KIMYO INTERNATIONAL UNIVERSITY IN

CENTRAL ASIAN JOURNAL OF STEM

ISSN 2181-2934 http://stem.kiut.uz/

https://doi.org/10.5281/zenodo.21978532

Compositional Lipschitz Bounds and Robustness for Multi-Stream Skeleton-Based Action Recognition

Kabul Khudaybergenov

Kimyo International University in Tashkent, Tashkent, Uzbekistan

q.xudaybergenov@kiut.uz

Abstract. In skeleton-based human action recognition the highest reported accuracies come almost invariably from multi-stream ensembles that average the softmax outputs of parallel joint, bone and velocity branches. The operator defined by such an ensemble, however, has never been supplied with a provable stability guarantee. This paper closes that gap with a self-contained compositional Lipschitz analysis. We model the multi-stream recogniser as a convex aggregation of parallel branches, each of which composes a hand-designed stream generator, a deep graph-temporal encoder and a softmax classifier. Our central finding is that the complete operator's Lipschitz constant equals the weight-averaged branch constant. Convex aggregation can therefore never degrade stability below that of the least stable branch, while the factor predicted by a naive parallel-composition argument is cancelled exactly by the aggregation weights. We then derive explicit constants, readable from trained weights, for the layer types that occur in practice - multi-subset graph convolution, temporal convolution, residual connections, inference-time BatchNorm, pooling, the linear head and softmax, as well as for the joint, bone and velocity stream generators.

Keywords: skeleton-based action recognition; multi-stream ensemble; Lipschitz continuity; certified robustness; graph convolutional networks; compositional bounds.

1. Introduction

Skeleton-based human action recognition has become a core problem in video understanding. Representing a person by a sequence of joint coordinates discards appearance and background, which keeps the representation compact and largely privacy-preserving while leaving it relatively insensitive to changes in lighting and camera viewpoint. Since spatial-temporal graph convolutional networks were first proposed [1], the field has been dominated by graph convolutional network (GCN) architectures [2-4], and successive studies have refined the graph topology, the temporal modelling and the training signal [5-8]; the survey of Ren *et al*. [9] follows this trajectory in detail. Across nearly all of these systems the best accuracies are obtained not by a single network but by a multi-stream ensemble, in which separate branches are trained on the joint, bone and temporal-difference (velocity) representations and their softmax outputs are averaged [2,3]. Multi-stream design has thus become the standard route to state-of-the-art performance.

Once such models leave benchmark leaderboards and enter real deployments, surveillance, ambient assisted living, rehabilitation monitoring, human-robot interaction accuracy on clean test data ceases to be an adequate criterion. The skeletons are themselves the output of an upstream pose estimator and therefore carry estimation error, quantisation, missing joints and jitter. Beyond this ordinary noise, skeleton recognisers have been shown to break under small, deliberately constructed perturbations [10,11]. What is missing is *a quantitative and provable* statement of how far the output can move when the input is perturbed within a bounded budget, that is, a robustness guarantee.

Skeleton recognition and multi-stream models. The graph convolutional formulation of Yan *et al.* [1] established the joint-frame graph as the standard substrate for skeleton recognition. Subsequent architectures introduced adaptive, data-dependent

topologies [2], multi-scale and disentangled aggregation [3], deformable graph sampling [5], decoupled knowledge embedding [6], graph convolution augmented by self-attention [7,12] and semantic-aware spatio-temporal topology modelling [8]; efficient temporal pooling variants have appeared as well [13], and the design space as a whole is surveyed in [9]. One element recurs throughout this line of work [2-4]: the multi-stream ensemble, in which joint, bone and velocity branches are trained separately and their probabilities averaged. None of these studies asks what that ensemble structure implies for stability.

Lipschitz constants and certified robustness. Much of deterministic robustness certification rests on Lipschitz continuity. Tsuzuku *et al.* [14] linked the global Lipschitz constant to a classification margin, while randomized smoothing [15] provides a separate, probabilistic route to the same end. Estimating the constant tightly, or computing it exactly, has been approached through exact combinatorial computation for ReLU networks [16], semidefinite programming [17,18], operator-averaging certificates [19] and dynamical-system parameterisations [20]. Spectral normalisation [21] keeps the constant small during training, and the awkward case of self-attention was settled by Kim *et al.* [22]. Every one of these methods, however, certifies a *single feedforward network*.

Stability of graph neural networks. A neighbouring strand of research asks how stable graph neural networks are when the graph itself is perturbed. Gama *et al.* [23] established stability under graph deformations, Kenlay *et al.* [24] analysed edge rewiring, and more recent work bounds stability under manifold deformations [25] and controls generalization through Lipschitz loss functions [26]. In each case what is perturbed is the operator, namely the graph or the manifold underlying it.

Robustness of skeleton recognisers. The specific fragility of skeleton recognisers has been demonstrated empirically by black-box attacks [10] and by adversarial early-action prediction [11]. These studies construct perturbations that defeat the models, but they do not deliver a radius within which no perturbation can succeed.

Gap, aim and objectives. The literature reviewed above leaves one feature untouched that is at once specific to and universal in high-performance skeleton recognition: the operator of interest is a convex aggregation of several parallel branches, each branch being itself a composition of a hand-designed stream generator, a deep graph-temporal encoder and a softmax classifier, with the aggregation carried out on the probability simplex. Existing Lipschitz certificates address one feedforward network at a time, and existing stability results for graph networks perturb the graph rather than the signal. To our knowledge the Lipschitz behaviour of this compositional, multi-stream operator, and in particular the role played by the ensemble aggregation, has not been examined. Applied carelessly, parallel composition bounds even suggest that ensembling should inflate the Lipschitz constant by a factor **Ошибка! Объект не может быть создан из кодов полей редактирования.** in the number of streams, which would make the multi-stream design a liability rather than an asset. The aim of this work is accordingly to establish a complete, self-contained and weight-computable compositional Lipschitz analysis of the multi-stream skeleton recognition operator, and to convert it into a certified robustness radius together with an explicit account of when that radius is meaningful. Four objectives follow:

- to formalise the multi-stream recogniser as a composition of stream generation, graph-temporal encoding, classification and ensemble aggregation, and to determine the Lipschitz constant of the resulting operator;
- to derive explicit constants, computable from trained weights, for every layer type and every stream generator encountered in practice;
- to compare the available compositional bounds, establishing their ordering and the exact conditions under which each is attained;
- to translate the tightest bound into a certified radius and to characterise when a global compositional bound ceases to carry information.

