C}_4}(x) = 3$

- $X_1, X_2, X_3 \sim \text{Binomial}(\theta)$
- Sample space $\mathcal{X} = \{(x,y,z) \mid x,y,z \in \{0,1\}\}$
- (Minimal) Sufficient Statistic $S = x_1 + x_2 + x_3$
- $S(x)$ induced partition $\mathcal{X} = \mathcal{C}_1 \bigcup \mathcal{C}_2 \bigcup \mathcal{C}_3 \bigcup \mathcal{C}_4$

Continuous Case



- $X_1, X_2 \sim \mathcal{N}(\mu, \sigma^2)$
- Sample space: $\mathcal{X} = \mathbb{R}^2$
- (Minimal) sufficient statistics:
  Mean: $S_\mu = x_1 + x_2$
  Variance: $S_{\sigma^2} = x_1^2 + x_2^2$
- S_μ induced partition $\mathcal{X} = \bigcup \mathcal{C}_t$
  $\mathcal{C}_t = \{(x,y) \mid x + y = t\}_{t\in\mathbb{R}}$
- S_{σ^2} induce $\mathcal{X} = \bigcup_{t\in\mathbb{R}^+} \mathcal{C}_t$, $\mathcal{C}_t = \{(x,y) \mid x^2 + y^2 = t\}$

Figure: Discrete and continuous case of sample space partition induced by sufficient statistics. (Left): The sufficient statistic generates a partition of black, red, blue green dots. (Right): The sufficient statistic generates a partition of red isocontours for the variance and blue isocontours for the mean parameter.

- *Probabilistic Symmetry* is defined on random structures X_∞ (random variables, random graphs, random partitions,...). A random structure is *symmetric* to G if $g(X) \overset{d}{=} X, \forall X \in X_\infty, g \in G$. The canonical example being exchangeability [2].
- Sufficiency describe information relevant to inference, Invariance introduces irrelevance and needs to be quotient out.

Overview

Metric Measure Symmetry & Measuring symmetry

Common symmetries in metric measures:

- Reparametrization symmetry.
  - Physical properties (Curve length, Regional areas, Solid volumes) is independent of coordinate transformations.
  - Canonical variable transform is coupled with a Jacobian term (r.v. Bivariate transform $f_{UV}(u, v) = f_{XY}(x(u, v), y(u, v))|J|$). When the $|J|$ factor is absorbed, the quantity is reparametrization invariant (Fisher information, Mutual information).
- Geometrical Transformation Symmetry
  - Rotations (Cosine similarity, L2 logistic regression), Affine (Amari-Chentsov tensor[1])
- Problem specific symmetries:
  - Optimal policy invariance under reward shaping $\tilde{R} = R + F(x, a, x') = R + \gamma\phi(x') - \phi(x)$: This non-classical invariance is generated from the Bellman objective function form.

Measuring Symmetry

One can use Lie derivative to quantify how much symmetry is aligned/violated (Locally) by rearranging the equivariance condition: [7]

$$\rho_{21}(g)[f](x) = \rho_2(g)^{-1} f(\rho_1(g))(x) \tag{2}$$

The Lie derivative generated by a vector field Y can be expanded using the rewritten condition

$$\mathcal{L}_Y(f) = \lim_{t \to 0} \frac{\rho_{21}(\phi_Y^t)[f] - f}{t} = \lim_{t \to 0} \frac{\psi^*_{exp(-tY)} \circ f \circ \psi_{exp(tY)} - f}{t} \tag{3}$$

- ϕ_Y^t is the local 1-parameter group generated by Y (flowing along the vector field Y with time t)
- $\psi_{exp(tY)} : \mathcal{M} \to \mathcal{M}$ is the manifold pushforward defined by the group action
- $\psi^*_{exp(-tY)} : T^*_{\phi_Y^t(p)}\mathcal{M} \to T^*_p\mathcal{M}$ is the pullback of the cotangent space. Namely, it pulls back the cotangent space at $\phi_Y^t(p)$ to p

Optimization Practice [3]

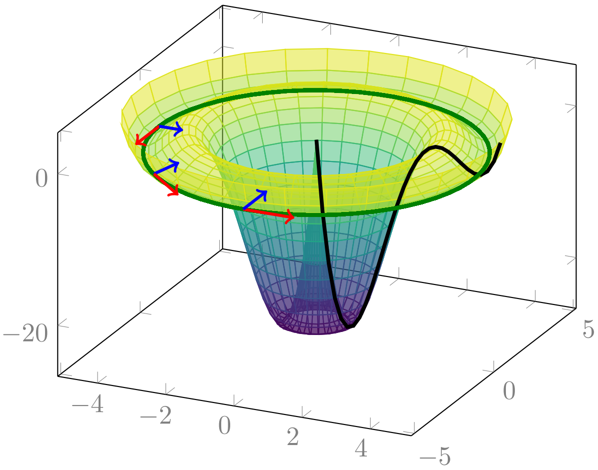


Figure: An illustration of a non-convex loss landscape with radial symmetry. $\mathcal{M}$: surface, $\mathcal{M}/G$: black curve

Symmetries on the functional landscape often entails non-convexity. In terms of optimizing on the total space $\mathcal{M}$ or quotient space $\mathcal{M}/G$

- For *first order* Riemannian gradient descent method, there is no difference utilizing the quotient structure or using the algorithm in the original space.
- For *second order methods*, Newton's method would be catastrophic for optimizing the loss in the original space $\mathcal{M}$, since Newton's method solves step direction in one shot.
- Using conjugate gradient to minimize the second order expansion mitigates the problem of solving an underdetermined system when optimizing in original space $\mathcal{M}$.

