



1 Introduction: 長時間動画の生成における「崩壊」問題

動画生成モデルにおいて、15 秒を超えるような長時間生成は極めて困難な課題である。従来のモデルでは時間ステップが進むにつれて、物体の形状崩壊と呼ばれるアイデンティティの消失が頻発する。本レジュメでは、Seedance 2.0 が導入した全域注意制約 (**Global Attention Constraint**) がいかにして誤差の指数関数的増大を抑制し、長期間の一致性を数学的に保証しているかを力学系の観点から解説する。

Def. 1.1 (自己回帰的生成の不安定性): 動画の予測変数 \hat{z}_t が直前のフレーム \hat{z}_{t-1} のみに依存して生成される場合、各ステップで生じる微小な推論誤差が時間とともに累積し、非線形な系において指数関数的に増幅される現象を指す。

2 Mathematical Modeling: 誤差伝播の解析

動画生成プロセスを、潜在空間 \mathbb{R}^d 上の離散時間力学系として定式化し、その安定性を解析する。

2.1 自己回帰生成の力学方程式

従来の動画生成モデルは、一般にマルコフ連鎖 (**Markov Chain**) として記述される。すなわち、時刻 t における生成フレーム (または予測変数) \hat{z}_t は、直前の生成結果 \hat{z}_{t-1} を入力とする生成関数 F によって決定される。

$$\hat{z}_t = F(\hat{z}_{t-1}) + \xi_t \quad [1]$$

ここで、各変数の定義は以下の通りである。

- $F^1 : \mathbb{R}^d \rightarrow \mathbb{R}^d$: ニューラルネットワークによってパラメータ化された遷移関数 (生成モデル)。
- 独立同分布 (i.i.d.) なノイズ $\xi_t \sim \mathcal{N}(0, \Sigma)^2$: モデル化できない確率的なノイズ (固定誤差)。

2.2 テイラー展開による線形化解析

¹ Sora, Seedance といった AI 生成モデルで使われた F の構造は自己注意機構 (Self-Attention) + フィードフォワードネットワーク (MLP) であり, *i.e.* $y = \text{Attention}(z_{t-1}) + z_{t-1} \implies F(z_{t-1}) = \text{MLP}(\text{LayerNorm}(y)) + y$ である。ここでの F には膨大なパラメータが含まれている。これは z_{t-1} の各要素間の関連 (例えば「猫の手」と「猫の頭」の位置関係) を計算し、その後それらがどのように一緒に動くかを予測する。

² $\xi_t \sim \mathcal{N}(0, \Sigma)$ という表現において、 $\Sigma = \sigma^2 I$ は多次元正規分布の共分散行列 (Covariance Matrix) を表す。

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

生成された軌道 \hat{z}_t と、理想的な真の軌道 z_t との間の誤差 $\epsilon_t := \hat{z}_t - z_t$ を解析するため、生成関数 $F(\hat{z}_{t-1})$ を真の値 z_{t-1} の近傍で 1 次のテイラー展開（**Taylor Expansion**）を用いて線形近似する。

$$F(\hat{z}_{t-1}) \approx F(z_{t-1}) + \mathbf{J} \cdot (\hat{z}_{t-1} - z_{t-1}) \quad [2]$$

この近似式における各項の意味は以下の通りである。

1. 理想遷移: $F(z_{t-1}) \approx z_t$

理想的な学習が行われている場合、真の入力 z_{t-1} に対する関数の出力は、真の次の状態 z_t に等しいと仮定できる。

2. ヤコビ行列: $\mathbf{J} = \nabla F|_{z_{t-1}} \in \mathbb{R}^{d \times d}$

これは生成関数の入力に対する感度を表すヤコビ行列（**Jacobian Matrix**）であり、入力の微小な変化が主力にどの程度影響するか（拡大するか縮小するか）を決定する線形写像である。

3. 誤差項: $\epsilon_{t-1} = \hat{z}_{t-1} - z_{t-1}$

前ステップまでの累積誤差ベクトル。

これらを [1] の力学方程式に代入することで、誤差の発展方程式が得られる。

$$\begin{aligned} \hat{z}_t &\approx z_t + \mathbf{J} \cdot \epsilon_{t-1} + \xi_t \\ &\Downarrow \\ \underbrace{\hat{z}_t - z_t}_{\epsilon_t} &\approx \mathbf{J} \cdot \epsilon_{t-1} + \xi_t \end{aligned} \quad [3]$$