The rest of the paper structured as follows. Section 2 sets out the analytical framework, the assumptions and the layer-wise machinery; Section 3 reports the

theorems, the bound hierarchy, the radius, the non-vacuity condition and a fully worked numerical example; Section 4 interprets these results, relates them to earlier work and states the limitations; Section 5 concludes.

2. Methods

This is an analytical study. The object under investigation is the input-output operator of a trained multi-stream skeleton recogniser, and the instrument is compositional Lipschitz analysis in normed spaces. Section 2.1 specifies the object together with the assumptions it must satisfy; Section 2.2 describes the analytical procedure and proves the elementary composition rules on which it rests; Section 2.3 makes that procedure executable on real weights by supplying explicit per-layer and per-stream constants, and closes with the numerical routine used when those constants turn out to be too coarse. Each step is stated so that the resulting numbers can be reproduced from a trained checkpoint without any retraining.

2.1 Problem setting

Throughout, $(\mathrm{X}, \|\cdot\|_{\mathrm{X}})$ and $(\mathrm{Y}, \|\cdot\|_{\mathrm{Y}})$ denote real normed spaces. A map $T : \mathrm{X} \to \mathrm{Y}$ is *Lipschitz* with constant $L \geq 0$ if

$$\|T(x) - T(y)\|_{\mathrm{Y}} \leq L \|x - y\|_{\mathrm{X}}, \ \forall x, y \in \mathrm{X}, \tag{1}$$

and $\mathrm{Lip}(T)$ denotes the smallest such constant. For a matrix A we write $\|A\|_2$ for the spectral norm and $\|A\|_F$ for the Frobenius norm, and we shall repeatedly invoke the standard inequality

$$\|AXB\|_F \leq \|A\|_2 \|X\|_F \|B\|_2 . \tag{2}$$

A skeleton sequence is a tensor $X \in \mathrm{X} := \mathbb{R}^{T \times V \times C}$, where T is the number of frames, V the number of joints and C the coordinate dimension ($C = 2$ or 3). We endow X with the Frobenius norm

$$\|X\|_F = \Big(\sum_{t,v,c} X_{tvc}^2 \Big)^{1/2} .$$

Remark 1. (*Why the input space is not the video space*). It is tempting to treat raw video as the input space and to postulate a Lipschitz *skeleton extraction* operator $\mathrm{S}:\mathrm{V}\to\mathrm{X}$, but this is not admissible. Modern pose estimators rely on heatmap argmax decoding, person detection and identity association across frames, and each of these operations is discontinuous, so an arbitrarily small pixel perturbation can flip a detection or reassign a track. We therefore take the (normalised) skeleton sequence as the input and treat pose estimation error as an input perturbation budget ε, which can be calibrated empirically from the MPJPE of the estimator in use. Every statement below then applies to the deployed pipeline, with ε absorbing the extraction stage; and if a Lipschitz extraction operator does happen to be available on a restricted domain, each bound below is simply rescaled by L_{S}.

Let $M\in\mathbb{N}$ be the number of streams and K the number of action classes, and let $\Delta^K=\{p\in\mathbb{R}^K_{\geq 0}:\sum_k p_k=1\}$ denote the probability simplex. The model consists of

$$\varphi_m:\mathrm{X}\to\mathrm{Z}_m \qquad \text{(stream / latent-feature generation)} \tag{3}$$

$$\Psi_m:\mathrm{Z}_m\to\mathrm{H}_m \qquad \text{(graph-temporal encoder),} \tag{4}$$

$$\mathrm{C}_m:\mathrm{H}_m\to\Delta^K \qquad \text{(pooling, classifier head, softmax),} \tag{5}$$

$$\mathrm{A}_w:(\Delta^K)^M\to\Delta^K \qquad \text{(ensemble aggregation),} \tag{6}$$

for $m=1,\dots,M$, and the recognition operator is

$$F=\mathrm{A}_w\circ\left(\mathrm{C}_1\circ\Psi_1\circ\varphi_1,\dots,\mathrm{C}_M\circ\Psi_M\circ\varphi_M\right). \tag{7}$$

We write $G_m:=\mathrm{C}_m\circ\Psi_m\circ\varphi_m$ for the m-th branch and set $G:=(G_1,\dots,G_M)$. On the product space $\mathrm{P}:=(\Delta^K)^M$ two norms will be needed,

$$\|p\|_{2,2}:=\Big(\sum_{m=1}^{M}\big\|p^{(m)}\big\|_2^2\Big)^{1/2},\quad \|p\|_{\infty,2}:=\max_{1\leq m\leq M}\big\|p^{(m)}\big\|_2, \tag{8}$$

whereas Δ^K itself always carries $\|\cdot\|_2$.

Assumption 1. For every $m\in\{1,\dots,M\}$ there exist finite constants $L_{\varphi_m},L_{\Psi_m},L_{\mathrm{C}_m}$ such that:

(A1) φ_m is Lipschitz with constant L_{φ_m} on X;

(A2) Ψ_m is Lipschitz with constant L_{Ψ_m} on $\varphi_m(\mathrm{X})$;

(A3) C_m is Lipschitz with constant L_{C_m} on $\Psi_m(\varphi_m(\mathrm{X}))$, the target being equipped with $\|\cdot\|_2$;

(A4) $\mathrm{A}_w(p^{(1)},\ldots,p^{(M)}) = \sum_{m=1}^{M} w_m p^{(m)}$ with $w \in \Delta^M$, i.e. $w_m \geq 0$ and $\sum_m w_m = 1$.