Overview

1. Mathematical Foundations
   - Review: Group (Representation) Theory, Deep Network Architectures
   - Relaxing the exact G-equivariant condition
2. Statistical and Optimization practice under symmetry
   - Invariance, probabilistic symmetry and statistical inference
   - Optimization symmetry practice
3. **Biomedical Applications & Case Studies**
   - Case Study: Group Invariance Case Study on Kreuzer Skarke Dataset
   - Case Study: fMRI music dataset
4. References

Overview

Group Invariant Learning on Kreuzer Skarke Dataset

Work with Christian Ewert, Sumner Magruder, Vera Maiboroda, Pragya Singh, and Daniel Platt. [2]

- Problem: Regression $\mathbb{R}^{4\times 26} \to \mathbb{Z}_+$
  - Symmetry Group: $S_4 \times S_{26}$.
  - Cardinality: $4! \times 26! = 9.7 \times 10^{27}$

  Data Augmentation impossible Arch-review
- Models: CNN, Xgboost, Invariant MLP, (Vision) Transformer, PointNet++, MLP with invariant features
- Data Preprocessing: Original, Original (Random) Permuted, Preprocessed, Preprocessed Permuted
- Main Findings:
  1. Approximately Invariant models outperform fully invariant models
  2. Group Invariant Preprocessing improves performance
  3. Building group invariance improves performance

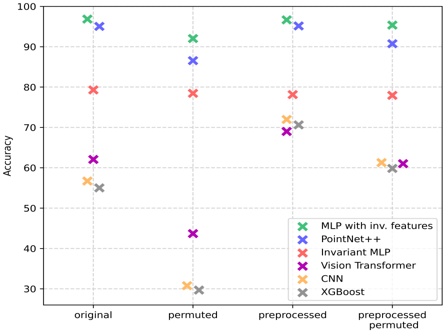
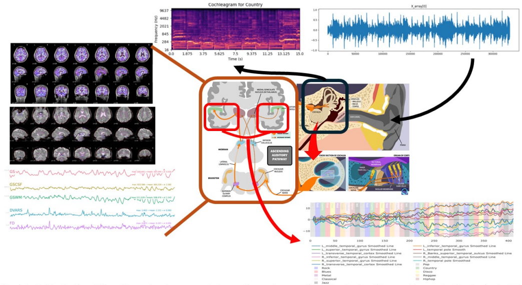


Figure: Different Architectures across group invariant preprocessing

[2] Christian Ewert et al. "Group-invariant machine learning on the Kreuzer-Skarke dataset". In: *Physics Letters B* 858 (2024), p. 138996

Overview

Auditory Neuroscience, Physiology, AI applications



Physiology illustration: https://pressbooks.umn.edu/sensationandperception/chapter/auditory-pathways-to-the-brain-draft/

Figure: The biological pathway and the signal processing & ML aspect of music melody recognition

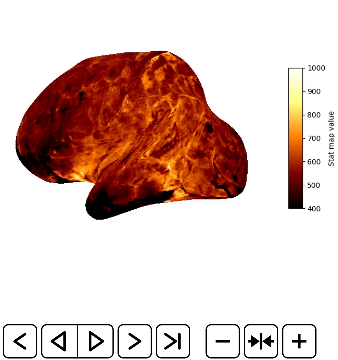


Figure: Animation of the music melody brain response

Biomedical Applications - fMRI music genre dataset

Heterogeneity data challenge:

- Music genre heterogeneity: music clip selected from GTZAN randomly clipped from the genre
- Subject heterogeneity:
  1. Brain Spatial Configuration variation
  2. Individual temporal experience/perception variation

▶Goal: Coexplain music stimulus and brain response

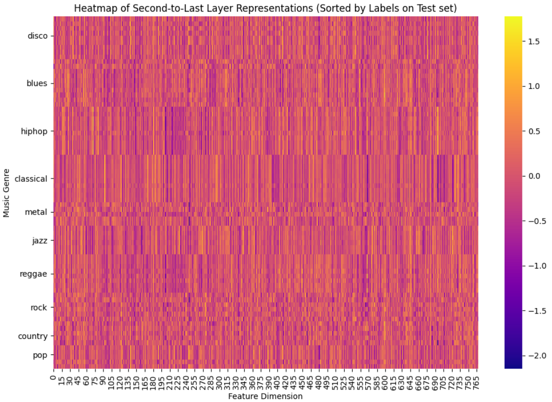


Figure: Homologies captured from finetuned pretrained speech transformer model (Distil-Hubert [4]) visualized on test set not seen in training

Biomedical Applications - fMRI music genre dataset

Biomedical Datasets demonstrate various classical and non-classical invariances

- fMRI preprocessing (e.g., registering the hypervolumetric data into a common 3D/4D spatiotemporal) atlas space to align the fMRI data and facilitate a form of inference invariance can be regarded as the "group invariant" preprocessing step.

