Gyrogroup Batch Normalization

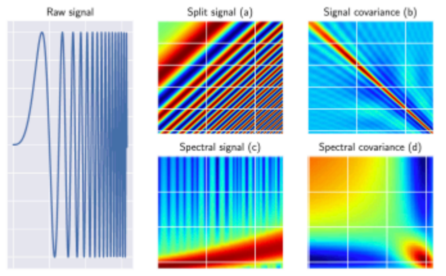
Ziheng Chen, Yue Song, Xiao-Jun Wu, Nicu Sebe

Homepage LinkIn

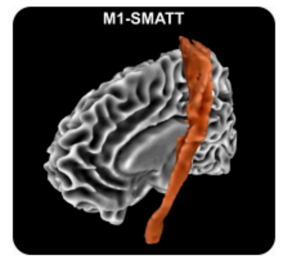
ICLR International Conference On Learning Representations

Motivation

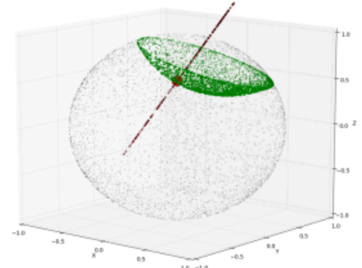
Manifold-valued Measurements in Applications



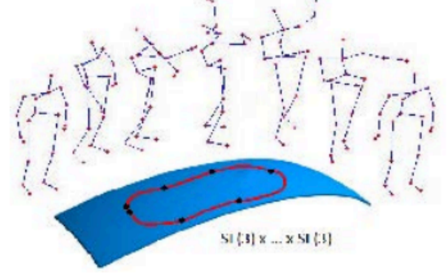
Radar Classification (Brooks, 2020)



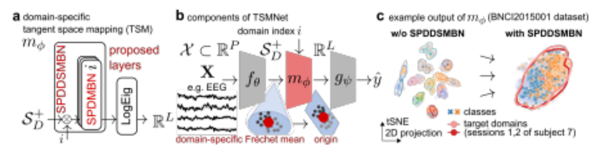
Medical Imaging (Chakraborty et al., 2020)



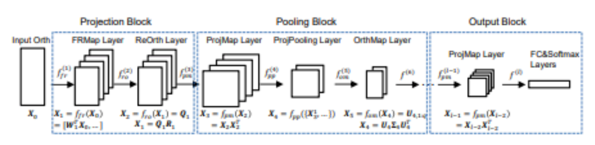
NLP, Graph... (Ganea et al., 2018)



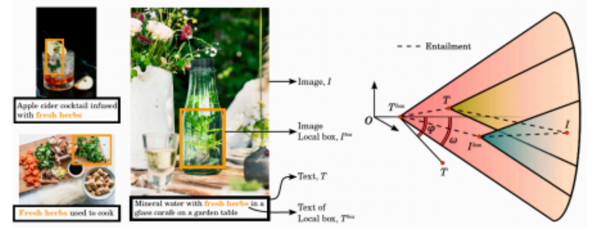
Action Recognition (Vemulapalli et al., 2014)



EEG (Kobler et al., 2022a)



GrNet (Huang et al., 2018)



Vision-Language Models (Huang et al., 2018)

Euclidean Normalization: controlling mean and variance

$$\forall i \leq N, x_i \leftarrow \gamma \frac{x_i - \mu_b}{\sqrt{v_b^2 + \epsilon}} + \beta$$

- Centering
- Biasing
- Scaling

Existing Riemannian Normalization:

- Fails to normalize statistics in a general manner

Method	Controllable Statistics	Applied Geometries	Incorporated by GyroBN
SPDBN (Brooks et al., 2019)	M	SPD manifolds under AIM	✓
SPDBN (Kobler et al., 2022b)	M+V	SPD manifolds under AIM	✓
SPDDSMBN (Kobler et al., 2022a)	M+V	SPD manifolds under AIM	✓
ManifoldNorm (Chakraborty, 2020, Algs. 1-2)	N/A	Riemannian homogeneous space	✗
ManifoldNorm (Chakraborty, 2020, Algs. 3-4)	M+V	Matrix Lie groups under the distance $d(X,Y) = \lVert \mathrm{mlog}(X^{-1}Y) \rVert$	✓
RBN (Lou et al., 2020, Alg. 2)	N/A	Geodesically complete manifolds	✗
LieBN (Chen et al., 2024b)	M+V	General Lie groups	✓
GyroBN	M+V	Pseudo-reductive gyrogroups with gyro isometric gyrations	N/A

Gyro Structures

Definition 2.1 (Gyrogroups (Ungar, 2009)). Given a nonempty set G with a binary operation $\oplus: G \times G \to G$, $\{G, \oplus\}$ forms a gyrogroup if its binary operation satisfies the following axioms for any $a, b, c \in G$:

(G1) There is at least one element $e \in G$ called a left identity (or neutral element) such that $e \oplus a = a$.