Remark 2 (Domain restriction is essential). Assumption (A2) is stated on the *image* of φ_m rather than on all of Z_m for a concrete reason: standard dot-product self-attention is not globally Lipschitz on an unbounded domain [22]. Whenever Ψ_m contains attention blocks, (A2) must therefore be read as a local statement on a bounded set $\mathrm{B} \supseteq \varphi_m(\mathrm{X})$. In practice this costs nothing, because normalised skeleton coordinates already lie in a compact set; the alternative is to adopt L_2-attention, for which a global constant is known [22]. For purely convolutional graph encoders, Section 2.3 makes L_{Ψ_m} explicit and no restriction of this kind is needed.

2.2 Analytical strategy and composition rules

The analysis proceeds in four steps. First, the operator (7) is decomposed along its two structural axes: sequentially within each branch ($\varphi_m \to \Psi_m \to \mathrm{C}_m$), and in parallel across branches, followed by the aggregation A_w. Second, constants are propagated along these two axes using the elementary rules of Lemmas 1-3 below; since the parallel step admits more than one product norm, three distinct compositional bounds arise, and rather than fixing one in advance we retain and compare all three. Third, the abstract constants of Assumption 1 are replaced by quantities that can be read directly off a checkpoint – spectral norms of convolution kernels and adjacency matrices, BatchNorm scale parameters, pooling sizes and the softmax temperature as collected in Propositions 1 and 2; obtaining them requires a single pass over the stored weights and no data at all. Fourth, the resulting constant is converted into a radius within which the predicted class

cannot change, and the diameter of the output simplex decides when that radius is non-trivial; where the global constant proves too coarse, the local constant (10) is estimated numerically by Algorithm 1. All bounds are stated with the Frobenius norm on X and on every intermediate space, and with the Euclidean norm on Δ^K. Because compositional constants depend on the norms chosen, this choice is fixed once and applied consistently; the worked example of Section 3.3 follows precisely the same four steps and reports every intermediate quantity, so that the computation can be reproduced.

The three facts on which the propagation rests are classical, and the short proofs below serve only to pin down the constants and the norms that will be used later.

Lemma 1 (Sequential composition). *Let* $T_1 : \mathrm{X} \to \mathrm{Y}$ *and* $T_2 : \mathrm{Y} \to \mathrm{Z}$ *be Lipschitz with constants* L_1 *and* L_2. *Then* $T_2 \circ T_1$ *is Lipschitz, with* $\mathrm{Lip}(T_2 \circ T_1) \le L_2 L_1$.

Proof. For any $x, y \in \mathrm{X}$,

$$\|T_2(T_1(x)) - T_2(T_1(y))\| \le L_2 \|T_1(x) - T_1(y)\| \le L_2 L_1 \|x - y\|. \square$$

Lemma 2 (Parallel aggregation, Euclidean product norm). *Let* $G_m : \mathrm{X} \to \mathrm{Y}_m$ *be Lipschitz with constants* L_m *for* $m = 1, \ldots, M$, *and let* $G = (G_1, \ldots, G_M) : \mathrm{X} \to \prod_m \mathrm{Y}_m$, *the product carrying the norm* $\|\cdot\|_{2,2}$ *of (8). Then* $\mathrm{Lip}(G) \le \left(\sum_m L_m^2 \right)^{1/2}$.

Proof. Squaring and summing termwise,

$$\|G(x) - G(y)\|_{2,2}^2 = \sum_m \|G_m(x) - G_m(y)\|^2 \le \sum_m L_m^2 \|x - y\|^2,$$

and taking square roots gives the claim. $\square$

Lemma 3 (Parallel aggregation, max product norm). *Under the hypotheses of Lemma 2, but with the product carrying* $\|\cdot\|_{\infty,2}$, *one has* $\mathrm{Lip}(G) \le \max_m L_m$.

Proof.

$$\|G(x) - G(y)\|_{\infty,2} = \max_m \|G_m(x) - G_m(y)\| \le \max_m L_m \|x - y\|. \ \square$$

2.3 Constants computable from trained weights

Assumption 1 has little practical value unless its constants can be recovered from the trained weights, and the two propositions of this section supply them. Throughout, layers act on tensors reshaped as required, and Frobenius norms are used on every intermediate space.

Proposition 1 (Layer-wise Lipschitz constants). *Each map below is Lipschitz, with the constant stated.*

1. (Activations) *Any elementwise map* σ *with* $|\sigma'| \le 1$ *a.e. – in particular ReLU, LeakyReLU with slope in* $[-1,1]$, *and* tanh *satisfies* $\mathrm{Lip}(\sigma) \le 1$. *For* GELU, $\mathrm{Lip}(\sigma) \le 1.0899$*; for the logistic sigmoid,* $\mathrm{Lip}(\sigma) \le 1/4$.

2. (Multi-subset graph convolution) $H \mapsto \sigma\big(\sum_{s=1}^{S} \hat{A}_s H W_s\big)$, with $\hat{A}_s \in \mathbb{R}^{V \times V}$, $W_s \in \mathbb{R}^{d \times d'}$, has Lipschitz constant at most $\sum_{s=1}^{S} \|\hat{A}_s\| \|W_s\|$. The adjacency norm $\|\hat{A}_s\|$ is fixed by the normalisation scheme: under the symmetric renormalization $\hat{A}_s = \tilde{D}_s^{-1/2}(A_s + I)\tilde{D}_s^{-1/2}$ one has $\|\hat{A}_s\| \le 1$, while for a learnable adaptive topology $\hat{A}_s = \hat{A}_s^0 + \Delta A_s$ (a normalised base term plus a data-dependent residual) the triangle inequality yields $\|\hat{A}_s\| \le \|\hat{A}_s^0\| + \|\Delta A_s\| \le 1 + \|\Delta A_s\|$.