この線形漸化式 $\epsilon_t \approx \mathbf{J}\epsilon_{t-1} + \xi_t$ こそが、動画生成の安定性を分析する基礎となる。行列 \mathbf{J} の性質（特にそのスペクトル半径³ $\rho(\mathbf{J})$ ）が、誤差 ϵ_t が時間とともに減衰するか、あるいは爆発するかを決める要因となる。先ほどの誤差伝播の式 $\epsilon_t \approx \mathbf{J} \cdot \epsilon_{t-1}$ において、ヤコビ行列 \mathbf{J} のスペクトル半径 $\rho(\mathbf{J})$ は、「誤差の大きさ」を決定する。以下のケースに分けて考える。

- $\rho(\mathbf{J}) > 1$:
 - 誤差はステップごとに拡大されます。
 - これが「指数関数的発散（**Exponential Divergence**）」である。動画は数秒で崩壊される。
- $\rho(\mathbf{J}) < 1$:
 - 誤差はステップごとに縮小され、最終的に 0 になる。
 - 一見良さそうであるが、これは「情報が消える」ことを意味し、動画が静止画になったり、動きがなくなったりする。
- $\rho(\mathbf{J}) \approx 1$:
 - 誤差は拡大も縮小も言えない。
 - Seedance 2.0 が目指しているのがこの状態である。これにより、長時間にわたって動画の形（アイデンティティ **identity**）を保ちつつ、動き続けることができる。

³ 正方行列 A が n 個の固有値 $\lambda_1, \lambda_2, \dots, \lambda_n$ を持っているとする。スペクトル半径 $\rho(A)$ は次のように定義されます：

$$\rho(A) = \max(|\lambda_1|, |\lambda_2|, \dots, |\lambda_n|).$$

i.e. 原点から最も遠い固有値までの距離を表す。

Prop. 2.1 (誤差の指数的爆発): ヤコビ行列 \mathbf{J} のスペクトル半径（最大固有値の絶対値）を $\rho(\mathbf{J}) = \mu$ とし、 $\mu = 1 + \lambda$ （ただし $\lambda > 0$ ）であるとする。このとき、十分大きな時間 t において、累積誤差のノルムは以下のように成長する。

$$\|\epsilon_t\| \approx C \cdot e^{\lambda t} \quad [4]$$

ここで C は定数である。すなわち、誤差は時間とともに指数関数的に増大する。

Proof.

誤差の漸化式 $\epsilon_t = \mathbf{J}\epsilon_{t-1} + \xi_t$ を $t = 0$ まで再帰的に展開し、そのノルムを評価することで証明する。

■漸化式の展開（一般解の導出） 漸化式を繰り返し代入する。

$$\begin{aligned} \epsilon_1 &= \mathbf{J}\epsilon_0 + \xi_1 \\ \epsilon_2 &= \mathbf{J}(\underbrace{\mathbf{J}\epsilon_0 + \xi_1}_{\epsilon_1}) + \xi_2 = \mathbf{J}^2\epsilon_0 + \mathbf{J}\xi_1 + \xi_2 \\ &\vdots \\ \epsilon_t &= \mathbf{J}^t\epsilon_0 + \sum_{k=1}^t \mathbf{J}^{t-k}\xi_k \end{aligned} \quad [5]$$

ここで、右辺第 1 項は「初期誤差 ϵ_0 の影響」を表し、第 2 項は「各ステップで加わるノイズ ξ_k の和」を表す。

■スペクトル半径によるノルム評価 行列 \mathbf{J} の最大固有値に対応する固有ベクトル方向の成分が支配的になると仮定する。線形代数の基本性質より、十分大きな n に対して $\|\mathbf{J}^n\| \approx \rho(\mathbf{J})^n = \mu^n$ と近似できる。三角不等式 $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$ を用いて上界を評価する。

$$\begin{aligned} \|\epsilon_t\| &\leq \|\mathbf{J}^t\epsilon_0\| + \sum_{k=1}^t \|\mathbf{J}^{t-k}\xi_k\| \\ &\leq \|\mathbf{J}^t\| \|\epsilon_0\| + \sum_{k=1}^t \|\mathbf{J}^{t-k}\| \|\xi_k\| \\ &\leq \mu^t \|\epsilon_0\| + \bar{\sigma} \sum_{j=0}^{t-1} \mu^j \quad (\text{ここで } j = t - k, \forall k, \|\xi_k\| \leq \bar{\sigma} \text{ と置く}) \end{aligned} \quad [6]$$

