

A Lie Group Approach to Riemannian Batch Normalization

Ziheng Chen, Yue Song, Yunmei Liu, and Nicu Sebe

Motivation

Euclidean Batch Normalization:

Facilitating network training by controlling mean and variance

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

Existing Riemannian Batch Normalization:

Fails to normalize statistics in a general manner

Table 2: Summary of some representative RBN methods.

Methods	Involved Statistics	Controllable Mean	Controllable Variance	Application Scenarios
SPDBN [Brooks et al., 2019b]	Mean	✓	N/A	SPD manifolds under AIM
SPDBN [Kobler et al., 2022b]	Mean+Variance	✓	✓	SPD manifolds under AIM
Chakraborty [2020, Algs. 1-2]	Mean+Variance	✗	✗	Riemannian homogeneous space
Chakraborty [2020, Algs. 3-4]	Mean+Variance	✓	✓	A certain Lie group structure and distance
RBN [Lou et al., 2020, Alg. 2]	Mean+Variance	✗	✗	Geodesically complete manifolds
Ours	Mean+Variance	✓	✓	General Lie groups

Preliminaries

Lie Groups and Pullback

Definition 2.1 (Lie Groups). A manifold is a Lie group, if it forms a group with a group operation \odot such that $m(x, y) \mapsto x \odot y$ and $i(x) \mapsto x^{-1}$ are both smooth, where x^{-1} is the group inverse.

Definition 2.2 (Left-invariance). A Riemannian metric g over a Lie group $\{G, \odot\}$ is left-invariant, if for any $x, y \in G$ and $V_1, V_2 \in T_x M$,

$$g_y(V_1, V_2) = g_{L_x(y)}(L_{x*,y}(V_1), L_{x*,y}(V_2)), \quad (1)$$

where $L_x(y) = x \odot y$ is the left translation by x , and $L_{x*,y}$ is the differential map of L_x at y .

Definition 2.3 (Pullback Metrics). Suppose $\mathcal{M}_1, \mathcal{M}_2$ are smooth manifolds, g is a Riemannian metric on \mathcal{M}_2 , and $f : \mathcal{M}_1 \rightarrow \mathcal{M}_2$ is smooth. Then the pullback of g by f is defined point-wisely,

$$(f^*g)_p(V, W) = g_{f(p)}(f_{*,p}(V), f_{*,p}(W)), \quad (2)$$

where $p \in \mathcal{M}$, $f_{*,p}(\cdot)$ is the differential map of f at p , and $V, W \in T_p \mathcal{M}$. If f^*g is positive definite, it is a Riemannian metric on \mathcal{M}_1 , called the pullback metric defined by f .

Geometries on the SPD Manifold

Table 1: Lie group structures and the associated Riemannian operators on SPD manifolds.

Metric	(α, β)-LEM	(α, β)-AIM	LCM
$g_P(V, W)$	$\langle \text{mlog}_{*,P}(V), \text{mlog}_{*,P}(W) \rangle^{(\alpha, \beta)}$	$\langle P^{-1}V, WP^{-1} \rangle^{(\alpha, \beta)}$	$\sum_{i>j} V_{ij} W_{ij} + \sum_{j=1}^n V_{jj} W_{jj} L_j^{-2}$
$d(P, Q)$	$\ \text{mlog}(P) - \text{mlog}(Q)\ ^{(\alpha, \beta)}$	$\left\ \text{mlog}\left(Q^{-\frac{1}{2}}PQ^{-\frac{1}{2}}\right) \right\ ^{(\alpha, \beta)}$	$\ \psi_{\text{LC}} \circ \text{Chol}(P) - \psi_{\text{LC}} \circ \text{Chol}(Q)\ _{\text{F}}$
$Q \odot P$	$\text{mexp}(\text{mlog}(P) + \text{mlog}(Q))$	KPK^{\top}	$\text{Chol}^{-1}([L + K] + \mathbb{K}L)$
$\text{FM}\{P_i\}$	$\text{mexp}\left(\frac{1}{n} \sum_i \text{mlog} P_i\right)$	Karcher Flow	$\psi_{\text{LC}}^{-1}\left(\frac{1}{n} \sum_i \psi_{\text{LC}}(P_i)\right)$
$\text{Log}_P Q$	$(\text{mlog}_{*,P})^{-1}[\text{mlog}(Q) - \text{mlog}(P)]$	$P^{\frac{1}{2}} \text{mlog}\left(P^{-\frac{1}{2}}QP^{-\frac{1}{2}}\right) P^{\frac{1}{2}}$	$(\text{Chol}^{-1})_{*,L}[[K] - [L] + \mathbb{L} \text{Dlog}(\mathbb{L}^{-1}K)]$
Invariance	Bi-invariance	Left-invariance	Bi-invariance

Geometry on the Rotation Matrix

Table 8: The associated Riemannian operators on Rotation matrices.