3. (Temporal convolution) The linear map $(\mathrm{T}X)_t = \sum_{k=1}^{K_t} W_k^{\bullet}\, x_{t+k-1}$ with zero padding satisfies $\mathrm{Lip}(\mathrm{T}) \le \sum_{k=1}^{K_t} \|W_k\|$.

4. (Residual block) If B is Lipschitz with constant L_{B} then $x \mapsto x + \mathrm{B}(x)$ has constant at most $1 + L_{\mathrm{B}}$.

5. (BatchNorm at inference) The affine map $x_c \mapsto \gamma_c (x_c - \mu_c)/\sqrt{\varsigma_c^2 + \dot{\mathrm{o}}} + \beta_c$ has constant $\max_c |\gamma_c| / \sqrt{\varsigma_c^2 + \dot{\mathrm{o}}}$.

6. (Global average pooling) Averaging over N positions, $x \mapsto N^{-1}\sum_{i=1}^{N} x_i$, has constant $N^{-1/2}$.

7. (Linear head) $h \mapsto Wh + b$ has constant $\|W\|$.

8. (Softmax) $\mathrm{Lip}(\mathrm{softmax}) = 1/2$ with respect to $\|\cdot\|_2$; with inverse temperature $\tau > 0$, $\mathrm{Lip}(\mathrm{softmax}_\tau) \le 1/(2\tau)$.

Proof. (1) follows by applying the mean value theorem coordinatewise and summing,

$$\|\sigma(X) - \sigma(Y)\|_F^2 = \sum_i |\sigma(x_i) - \sigma(y_i)|^2 \le \sum_i |x_i - y_i|^2;$$

the numerical value quoted for GELU is $\max_t |\mathrm{GELU}'(t)|$.

(2) By (1) it suffices to bound the affine part, and the triangle inequality combined with (2) gives

$$\left\| \sum_s \hat{A}_s (H_1 - H_2) W_s \right\|_F \le \sum_s \|\hat{A}_s\|_2 \, \|H_1 - H_2\|_F \, \|W_s\|_2 .$$

Under the symmetric renormalisation $\hat{A}_s$ is symmetric and similar to the nonnegative random-walk matrix $\tilde{D}_s^{-1}(A_s + I)$, so its eigenvalues lie in $(-1,1]$ and $\|\hat{A}_s\|_2 \le 1$; the adaptive case is once more the triangle inequality.

(3) Write $\mathrm{T} = \sum_k \mathrm{L}_{W_k} \circ \mathrm{S}_k$, in which S_k shifts by $k-1$ frames with zero padding and L_{W_k} is the pointwise linear map $x \mapsto W_k^{\bullet} x$. Zero-padded shifts are contractions, whence $\mathrm{Lip}(\mathrm{S}_k) \le 1$, while $\mathrm{Lip}(\mathrm{L}_{W_k}) = \|W_k\|$; the triangle inequality for operator norms completes the argument.

(4) Directly,

$$|(x + \mathrm{B}(x)) - (y + \mathrm{B}(y))| \le \|x - y\| + L_{\mathrm{B}} \|x - y\|.$$

(5) The map is diagonal and affine, so its operator norm is the largest absolute diagonal entry.

(6) The triangle and Cauchy--Schwarz inequalities give

$$\left\| \tfrac{1}{N}\sum_{i=1}^{N}(x_i - y_i) \right\|_2 \le \tfrac{1}{N}\sum_{i=1}^{N}\|x_i - y_i\|_2 \le \tfrac{1}{N}\sqrt{N}\Big(\sum_{i=1}^{N}\|x_i - y_i\|_2^2\Big)^{1/2} = N^{-1/2}\,\|x - y\|_F .$$

(7) is the definition of the spectral norm.

(8) At $p = \mathrm{softmax}(z)$ the Jacobian of the softmax is $J(z) = \mathrm{diag}(p) - pp^{\top}$, which is symmetric and positive semidefinite. For an arbitrary unit vector u,

$$u^{\top} Ju = \sum_k p_k u_k^2 - \Big(\sum_k p_k u_k\Big)^2 = \mathrm{Var}_p(u) \le \tfrac{1}{4}(\max_k u_k - \min_k u_k)^2 \le \tfrac{1}{2},$$

by Popoviciu's inequality on variances together with $(\max_k u_k - \min_k u_k)^2 \le 2\|u\|_2^2 = 2$. Hence $\|J(z)\| \le 1/2$ for every z, and the mean value inequality along the segment $[z_1, z_2]$ yields

$$\|\mathrm{softmax}(z_1) - \mathrm{softmax}(z_2)\|_2 \le \tfrac{1}{2}\|z_1 - z_2\|_2 .$$

The bound is attained in the limit $p \to (\tfrac{1}{2}, \tfrac{1}{2}, 0, \ldots, 0)$, where the eigenvalues of J are 0 and $\tfrac{1}{2}$, so $1/2$ cannot be improved. The temperature version follows from $\mathrm{softmax}_\tau(z) = \mathrm{softmax}(z/\tau)$ and Lemma 1. □

Proposition 2 (Stream generation). *Let $d_{\max}$ be the maximum number of children of a joint in the kinematic tree.*

1. The identity (joint) stream has constant 1.

2. The bone stream $b_v = x_v - x_{\pi(v)}$, where π is the parent map, has constant at most $1 + \sqrt{d_{\max}}$.

3. The velocity stream $v_t = x_{t+1} - x_t$ has constant at most 2.

4. A composition of streams (e.g.\ bone velocity) obeys Lemma 1.

Proof. (ii) The map is $X \mapsto (I - P)X$, in which P is the $0/1$ matrix with $P_{v\pi(v)} = 1$ and zeros elsewhere. Each row of P has exactly one nonzero entry and each column j has