■等比級数の和と指数関数近似 第 2 項は公比 μ の等比数列の和であるため、以下のように計算できる。

$$\sum_{j=0}^{t-1} \mu^j = \frac{\mu^t - 1}{\mu - 1} = \frac{(1 + \lambda)^t - 1}{\lambda} \quad (\text{ここで、} \mu = 1 + \lambda) \quad [7]$$

$\lambda > 0$ （拡大系）であるため、時間 t が大きいとき、誤差の増加速度は $\mu^t = (1 + \lambda)^t$ の項が支配的となる。ネイピア数の定義およびテイラー展開 $\ln(1 + x) \approx x$ ($x \ll 1$ の場合)⁵より、

$$(1 + \lambda)^t = \exp(\ln((1 + \lambda)^t)) = \exp(t \ln(1 + \lambda)) \approx \exp(\lambda t) \quad [8]$$

したがって、誤差全体は $e^{\lambda t}$ のオーダーで成長することが示された。 ■

⁴ Appendix A を参照

⁵ テイラー展開より、 $\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots \implies \ln(1 + x) \rightarrow 0 \ (x \rightarrow 0)$ 、よって $\ln(1 + x) \approx x$ となる。

3 Seedance 2.0: 多様体制約による誤差制御

Seedance 2.0 の核は、生成される z_t を特定の多様体 (**Manifold**) \mathcal{M} 上に拘束することにある。

3.1 ラグランジュ未定乗数法による定式化

全域注意機構 (Global Attention) により定義される制約関数 $C(z_t, z_0) = 0$ を導入し、生成時の目的関数にラグランジュ項を追加する。

■制約なしの生成 (元の目的関数) 通常の動画生成モデル (拡散モデル等) において、時刻 t で生成される潜在変数 z_t は、本来の生成ロス $\mathcal{L}_{\text{diff}}(z_t)$ を最小化するように最適化 (更新) される。

$$\min_{z_t} \mathcal{L}_{\text{diff}}(z_t) \quad [9]$$

しかし、これのみでは時間経過に伴い、モデルが直前のフレームとの連続性のみを重視し、初期フレーム z_0 のアイデンティティから徐々にドリフトしてしまう。

■多様体制約 (ハード制約) の導入 そこで、「生成される z_t は、常に初期フレーム z_0 と一貫性を保たなければならない」という強い制約を課す。これを $C(z_t, z_0) = 0$ (C とは、誤差関数⁶) と表す。幾何学的には、 $C = 0$ を満たす点 z_t の集合は高次元空間上の特定の曲面 (多様体: **Manifold**) を形成する。つまり、「生成軌道はこの曲面から逸脱してはならない」というルールである。

■ラグランジュ関数の構築 目的関数に制約関数 C とラグランジュ乗数 λ を掛けた項を追加し、新しい関数 \mathcal{L} を定義する。

$$\mathcal{L}(z_t, \lambda) = \mathcal{L}_{\text{diff}}(z_t) + \lambda \cdot C(z_t, z_0) \quad [10]$$

ここで λ は、制約を破ろうとする力に対して復元力を提供する役割を持つ。

■「復元力」の物理的解釈と勾配 式 (10) が最適解 (極小値) をとるための必要条件は、 z_t による勾配がゼロになることである。

$$\nabla_{z_t} \mathcal{L} = \nabla_{z_t} \mathcal{L}_{\text{diff}}(z_t) + \lambda \nabla_{z_t} C(z_t, z_0) = 0 \quad [11]$$

これを移項すると、力の釣り合いを示す方程式が得られる。

$$-\nabla_{z_t} \mathcal{L}_{\text{diff}}(z_t) = \lambda \nabla_{z_t} C(z_t, z_0) \quad [12]$$