Operators	$d^2(R, S)$	$\text{Log}_I R$	$\text{Exp}_I(A)$	$\gamma_{(R, S)}(t)$	FM
Expression	$\ \text{mlog}(R^T S)\ _{\text{F}}^2$	$\text{mlog}(R)$	$\text{mexp}(A)$	$R \text{mexp}(t \text{mlog}(R^T S))$	Manton [2004, Alg. 1]



UNIVERSITÀ
DI TRENTO

UNIVERSITY OF
LOUISVILLE



International Conference On
Learning Representations



LieBN on General Lie Groups

Riemannian Gaussian: $p(X | M, \sigma^2) = k(\sigma) \exp\left(-\frac{d(X, M)^2}{2\sigma^2}\right)$

Centering to the neutral element E : $\forall i \leq N, \tilde{P}_i \leftarrow L_{M_{\odot}^{-1}}(P_i)$,

Key operations: Scaling the dispersion: $\forall i \leq N, \tilde{P}_i \leftarrow \text{Exp}_E\left[\frac{s}{\sqrt{v^2 + \epsilon}} \text{Log}_E(\tilde{P}_i)\right]$,

Biasing towards parameter $B \in \mathcal{M}$: $\forall i \leq N, \tilde{P}_i \leftarrow L_B(\tilde{P}_i)$,

Proposition 4.1 (Population). Given a random point X over \mathcal{M} , and the Gaussian distribution $\mathcal{N}(M, v^2)$ defined in Eq. (12), we have the following properties for the population statistics:

1. (MLE of M) Given $\{P_{i\dots N} \in \mathcal{M}\}$ i.i.d. sampled from $\mathcal{N}(M, v^2)$, the maximum likelihood estimator (MLE) of M is the sample Fréchet mean.
2. (Homogeneity) Given $X \sim \mathcal{N}(M, v^2)$ and $B \in \mathcal{M}$, $L_B(X) \sim \mathcal{N}(L_B(M), v^2)$

Proposition 4.2 (Sample). Given N samples $\{P_{i\dots N} \in \mathcal{M}\}$, denoting $\phi_s(P_i) = \text{Exp}_E[s \text{Log}_E(P_i)]$, we have the following properties for the sample statistics:

Homogeneity of the sample mean: $\text{FM}\{L_B(P_i)\} = L_B(\text{FM}\{P_i\}), \forall B \in \mathcal{M}$, (16)

Controllable dispersion from E : $\sum_{i=1}^N w_i d^2(\phi_s(P_i), E) = s^2 \sum_{i=1}^N w_i d^2(P_i, E)$, (17)

where $\{w_{1\dots N}\}$ are weights satisfying a convexity constraint, i.e., $\forall i, w_i > 0$ and $\sum_i w_i = 1$.

Algorithm 1: Lie Group Batch Normalization (LieBN) Algorithm

```

Input : A batch of activations  $\{P_{1\dots N} \in \mathcal{M}\}$ , a small positive constant  $\epsilon$ , and
        momentum  $\gamma \in [0, 1]$ 
        running mean  $M_r = E$ , running variance  $v_r^2 = 1$ ,
        biasing parameter  $B \in \mathcal{M}$ , scaling parameter  $s \in \mathbb{R}/\{0\}$ ,
Output : Normalized activations  $\{\tilde{P}_{1\dots N}\}$ 
if training then
    Compute batch mean  $M_b$  and variance  $v_b^2$  of  $\{P_{1\dots N}\}$ ;
    Update running statistics  $M_r \leftarrow \text{WFM}(\{1 - \gamma, \gamma\}, \{M_r, M_b\})$ ,  $v_r^2 \leftarrow (1 - \gamma)v_r^2 + \gamma v_b^2$ ;
end
if training then  $M \leftarrow M_b$ ,  $v^2 \leftarrow v_b^2$ ;
else  $M \leftarrow M_r$ ,  $v^2 \leftarrow v_r^2$ ;
for  $i \leftarrow 1$  to  $N$  do
    Centering to the neutral element  $E$ :  $\tilde{P}_i \leftarrow L_{M_{\odot}^{-1}}(P_i)$ 
    Scaling the dispersion:  $\tilde{P}_i \leftarrow \text{Exp}_E\left[\frac{s}{\sqrt{v^2 + \epsilon}} \text{Log}_E(\tilde{P}_i)\right]$ 
    Biassing towards parameter  $B$ :  $\tilde{P}_i \leftarrow L_B(\tilde{P}_i)$ 
end

```

Natural generalization of the EBN

Proposition D.1. The LieBN algorithm presented in Alg. I is equivalent to the standard Euclidean BN when $\mathcal{M} = \mathbb{R}^n$, both during the training and testing phases.