(G2) There is an element $\ominus a \in G$ called a left inverse of a such that $\ominus a \oplus a = e$.

(G3) There is an automorphism $\mathrm{gyr}[a,b]: G \to G$ for each $a, b \in G$ such that

$$a \oplus (b \oplus c) = (a \oplus b) \oplus \mathrm{gyr}[a,b]c \quad \text{(Left Gyroassociative Law).} \tag{1}$$

The automorphism $\mathrm{gyr}[a,b]$ is called the gyroautomorphism, or the gyration of G generated by a, b.
(G4) Left reduction law: $\mathrm{gyr}[a,b] = \mathrm{gyr}[a \oplus b, b]$.

$$\text{Gyro addition: } P \oplus Q = \mathrm{Exp}_P\left(\mathrm{PT}_{E \to P}\left(\mathrm{Log}_E(Q)\right)\right),$$
$$\text{Gyro scalar product: } t \odot P = \mathrm{Exp}_E\left(t \,\mathrm{Log}_E(P)\right),$$
$$\text{Gyro inner product: } \langle P, Q \rangle_{\mathrm{gr}} = \langle \mathrm{Log}_E(P), \mathrm{Log}_E(Q) \rangle_E,$$
$$\text{Gyro norm: } \lVert P \rVert_{\mathrm{gr}} = \langle P, P \rangle_{\mathrm{gr}},$$
$$\text{Gyrodistance: } \mathrm{d}_{\mathrm{gry}}(P, Q) = \lVert \ominus P \oplus Q \rVert_{\mathrm{gr}},$$

$$M = \mathrm{FM}(\{P_i\}) = \underset{Q \in \mathcal{M}}{\mathrm{argmin}} \frac{1}{N} \sum_{i=1}^{N} \mathrm{d}_{\mathrm{gry}}^2 (P_i, Q)$$

Examples

Geometry	Symbol	$P \oplus Q$ or $x \oplus y$	E	$\ominus P$ or $\ominus x$	Lie group	Gyrogroup	References
AIM $\mathcal{S}^n_{++}$	$\oplus^{\mathrm{AI}}$	$P^{\frac{1}{2}} Q P^{\frac{1}{2}}$	I_n	P^{-1}	✗	✓	(Nguyen, 2022b)
LEM $\mathcal{S}^n_{++}$	$\oplus^{\mathrm{LE}}$	$\mathrm{mexp}(\mathrm{mlog}(P) + \mathrm{mlog}(Q))$	I_n	P^{-1}	✓	✓	(Arsigny et al., 2005) (Nguyen, 2022b)
LCM $\mathcal{S}^n_{++}$	$\oplus^{\mathrm{LC}}$	$\psi_{\mathrm{LC}}^{-1}(\psi_{\mathrm{LC}}(P) + \psi_{\mathrm{LC}}(Q))$	I_n	$\psi_{\mathrm{LC}}(-\psi_{\mathrm{LC}}(P))$	✓	✓	(Lin, 2019) (Nguyen, 2022b) (Chen et al., 2024e)
$\widetilde{\mathrm{Gr}}(p,n)$ / $\mathrm{Gr}(p,n)$	$\widetilde{\oplus}^{\mathrm{Gr}}$ / $\oplus^{\mathrm{Gr}}$	$\mathrm{mexp}(\Omega) Q \,\mathrm{mexp}(-\Omega)$ / $\mathrm{mexp}(\Omega) V$	$\widetilde{I}_{p,n}$ / $I_{p,n}$	$\mathrm{mexp}(-\Omega) \widetilde{I}_{p,n} \,\mathrm{mexp}(\Omega)$ / $\mathrm{mexp}(-\Omega) I_{p,n}$	✗	Non-reductive	(Nguyen, 2022a) (Nguyen & Yang, 2023)
$\mathcal{M}_K$	$\oplus_K$	$\frac{(1-2K\langle x,y\rangle - K\lVert y \rVert^2)x + (1+K\lVert x \rVert^2)y}{1-2K\langle x,y\rangle + K^2 \lVert x \rVert^2 \lVert y \rVert^2}$	0	$-x$	✗(✓for K=0)	✓	(Ungar, 2009) (Ganea et al., 2018) (Skopek et al., 2019)