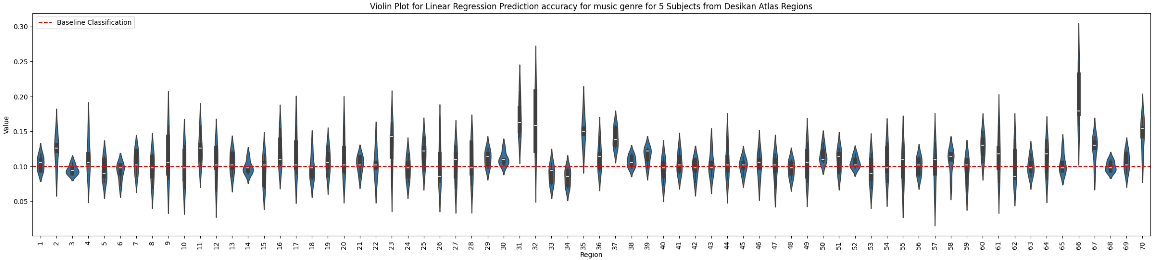


Figure: Different ROIs parcellated from Desikan atlas where the voxels are regressed on the training and statistics collected on the test set. Region-31 is Left Superior Temporal Gyrus. Region-32 is Left Supramarginal Gyrus. Region-35 is Left Transverse Temporal Cortex. Region-66 is right superior temporal Gyrus. Region-70 is the right transverse temporal cortex.

Predicting the Genre using brain spatiotemporal state tensor
$\mathbb{R}^{96\times 96\times 68\times 10} \to \{1, 2, .., 10\}$, where $96 \times 96 \times 68$ corresponds to the spatial dimensions and the 10 corresponds to the discrete time sampling points, and the range is indicated by the 10 music genres.

Models	Numerics	Invariance
Random Guessing	10%	No
Music Classification[3]	$73.24\% \pm 7.96\%$	Spectral Translation
Finetuning distilHubert	$72.07\%\pm3.40\%$	
k-NN[4]	$1.86\%\pm 0.73\%$	No
CNN on raw data	$\leq 15\%$	3D subspace translation
ROI-66 + Linear Regression	19.9%[5]	
ROI-66 + ML model	25.2%[6]	

[3] Caifeng Liu et al. "Bottom-up broadcast neural network for music genre classification". In: *Multimedia Tools and Applications* 80 (2021), pp. 7313–7331

[4] Selected among 1,3,5,7,9

[5] Averaged over five individuals

[6] Averaged over five individuals

Acknowledgements

SOCR Funding sources:
NIH: UL1 TR002240, R01 CA233487, R01 MH121079, R01 MH126137, **T32 GM141746 (BIDS-TP)**
NSF: 1916425, 1734853, 1636840, 1416953, 0716055, 1023115
SOCR Members: Zerihun Bekele, Milen Velev, Kaiming Cheng, Achu Shankar, Ryan Kwon, Shihang Li, Daxuan Deng, Zijing Li, Yongkai Qiu, Zhe Yin, Yufei Yang, Yuxin Wang, Rongqian Zhang, Yuyao Liu, Yupeng Zhang, Yunjie Guo, Simeone Marino, Ivo Dinov

www.SOCR.umich.edu

References

[1] Nihat Ay et al. "Information geometry and sufficient statistics". In: *Probability Theory and Related Fields* 162 (2015), pp. 327–364.

[2] Benjamin Bloem-Reddy, Yee Whye, et al. "Probabilistic symmetries and invariant neural networks". In: *Journal of Machine Learning Research* 21.90 (2020), pp. 1–61.

[3] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[4] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7087–7091.

[5] Christian Ewert et al. "Group-invariant machine learning on the Kreuzer-Skarke dataset". In: *Physics Letters B* 858 (2024), p. 138996.

[6] Shivam Garg et al. "What can transformers learn in-context? a case study of simple function classes". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 30583–30598.

[7] Nate Gruver et al. "The lie derivative for measuring learned equivariance". In: *arXiv preprint arXiv:2210.02984* (2022).

[8] Jared Kaplan et al. "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (2020).

[9] Caifeng Liu et al. "Bottom-up broadcast neural network for music genre classification". In: *Multimedia Tools and Applications* 80 (2021), pp. 7313–7331.

[10] Rui Wang, Robin Walters, and Rose Yu. "Approximately equivariant networks for imperfectly symmetric dynamics". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 23078–23091.

[11] Maurice Weiler et al. "3d steerable cnns: Learning rotationally equivariant features in volumetric data". In: *Advances in Neural Information Processing Systems* 31 (2018).