$\left|\pi^{-1}(j)\right| \le d_{\max}$ of them, so $\|P\|_2^2 = \left\|P^{\bullet} P\right\|_2 \le d_{\max}$ (by Gershgorin, or directly from $\|P\|_2 \le \sqrt{\|P\|_1 \|P\|_\infty} = \sqrt{d_{\max} \cdot 1}$). Consequently $\|I - P\| \le 1 + \sqrt{d_{\max}}$, and (2) applies. (iii) Here the map is $\mathrm{S}_1 - \mathrm{I}$ with S_1 a zero-padded shift, so its norm is at most 2. Parts (i) and (iv) are immediate. □

Remark 3 (Scale normalization) Skeleton pipelines often divide by a scale statistic $s(X) > 0$, such as a reference bone length. The map $X \mapsto X / s(X)$ is *not* globally Lipschitz, but it is Lipschitz on the set $\{\|X\|_F \le R, s(X) \ge s_{\min} > 0\}$ provided s is itself Lipschitz with constant L_s, and then

$$\mathrm{Lip}\big(X \mapsto X / s(X)\big) \le \frac{1}{s_{\min}} + \frac{R L_s}{s_{\min}^2}. \tag{9}$$

To see this, write $Xs(Y) - Ys(X) = s(Y)(X - Y) + Y\big(s(Y) - s(X)\big)$, so that

$$\left\| \frac{X}{s(X)} - \frac{Y}{s(Y)} \right\| = \frac{\|Xs(Y) - Ys(X)\|_F}{s(X)s(Y)} \le \|X - Y\|_F \left(\frac{1}{s(X)} + \frac{\|Y\|_F L_s}{s(X)s(Y)} \right),$$

from which the stated bound follows.

Finally, when the global product of spectral norms proves too coarse to be useful, the precise condition is given in Proposition 4, the global constant $L^{å}$ may be replaced by the local constant

$$L^{å}_{\mathrm{loc}}(X, \rho) := \sum_{m=1}^{M} w_m \sup_{X' \in B(X, \rho)} \|\mathrm{D}G_m(X')\|_2, \tag{10}$$

where $B(X, \rho)$ is the Frobenius ball of radius ρ centred at X. The inner spectral norm is estimated by power iteration on the branch Jacobian, as set out in Algorithm 1. Each iteration costs one forward-mode Jacobian--vector product and one reverse-mode vector--Jacobian product, so the total expense is $O(N_{\mathrm{it}})$ network passes per branch, again with no access to the training data.

Algorithm 1. Power iteration estimate of the branch Jacobian norm

Require: branch G_m, point X, iterations N_{it}

1: $u \leftarrow$ random tensor of the shape of X; $u \leftarrow u / \|u\|_2$

2: **For** $i = 1$ to N_{it} **do**

3: $v \leftarrow \mathrm{D}G_m(X)u$ // one forward-mode JVP

4: $u \leftarrow \mathrm{D}G_m(X)^{\cdot} v$ //one reverse-mode VJP

5: $\lambda \leftarrow \|u\|_F$; \; $u \leftarrow u / \lambda$

6: **EndFor**

7: **Return** $\sqrt{\lambda}$ //estimate of $\|\mathrm{D}G_m(X)\|_2$

3. Results

3.1 The Lipschitz constant and the hierarchy of bounds

Theorem 1 (Stability of the multi-stream operator). *Let F be as in (7) and suppose Assumption (1) holds. Then F is well defined, takes values in Δ^K, and is Lipschitz with*

$$\mathrm{Lip}(F) \le L^{\text{å}} := \sum_{m=1}^{M} w_m \underbrace{L_{\varphi_m} L_{\Psi_m} L_{\mathrm{C}_m}}_{=:L_m}. \tag{11}$$

In particular $\mathrm{Lip}(F) \le \max_m L_m$, *so aggregating streams never worsens the stability guarantee beyond that of the least stable stream. If, in addition, a Lipschitz extraction operator* S *with constant* L_{S} *is prepended (cf. Remark 1), then* $\mathrm{Lip}(F \circ \mathrm{S}) \le L_{\mathrm{S}} L^{\text{å}}$.

Proof. Step 1 (branches). Fix m and apply Lemma (1) twice under (A1)-(A3):

$$\mathrm{Lip}(G_m) = \mathrm{Lip}\left(\mathrm{C}_m \circ \Psi_m \circ \varphi_m\right) \le L_{\varphi_m} L_{\Psi_m} L_{\mathrm{C}_m} = L_m. \tag{12}$$

The constant L_m depends on the m-th branch alone, and the composition is well defined because (A2)-(A3) place each map on the image of its predecessor.

Step 2 (aggregation). Let $X, Y \in \mathrm{X}$. Combining (A4), the linearity of A_w, the triangle inequality, homogeneity ($w_m \ge 0$) and (12),

$$\|F(X) - F(Y)\|_2 = \left\| \sum_{m=1}^{M} w_m \left(G_m(X) - G_m(Y) \right) \right\|_2 \le \sum_{m=1}^{M} w_m \|G_m(X) - G_m(Y)\|_2$$

$$\le \Big(\sum_{m=1}^{M} w_m L_m\Big)\|X-Y\|_F = L^{\text{å}}\,\|X-Y\|_F. \tag{13}$$

Since X and Y were arbitrary, (11) follows.

Step 3 (range). Every $G_m(X)$ lies in Δ^K, which is convex; as $w \in \Delta^M$, the point $\sum_m w_m G_m(X)$ is a convex combination of points of Δ^K and therefore lies in Δ^K as well, so F maps into Δ^K.