- 左辺 ($-\nabla \mathcal{L}_{\text{diff}}$): 生成モデルが z_t を動かそうとする力 (動画の進行方向)。
- 右辺 ($\lambda \nabla C$): 多様体に対する法線ベクトル (垂直方向の力)。

生成プロセスが多様体から逸脱しようとする、ラグランジュ乗数 λ が「安全ロープの張力」として自動的に働き、多様体の垂直方向へ強制的に引き戻す。これが復元力 (**Restoring Force**) であり、15 秒以上の長時間動画の生成においても動画が崩壊しない数学的根拠である。

⁶ $C(z_t, z_0)$ を特徴の差異を計算する非負の誤差関数 (distance/penalty function) とする。

$$C(z_t, z_0) \begin{cases} = 0, & z_t \text{ が重要な特徴において } z_0 \text{ と完全に一致する場合 (制約を満たす)} \\ > 0, & z_t \text{ が崩壊し始め、} z_0 \text{ から逸脱した場合 (制約を破る)} \end{cases}$$

$$\text{i.e. } C(z_t, z_0) \geq 0$$

4 Derivation: 次線形誤差成長への抑制

Thm. 4.1 (\sqrt{t} 安定性の証明): 全域注意制約下では、誤差伝播のヤコビ行列 \mathbf{J} の最大固有値が実質的に 1 に固定され、累積誤差 $\|\epsilon_t\|$ は $t^{1/2}$ のオーダーで成長する。

Proof.

■制約の効果（接空間への射影） ラグランジュ未定乗数法を用いて制約 $C(z_t, z_0) = 0$ を課すことは、生成される潜在変数 z_t を特定の多様体（Manifold） \mathcal{M} 上に束縛することを意味する。この制約下において、誤差の伝播を支配するヤコビ行列 \mathbf{J} は、制約の復元力によって多様体 \mathcal{M} の接空間（Tangent Space） $\mathcal{T}_z\mathcal{M}$ へと誤差ベクトルを射影する実効的な演算子 $\mathbf{J}_{\text{eff}} = \mathcal{P}_{\mathcal{M}}\mathbf{J}$ として機能する。

■安定性の変容（固有値の正規化） 射影演算子 $\mathcal{P}_{\mathcal{M}}$ の働きにより、多様体に直交する方向（制約を破り、動画が崩壊する方向）への微小な誤差変動はラグランジュ乗数によって即座に減衰させられる（該当する固有値 $\ll 1$ ）。一方、多様体に接する方向（アイデンティティを保ったまま動画が進行する方向）への変動のみが保存される。これにより、ヤコビ行列の最大の固有値（スペクトル半径）は実質的に 1 に固定される。すなわち、指数増幅の要因であった λ が 0 に強制され、 $\rho(\mathbf{J}_{\text{eff}}) \leq \max(0, 1) = 1$ が成立する。従って、 $\rho(\mathbf{J}_{\text{eff}}) \approx 1$ となる。