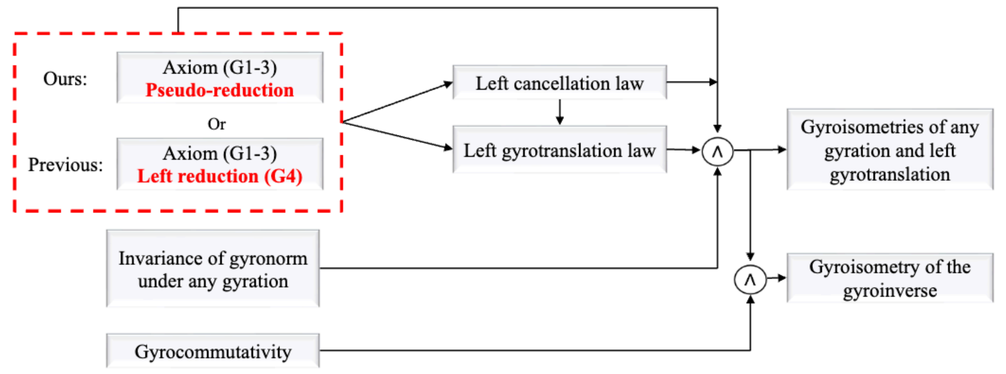
Pseudo-reductive Gyrogroups

Definition

Definition 3.1 (Pseudo-reductive Gyrogroups). A groupoid $\{G, \oplus\}$ is a pseudo-reductive gyrogroup if it satisfies axioms (G1), (G2), (G3) and the following pseudo-reductive law:

$$\mathrm{gyr}[X, P] = \mathbb{1}, \text{ for any left inverse } X \text{ of } P \text{ in } G, \tag{12}$$

where $\mathbb{1}$ is the identity map.



Theorem D.1 (First Pseudo-reductive Gyrogroups Properties). *Let $\{G, \oplus\}$ be a pseudo-reductive gyrogroup. For any elements $P, Q, R, X \in G$, we have:*

1. *If $P \oplus Q = P \oplus R$, then $Q = R$ (General Left Cancellation law; see (9) below).*
2. $\mathrm{gyr}[E, P] = \mathbb{1}$ *for any left identity E in G.*
3. $\mathrm{gyr}[X, P] = \mathbb{1}$ *for any left inverse X of P in G.*
4. *There is P left identity which is P right identity.*
5. *There is only one left identity.*
6. *Every left inverse is P right inverse.*
7. *There is only one left inverse, $\ominus P$, of P, and $\ominus(\ominus P) = P$.*
8. *The left cancellation law: $\ominus P \oplus (P \oplus Q) = Q$.*
9. *The gyrator identity:* $\mathrm{gyr}[P, Q]X = \ominus(P \oplus Q) \oplus \{P \oplus (Q \oplus X)\}$.
10. $\mathrm{gyr}[P, Q]E = E$.
11. $\mathrm{gyr}[P, Q](\ominus X) = \ominus \mathrm{gyr}[P, Q]X$.
12. $\mathrm{gyr}[P, E] = \mathbb{1}$.
13. *The gyrosum inversion law:* $\ominus(P \oplus Q) = \mathrm{gyr}[P, Q](\ominus Q \ominus P)$

Theories

Proposition 3.2. [↓] *$\mathrm{Gr}(p,n)$ and $\widetilde{\mathrm{Gr}}(p,n)$ form pseudo-reductive and gyrocommutative gyrogroups.*

Proposition 3.6. [↓] *For every (pseudo-reductive) gyrogroup in Tab. 2, the gyrodistance is identical to the geodesic distance (therefore symmetric). The gyroinverse, any gyration and any left gyrotranslation are gyroisometries.*

GyroBN

$$\forall i \leq N, x_i \leftarrow \gamma \frac{x_i - \mu_b}{\sqrt{v_b^2 + \epsilon}} + \beta \quad \xrightarrow{\text{Gyro}} \quad \forall i \leq N, \tilde{P}_i = \overbrace{B \oplus}^{\text{Biasing}} \left(\overbrace{\frac{s}{\sqrt{v^2 + \epsilon}} \odot}^{\text{Scaling}} \left(\overbrace{\ominus M \oplus P_i}^{\text{Centering}} \right) \right)$$