Step 4 (the max *bound).* Finally, $\sum_m w_m L_m \le \Big(\sum_m w_m\Big)\max_m L_m = \max_m L_m$, and the last assertion is immediate from Lemma (1). □

Proposition 3 (Hierarchy of compositional bounds). *Under Assumption 1, the three natural bounds that follow from Lemmas 1-3 satisfy*

$$\underbrace{\sum_{m=1}^{M} w_m L_m}_{L^{\text{å}}\ (Thm.1)} \le \min\Big\{\underbrace{\|w\|_2\Big(\sum_{m=1}^{M} L_m^2\Big)^{1/2}}_{via\ Lemma\ 2},\ \underbrace{\max_{1\le m\le M} L_m}_{via\ Lemma\ 3}\Big\}. \tag{14}$$

The first term is attained exactly when w *is proportional to* $(L_1,\ldots,L_M)$*, and the second exactly when* w *is supported on* $\operatorname{argmax}_m L_m$*. In the homogeneous case* $L_m \equiv \ell$ *with* $w_m \equiv 1/M$*, all three coincide and equal* ℓ*.*

Proof. The first inequality is Cauchy-Schwarz, $\langle w, L\rangle \le \|w\|_2\,\|L\|_2$, with its usual equality condition, and the second was obtained in Step 4 of the proof of Theorem 1. As for the middle term, under $\|\cdot\|_{2,2}$ on P one has $\mathrm{Lip}(\mathrm{A}_w) \le \|w\|_2$, since Cauchy-Schwarz gives

$$\Big\|\sum_m w_m (p^{(m)} - q^{(m)})\Big\|_2 \le \sum_m w_m \|p^{(m)} - q^{(m)}\|_2 \le \|w\|_2\,\|p-q\|_{2,2}$$

combining this with Lemma 2 produces the term. Similarly, under $\|\cdot\|_{\infty,2}$ one has

$$\mathrm{Lip}(\mathrm{A}_w) \le \sum_m w_m = 1,$$

so Lemma 3 yields $\max_m L_m$.□

3.2 Certified radius and the limits of the global bound

For $p \in \Delta^K$ let $p_{(1)} \geq p_{(2)} \geq \dots$ denote the components in decreasing order, and define the *margin* $\gamma(p) := p_{(1)} - p_{(2)}$.

Corollary 1 (Radius). *Suppose F satisfies Theorem 1 with constant $L^{å} > 0$, let $X \in \mathrm{X}$, and let $p = F(X)$ have a unique maximiser $k^* = \operatorname{argmax}_k p_k$. If*

$$\|X' - X\|_F < r(X) := \frac{\gamma(p)}{\sqrt{2}\,L^{å}}, \tag{15}$$

then $\operatorname{argmax}_k F(X')_k = k^*$*; that is, the predicted action class provably does not change.*

Proof. Put $\Delta := F(X') - F(X)$, so that $\|\Delta\|_2 \leq L^{å}\varepsilon$ with $\varepsilon := \|X' - X\|_F$, and fix $j \neq k^*$. Then

$$F(X')_{k^*} - F(X')_j = \big(p_{k^*} - p_j\big) + \big(\Delta_{k^*} - \Delta_j\big) \geq \gamma(p) - \big(|\Delta_{k^*}| + |\Delta_j|\big).$$

For any two coordinates, $|\Delta_{k^*}| + |\Delta_j| \leq \sqrt{2}\big(\Delta_{k^*}^2 + \Delta_j^2\big)^{1/2} \leq \sqrt{2}\Delta_2 \leq \sqrt{2}L^{å}\varepsilon$, whence $F(X')_{k^*} - F(X')_j \geq \gamma(p) - \sqrt{2}L^{å}\varepsilon > 0$ whenever $\varepsilon < r(X)$.

Remark 4 (The factor $\sqrt{2}$). Had $|\Delta_{k^*}| + |\Delta_j|$ instead been bounded by $2\|\Delta\|_\infty \leq 2\|\Delta\|_2$, the weaker radius $\gamma / (2L^{å})$ would have resulted. The gain of a factor $\sqrt{2}$ comes from the observation that only two coordinates of Δ enter the estimate. □

Proposition 4 (Non-vacuity). $\operatorname{diam}(\Delta^K, \|\cdot\|_2) = \sqrt{2}$. *As a result, the inequality $\|F(X) - F(X')\|_2 \leq L^{å}\,\|X - X'\|_F$ is informative only on the range*

$$\|X - X'\|_F < \sqrt{2} / L^{å}, \tag{16}$$

and holds trivially outside it.

Proof. For $p, q \in \Delta^K$ one has $\|p - q\|_2^2 = \|p\|_2^2 + \|q\|_2^2 - 2\langle p, q\rangle \leq 1 + 1 - 0 = 2$, because $\|p\|_2 \leq \|p\|_1 = 1$ and $\langle p, q\rangle \geq 0$; the value $\sqrt{2}$ is attained at two distinct vertices. The second claim then follows at once. □

3.3 A worked four-stream example

Example 1. Consider a four-stream configuration ($M = 4$: joint, bone, joint velocity and bone velocity) with uniform ensemble weights $w_m = 1/4$, a kinematic tree with $d_{\max} = 3$, and sequences of length $T = 64$ over $V = 25$ joints. The encoder is spectrally normalised and consists of 10 residual graph-temporal blocks of per-block constant 1.3 ; the classifier head has $\|W_{\mathrm{fc}}\| = 4$, and the softmax uses $\tau = 1$.

Proposition 2 gives $L_{\varphi_1} = 1$, $L_{\varphi_2} \le 1 + \sqrt{3} = 2.732$, $L_{\varphi_3} \le 2$ and $L_{\varphi_4} \le 2 \cdot 2.732 = 5.464$. Proposition 1(4) yields $L_{\Psi_m} \le 1.3^{10} = 13.79$ for every m, while Proposition 1(6),(7),(8) with $N = TV = 1600$ give

$$L_{\mathrm{C}_m} \le N^{-1/2} \cdot \|W_{\mathrm{fc}}\| \cdot \tfrac{1}{2} = \tfrac{1}{40} \cdot 4 \cdot \tfrac{1}{2} = 0.05.$$

The resulting branch constants $L_m = L_{\varphi_m} L_{\Psi_m} L_{\mathrm{C}_m}$ are collected in Table 1.