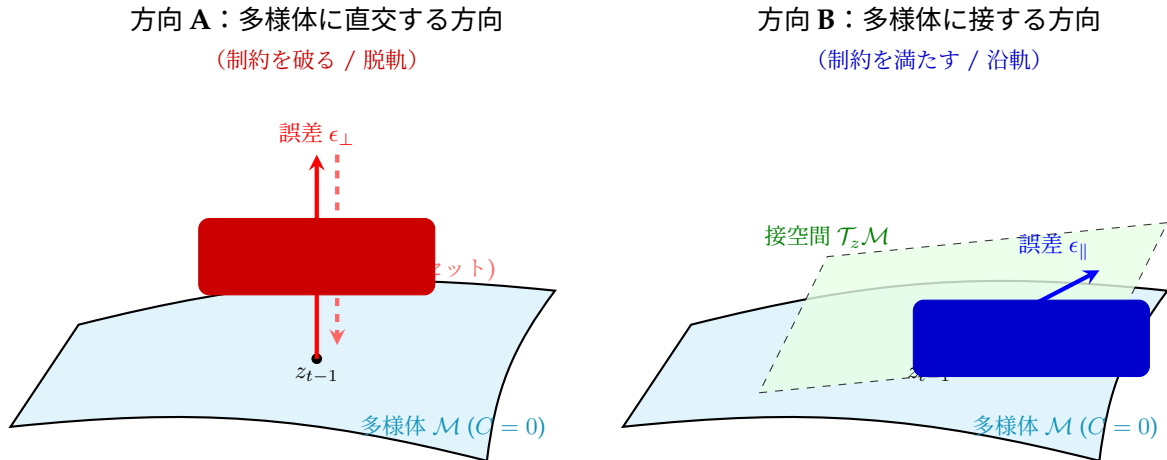


図 1: 多様体から逸脱する誤差（方向 A）は射影によって消去され、多様体に沿う誤差（方向 B）のみが等倍で保存されるため、最大増幅率は 1 となる。

■誤差方程式の簡略化 以前のセクションで導出した誤差の漸化式 $\epsilon_t \approx \mathbf{J}\epsilon_{t-1} + \xi_t$ に対して、実効ヤコビ行列 \mathbf{J}_{eff} を適用する。 $\rho(\mathbf{J}_{\text{eff}}) \approx 1$ であるため、 \mathbf{J}_{eff} を繰り返し掛けることによる指数関数的な拡大（ $\mu^t = 1^t = 1$ ）が消失する。その結果、誤差の更新式は過去の誤差を増幅することなく、単なる加算の形へと簡略化される。

$$\epsilon_t \approx \epsilon_{t-1} + \xi_t \quad [13]$$

■ランダムウォークの形成 初期誤差を $\epsilon_0 = 0$ と仮定し、簡略化された上記の漸化式を時刻 $t = 0$ から再帰的に展開すると、時刻 t における累積誤差は、各ステップで生じた不可避なノイズ ξ_i の単純な総和として表される。

$$\epsilon_t \approx \sum_{i=1}^t \xi_i \quad [14]$$

これは、誤差が特定の方向へ指数的に発散するのではなく、離散時間のブラウン運動（ランダムウォーク）に従って空間内を漂うことを意味している。

■分散の加法性 各ステップの生成ノイズ ξ_i は互いに独立同分布（i.i.d.）であり、平均ベクトル $\mathbb{E}[\xi_i]$ が $\mathbf{0}$ 、分散共分散行列が $\text{Var}(\xi_i) = \sigma^2 \mathbf{I}$ であると仮定する。確率論における独立な確率変数の和の性質より、累積誤差の分散は各ノイズの分散の和に等しくなる。

$$\text{Var}(\epsilon_t) = \text{Var}\left(\sum_{i=1}^t \xi_i\right) = \sum_{i=1}^t \text{Var}(\xi_i) = \sum_{i=1}^t \sigma^2 \mathbf{I} = t\sigma^2 \mathbf{I} \quad [15]$$

つまり、累積誤差の「広がり具合（分散）」は、時間 t に対して指数関数的ではなく、線形にしか増大しない。

■次線形性の導出 動画生成における実際の誤差の大きさ（振幅）は、分散そのものではなく、二乗平均平方根（RMS: Root Mean Square）として評価される。潜在空間の次元数を d とすると、誤差ノルムの期待値は分散共分散行列のトレース（対角成分の和）の平方根で近似できる。

$$\|\epsilon_t\|^2 = \epsilon_t^T \epsilon_t = \text{Tr}(\epsilon_t \epsilon_t^T) = \text{Tr}(\epsilon_t^T \epsilon_t)$$

よって、

$$\mathbb{E}[\|\epsilon_t\|^2] = \mathbb{E}[\text{Tr}(\epsilon_t^T \epsilon_t)] = \text{Tr}(\mathbb{E}[\epsilon_t \epsilon_t^T]) = \text{Tr}(\text{Var}(\epsilon_t))$$

したがって、

$$\underbrace{\mathbb{E}[\|\epsilon_t\|] \approx \sqrt{\mathbb{E}[\|\epsilon_t\|^2]}}_{\text{Jensen's ineq: } \mathbb{E}[\sqrt{X}] \leq \sqrt{\mathbb{E}[X]}} = \sqrt{\text{Tr}(\text{Var}(\epsilon_t))} = \sqrt{\text{Tr}(t\sigma^2 \mathbf{I})} = \sqrt{d \cdot t\sigma^2} = \sqrt{d}\sigma\sqrt{t}$$