Properties

Theorem 4.1 (Homogeneity). [↓] *Supposing $\{\mathcal{M}, \oplus\}$ is a pseudo-reductive gyrogroup with any gyration $\mathrm{gyr}[\cdot,\cdot]$ as a gyroisometry, for N samples $\{P_{i \ldots N} \in \mathcal{M}\}$, we have the following properties:*

$$\textit{Homogeneity of gyromean: } \mathrm{FM}(\{B \oplus P_i\}) = B \oplus \mathrm{FM}(\{P_i\}), \forall B \in \mathcal{M}, \tag{16}$$

$$\textit{Homogeneity of dispersion from } E\text{: } \frac{1}{N} \sum_{i=1}^{N} \mathrm{d}_{\mathrm{gry}}^2 (t \odot P_i, E) = \frac{t^2}{N} \sum_{i=1}^{N} \mathrm{d}_{\mathrm{gry}}^2 (P_i, E), \tag{17}$$

Algorithm

Recovers
- Euc. BN
- LieBN
- ....

Algorithm 1: Gyrogroup Batch Normalization (GyroBN)

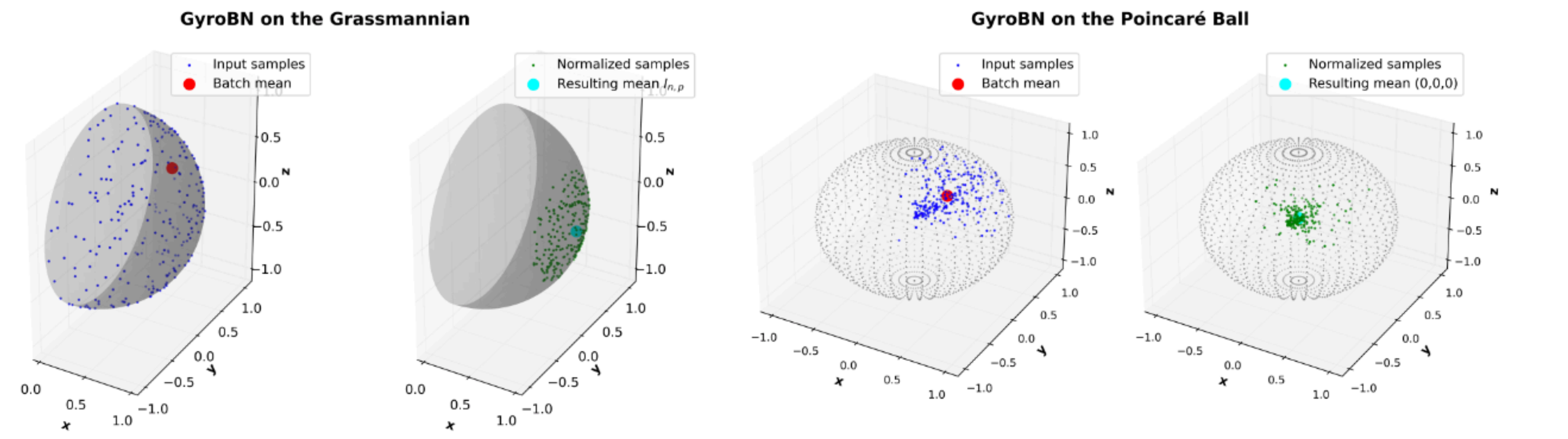
Require : batch of activations $\{P_{1 \ldots N} \in \mathcal{M}\}$, small positive constant ϵ, and momentum $\eta \in [0, 1]$, running mean M_r, running variance v_r^2, biasing parameter $B \in \mathcal{M}$, scaling parameter $s \in \mathbb{R}$.

Return : normalized batch $\{\tilde{P}_{1 \ldots N} \in \mathcal{M}\}$

```
if training then
    Compute batch mean M_b and variance v_b^2 of {P_1...N};
    Update running statistics M_r = Bar_γ(M_b, M_r), v_r^2 = γ v_b^2 + (1 − γ) v_r^2;
end
(M, v^2) = (M_b, v_b^2) if training else (M_r, v_r^2)
∀i ≤ N, P̃_i = B ⊕ ( s/√(v^2+ε) ⊙ (⊖M ⊕ P_i) )
```