Table 1. Branch constants for the configuration of Example 1.

Stream m	L_{φ_m}	L_{Ψ_m}	L_{C_m}	L_m
joint	1.000	13.79	0.05	0.690
bone	2.732	13.79	0.05	1.884
joint velocity	2.000	13.79	0.05	1.379
bone velocity	5.464	13.79	0.05	3.768

Theorem 1 then gives

$$L^{\text{å}} = \tfrac{1}{4}(0.690 + 1.884 + 1.379 + 3.768) = 1.930,$$

whereas the Euclidean-product route of Proposition 3 returns $\|w\|_2 \left(\sum_m L_m^2\right)^{1/2} = \tfrac{1}{2}\sqrt{20.12} = 2.243$ and the max route returns 3.768. The bound (11) is accordingly 14% tighter than the first and 49% tighter than the second; had $\mathrm{Lip}(\mathrm{A}_w)$ been left unquantified at 1, the reported constant would have been 4.486, roughly 2.3 times larger.

By Proposition 4 the bound remains informative for perturbations up to $\sqrt{2}/1.930 = 0.733$ in normalised coordinate units. For a sample with margin $\gamma = 0.6$, Corollary 1 certifies the prediction against every perturbation whose Frobenius norm falls below

$$r = \frac{0.6}{\sqrt{2}\cdot 1.930} = 0.220.$$

Since the perturbation is measured over the whole $64\times 25\times 3$ tensor, this corresponds to a per-joint, per-frame root-mean-square displacement of $0.220/\sqrt{4800} = 3.2\times 10^{-3}$.

4 Discussion

4.1 Interpretation and relation to earlier work

Theorem 1 identifies the Lipschitz constant of a multi-stream recogniser as the *weighted mean* of the branch constants, and Proposition 3 places that quantity below both competing bounds. Two consequences follow, each of which dispels a misconception that survives only while the aggregation constant is left unspecified.

*Remark 5 (*Two consequences for model design).

1. Invoking Lemma 2 on its own, with $\mathrm{Lip}(\mathrm{A}_w)$ treated as $O(1)$, creates the impression that ensembling inflates the Lipschitz constant by a factor $\sqrt{M}$. It does not: for a convex aggregation the $\sqrt{M}$ produced by the parallel branching is cancelled exactly by $\|w\|_2 = M^{-1/2}$ in the uniform case. The multi-stream design that dominates the benchmarks [2-4] is therefore no robustness liability, and the numerical example puts the error committed by the naive argument at a factor of 2.3.

2. Because $L^{\text{å}}$ is a weighted mean, the ensemble weights also act as a stability control: down-weighting a stream with a large L_m - typically one of the second-order streams, cf. Proposition 2 – reduces the radius denominator linearly, at a measurable cost in clean accuracy. In the configuration of Example 1 the bone-velocity branch alone accounts for almost half of $L^{\text{å}}$, so shifting weight away from it is by some margin the

cheapest lever available. This is a practical alternative to Lipschitz-constrained retraining or spectral normalization [21], neither of which can be applied after the fact.

The analysis is orthogonal to the architectural line of research reviewed in Section 1: it takes the ensemble structure as given and asks what it implies for stability, so the conclusions apply to any member of the family [1-8] whose layers admit the constants of Section 2.3. Set against the certification literature [14-20], which certifies one feedforward network at a time, the contribution here is the composition of per-layer and per-stream constants along the particular parallel-then-aggregate structure of the multi-stream recogniser, together with an explicit constant for the aggregation step itself. The margin mechanism behind Corollary 1 follows [14], but the factor $\sqrt{2}$ of Remark 4 improves on the customary 2, because only two coordinates of the output perturbation enter the estimate.

4.2 Limitations and future work

Four limitations deserve to be stated plainly. The first, and the strongest objection that can be raised against purely global compositional bounds, is the one recorded in Proposition 4.

Remark 6 (Why local constants are indispensable). In a deep encoder the product appearing in Proposition 1(4) grows geometrically with depth, so an $L^{\text{å}}$ assembled from spectral norms alone can easily exceed $\sqrt{2}/\varepsilon$ for realistic ε; at that point (11) remains formally correct but is practically empty. Three remedies are available: (i) constraining $\left\|\hat{A}_s\right\|_2 \|W_s\|_2$ during training by spectral normalisation [21]; (ii) substituting the local constant (10), estimated by Algorithm 1, for $L^{\text{å}}$; and (iii) re-weighting w as described in Remark 5.

Second, the bounds are *upper* bounds obtained through a chain of triangle and Cauchy-Schwarz inequalities. Each step is tight when taken in isolation, yet their composition need not be, so the certified radius is conservative. Third, Assumption (A2) is only a local statement whenever the encoder contains dot-product self-attention

[7,12], since such attention is not globally Lipschitz [22], as noted in Remark 2; by the same token, scale normalisation is Lipschitz only on a restricted domain, with the constant (9) of Remark 3, which is the formal counterpart of the empirical observation that scale normalisation amplifies noise for small subjects or distant cameras. Fourth, the study is purely analytical: the numerical example of Section 3.3 illustrates the framework but does not validate it on recorded data.

5. Conclusion

This paper has developed a self-contained compositional Lipschitz analysis of the multi-stream skeleton recognition operator and shown that, because the ensemble aggregation is convex, its Lipschitz constant is the weight-averaged branch constant $L^{å} = \sum_m w_m L_m$ (Theorem 1) - no worse than that of the least stable stream. The apparent $\sqrt{M}$ inflation is thus an artefact that disappears as soon as the aggregation weights are taken into account. Together with the explicit per-layer and per-stream constants of Propositions 1 and 2, this yields a weight-computable certified radius (Corollary 1) and turns the ensemble weights into an inexpensive stability control. The non-vacuity condition (Proposition 4) marks an honest boundary around the regime in which the global bound remains informative, thereby motivating the local constant of Remark 6 and leaving empirical validation on NTU RGB+D and NW-UCLA to future work.