Gaussian Preservation

$$\phi_s(P) = \text{Exp}_E[s \text{Log}_E(P)]$$

Lemma C.1. Given a random point X distributed over \mathcal{M} with P.D.F. p_X , the P.D.F. of $Y = \phi_s(X)$ is given by:

$$p_Y(Q) = \Delta p_X(\phi_s^{-1}(Q)). \quad (26)$$

where $\Delta = \frac{|\phi_s^{-1}|}{L_{\phi_s^{-1}(Q) \odot Q^{-1}*}}$. Here $|\cdot|$ denotes the determinant, and ϕ_s^{-1} and $L_{\phi_s^{-1}(Q) \odot Q^{-1}*}$ are the differentials.

Corollary C.2. Following the notations in Lem. C.1, if $\Delta = c$ is a constant and $X \sim \mathcal{N}(E, \sigma^2)$, then Y also follows a Gaussian distribution, i.e., $Y \sim \mathcal{N}(E, s^2 \sigma^2)$

Corollary C.4. Given a Lie group \mathcal{M} pulled back from the Euclidean space, and a random point $X \sim \mathcal{N}(E, \sigma^2)$ over \mathcal{M} , $Y = \phi_s(X) \sim \mathcal{N}(E, s^2 \sigma^2)$

LEM, LCM and their variants: $\mathcal{N}(M, \sigma^2) \rightarrow \mathcal{N}(E, \sigma^2) \rightarrow \mathcal{N}(E, s^2) \rightarrow \mathcal{N}(B, s^2)$

LieBN on the deformed SPD Lie groups

Power deformation: $\tilde{g} = \frac{1}{\theta^2} \text{Pow}_{\theta}^* g$,

Table 3: Key operators in calculating LieBN on SPD manifolds.

Metric	(θ, α, β)-AIM	(α, β)-LEM	θ -LCM
Pullback Map	P_{θ}	mlog	$P_{\theta} \odot \psi_{\text{LC}}$
Codomain	$\{\mathcal{S}_{++}^n, \odot^{\text{AI}}, \frac{1}{\theta^2} g^{(\alpha, \beta)-\text{AI}}\}$	$\{\mathcal{S}^n, \langle \cdot, \cdot \rangle^{(\alpha, \beta)}\}$	$\{\mathcal{L}^n, \frac{1}{\theta^2} \langle \cdot, \cdot \rangle\}$
$L_Q(P)$	KPK^{\top}	$P + Q$	$P + Q$
$L_{Q_{\odot}}(P)$	$K^{-1}PK^{-\top}$	$P - Q$	$P - Q$
$\text{Exp}_E[s \text{Log}_E(P)]$	P^s	sP	sP
FM	Karcher Flow	Arithmetic average	Arithmetic average
$\text{WFM}(\{1 - \gamma, \gamma\}, \{P_1, P_2\})$	$P_2^{\frac{1}{2}} (P_2^{-\frac{1}{2}} P_1 P_2^{-\frac{1}{2}})^{\gamma} P_2^{\frac{1}{2}}$	Arithmetic weighted average	Arithmetic weighted average

Proposition 5.1 (Deformation). (θ, α, β)-LEM is equal to (α, β) -LEM. θ -LCM interpolates between \tilde{g} -LEM ($\theta = 0$) and LCM ($\theta = 1$), with \tilde{g} -LEM defined as

$$\langle V, W \rangle_P = \tilde{g}(\text{mlog}_{*,P}(V), \text{mlog}_{*,P}(W)), \forall P \in \mathcal{S}_{++}^n, \forall V, W \in T_P \mathcal{S}_{++}^n, \quad (18)$$

where $\tilde{g}(V_1, V_2) = \frac{1}{2}(V_1, V_2) - \frac{1}{4}\langle \mathbb{D}(V_1), \mathbb{D}(V_2) \$