Manifestations

Operator	$\mathrm{Gr}(p,n)$	$\mathbb{P}^n_K$
Identity element	$I_{p,n}$	$0 \in \mathbb{R}^n$
$P \oplus^{\mathrm{Gr}} Q$ or $x \oplus_K y$	$\mathrm{mexp}(\Omega) V$	$\frac{(1-2K\langle x,y\rangle - K\lVert y \rVert^2)x + (1+K\lVert x \rVert^2)y}{1-2K\langle x,y\rangle + K^2 \lVert x \rVert^2 \lVert y \rVert^2}$
$\ominus^{\mathrm{Gr}} P$ or $\ominus_K x$	$\mathrm{mexp}(-\Omega) I_{p,n}$	$-x$
$t \odot^{\mathrm{Gr}} P$ or $t \odot_K x$	$\mathrm{mexp}(t\Omega) I_{p,n}$	$\frac{1}{\sqrt{\lvert K \rvert}} \tanh\left(t \tanh^{-1}(\sqrt{\lvert K \rvert} \lVert x \rVert)\right) \frac{x}{\lVert x \rVert}$
$\mathrm{Bar}^{\mathrm{Gr}}_\gamma(Q, P)$ or $\mathrm{Bar}^K_\gamma(y, x)$	$\mathrm{Exp}^{\mathrm{Gr}}_P(\gamma \,\mathrm{Log}^{\mathrm{Gr}}_P(Q))$	$x \oplus_K (-x \oplus_K y) \odot_K t$
Fréchet Mean	Karcher Flow (Karcher, 1977)	(Lou et al., 2020, Alg. 1)



Experiments

Gras.

Table 3: Comparison of GyroBN against other Grassmannian BNs under GyroGr backbone.

BN	None		ManifoldNorm-Gr		RBN-Gr		GyroBN-Gr	
Acc.	Mean±std	Max	Mean±std	Max	Mean±std	Max	Mean±std	Max
HDM05	48.97±0.24	49.23	49.67±0.76	50.41	48.64±0.77	49.49	**51.89±0.37**	**52.43**
NTU60	70.13±0.16	70.32	68.56±0.43	69.14	67.77±0.52	68.35	**72.60±0.04**	**72.65**
NTU120	53.76±0.18	53.96	51.41±0.38	51.92	50.56±0.22	50.82	**55.47±0.10**	**55.59**

Table 4: Ablation of Grassmannian GyroBN under various network architectures.

	HDM05				NTU60				NTU120			
Architecture	1Block	2Block	3Block	4Block	1Block	2Block	3Block	4Block	1Block	2Block	3Block	4Block
GyroGr	49.23	49.09	47.02	27.36	70.32	70.14	70.23	65.03	53.96	54.1	54.59	47.59
GyroGrBN	**52.43**	**50.62**	**51.56**	**30.29**	**72.65**	**71.93**	**72.25**	**66.67**	**55.59**	**56.15**	**54.63**	**48.9**

Poincaré

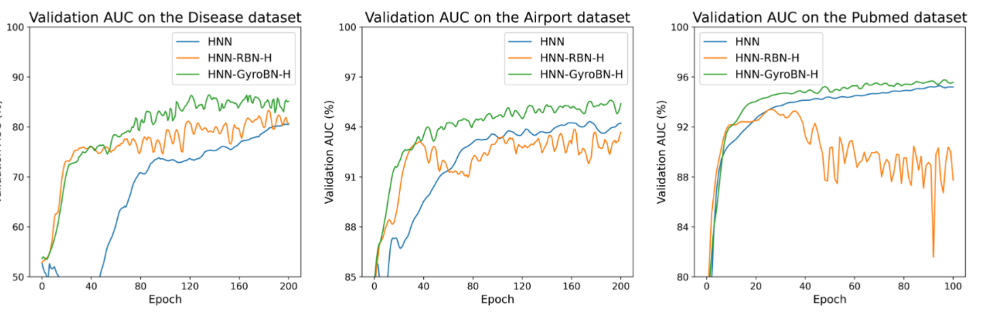


Table 5: Comparison of HNN with or without GyroBN-H or RBN-H on the link prediction task.

Dataset	HNN	HNN-RBN-H	HNN-GyroBN-H
Cora	89.0 ± 0.1	93.5 ± 0.5	**94.3 ± 0.2**
Disease	75.1 ± 0.3	76.6 ± 2.2	**81.2 ± 0.9**
Airport	90.8 ± 0.2	94.2 ± 0.4	**95.4 ± 0.2**
Pubmed	94.9 ± 0.1	93.4 ± 0.2	**95.8 ± 0.1**

Efficiency

Methods	HDM05 (47 × 10)	NTU60 (75 × 10)	NTU120 (75 × 10)
GyroGr	2.19	50.92	80.72
GyroGr-ManifoldNorm	4.98	242.12	409.48
GyroGr-RBN	5.16	242.63	410.08
GyroGr-GyroBN	3.10	59.55	108.92

Methods	Cora	Disease	Airport	Pubmed
HNN	0.0323	0.0271	0.054	0.1253
HNN-RBN-H	0.0905	0.0883	0.1215	0.3416
HNN-GyroBN-H	0.0757	0.0842	0.119	0.3351