References

1. Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 7444–7452. https://doi.org/10.1609/aaai.v32i1.12328

2. Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12026–12035. https://doi.org/10.1109/CVPR.2019.01230

3. Liu, Z., Zhang, H., Chen, Z., Wang, Z., & Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 143–152. https://doi.org/10.1109/CVPR42600.2020.00022

4. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., & Hu, W. (2021). Channel-wise topology refinement graph convolution for skeleton-based action recognition. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 13359–13368. https://doi.org/10.1109/ICCV48922.2021.01311

5. Myung, W., Su, N., Xue, J.-H., & Wang, G. (2024). DeGCN: Deformable graph convolutional networks for skeleton-based action recognition. IEEE Transactions on Image Processing, 33, 2477–2490. https://doi.org/10.1109/TIP.2024.3378886

6. Liu, Y., Li, Y., Zhang, H., Zhang, X., & Xu, D. (2024). Decoupled knowledge embedded graph convolutional network for skeleton-based human action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 34(10), 9445–9457. https://doi.org/10.1109/TCSVT.2024.3399126

7. Wu, Z., Ding, Y., Wan, L., Li, T., & Nian, F. (2025). Local and global self-attention enhanced graph convolutional network for skeleton-based action recognition. Pattern Recognition, 159, 111106. https://doi.org/10.1016/j.patcog.2024.111106

8. Cui, H., Huang, R., Zhang, R., & Hayama, T. (2025). DSTSA-GCN: Advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. Neurocomputing, 637, 130066. https://doi.org/10.1016/j.neucom.2025.130066

9. Ren, B., Liu, M., Ding, R., & Liu, H. (2024). A survey on 3D skeleton-based action recognition using learning method. Cyborg and Bionic Systems, 5, 0100. https://doi.org/10.34133/cbsystems.0100

10. Diao, Y., Shao, T., Yang, Y.-L., Zhou, K., & Wang, H. (2021). BASAR: Black-box attack on skeletal action recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7597–7607. https://doi.org/10.1109/CVPR46437.2021.00751

11. Li, X.-S., Zhang, N., Cai, B.-Q., et al. (2024). Adversarial graph convolutional network for skeleton-based early action prediction. Journal of Computer Science and Technology, 39(6), 1269–1280. https://doi.org/10.1007/s11390-023-2638-7

12. Tsuzuku, Y., Sato, I., & Sugiyama, M. (2018). Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. Advances in Neural Information Processing Systems, 31.

13. Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. Proceedings of the 36th International Conference on Machine Learning (PMLR, 97), 1310–1320. https://doi.org/10.48550/arXiv.1902.02918

14. Jordan, M., & Dimakis, A. G. (2020). Exactly computing the local Lipschitz constant of ReLU networks. Advances in Neural Information Processing Systems, 33, 7344–7353. https://doi.org/10.48550/arXiv.2003.01219

15. Fazlyab, M., Robey, A., Hassani, H., Morari, M., & Pappas, G. J. (2019). Efficient and accurate estimation of Lipschitz constants for deep neural networks. Advances in Neural Information Processing Systems, 32.

16. Fazlyab, M., Morari, M., & Pappas, G. J. (2022). Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. IEEE Transactions on Automatic Control, 6(1), 1–15. https://doi.org/10.1109/TAC.2020.3046193

17. Combettes, P. L., & Pesquet, J.-C. (2020). Lipschitz certificates for layered network structures driven by averaged activation operators. SIAM Journal on Mathematics of Data Science, 2(2), 529–557. https://doi.org/10.1137/19M1272780

18. Meunier, L., Delattre, B., Araujo, A., & Allauzen, A. (2022). A dynamical system perspective for Lipschitz neural networks. Proceedings of the 39th International Conference on Machine Learning (PMLR, 162).

19. Li, M., Chen, K., Bai, Y., & Pei, J. (2024). Skeleton action recognition via graph convolutional network with self-attention module. Electronic Research Archive, 32(4), 2848–2864. https://doi.org/10.3934/era.2024129

20. Gunasekara, S. R., Li, W., Yang, J., & Ogunbona, P. O. (2025). Asynchronous joint-based temporal pooling for skeleton-based action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 35(1), 357–366. https://doi.org/10.1109/TCSVT.2024.3465845

21. Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1802.05957

22. Kim, H., Papamakarios, G., & Mnih, A. (2021). The Lipschitz constant of self-attention. Proceedings of the 38th International Conference on Machine Learning (PMLR, 139), 5562–5571. https://doi.org/10.48550/arXiv.2006.04710

23. Gama, F., Bruna, J., & Ribeiro, A. (2020). Stability properties of graph neural networks. IEEE Transactions on Signal Processing, 68, 5680–5695. https://doi.org/10.1109/TSP.2020.3026980

24. Kenlay, H., Thanou, D., & Dong, X. (2021). On the stability of graph convolutional neural networks under edge rewiring. ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing, 8513–8517. https://doi.org/10.1109/ICASSP39728.2021.9413474

25. Wang, Z., Ruiz, L., & Ribeiro, A. (2024). Stability to deformations of manifold filters and manifold neural networks. IEEE Transactions on Signal Processing, 72, 2130–2146. https://doi.org/10.1109/TSP.2024.3378379

26. Wang, Z., Cerviño, J., & Ribeiro, A. (2025). Generalization of geometric graph neural networks with Lipschitz loss functions. IEEE Transactions on Sigsnal Processing, 73, 1549–1561. https://doi.org/10.1109/TSP.2025.3553378