ここで $\sqrt{d}\sigma$ は定数であるため、結果として $\|\epsilon_t\| \propto \sqrt{t} = t^{1/2}$ となる。 ■

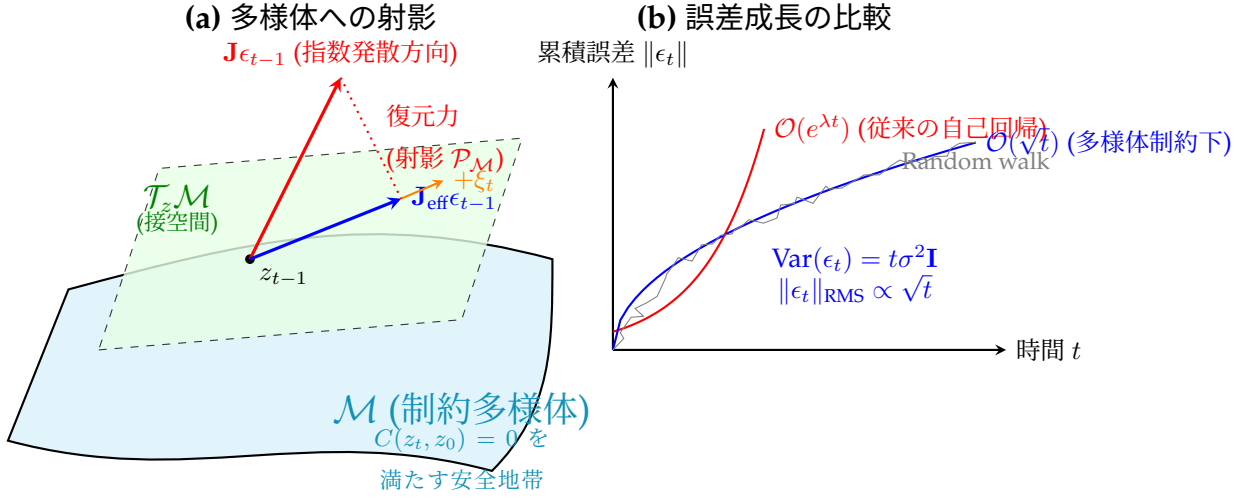


図 2: 多様体制約による実効ヤコビ行列の射影と、累積誤差の次線形 (\sqrt{t}) 成長への移行メカニズム。

Appendix A スペクトル半径とそのべき反復法

Ex 1: $\|\mathbf{J}^n\| \approx \rho(\mathbf{J})^n$ の証明

$\|\mathbf{J}^n\| \approx \rho(\mathbf{J})^n$ を証明せよ。ここで、 \mathbf{J} は $d \times d$ の行列である。 λ_i は \mathbf{J} の固有値であり、 $\rho(\mathbf{J}) = \max_{i \in \mathbb{N}} |\lambda_i|$ はスペクトル半径である。

Proof.

\mathbf{J} は $d \times d$ の行列であるため、固有ベクトルの集合 $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d\}$ を持つ。その対応する固有値は $\{\lambda_1, \lambda_2, \dots, \lambda_d\}$ である。任意のベクトル $\mathbf{x} \in \mathbb{R}^d$ は、これらの固有ベクトルの線形結合として表すことができる。

$$\mathbf{x} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \dots + c_d \mathbf{v}_d \quad [16]$$

ここで、 c_i はスカラー係数である。 \mathbf{J}^n を \mathbf{x} に作用させると、以下ようになる。

$$\begin{aligned} \mathbf{J}^n \mathbf{x} &= \mathbf{J}^n (c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \dots + c_d \mathbf{v}_d) \\ &= c_1 \mathbf{J}^n \mathbf{v}_1 + c_2 \mathbf{J}^n \mathbf{v}_2 + \dots + c_d \mathbf{J}^n \mathbf{v}_d \\ &= c_1 \lambda_1^n \mathbf{v}_1 + c_2 \lambda_2^n \mathbf{v}_2 + \dots + c_d \lambda_d^n \mathbf{v}_d \quad (\text{固有値の定義 } \mathbf{J}\mathbf{v} = \lambda\mathbf{v}) \end{aligned} \quad [17]$$

この式から、

$$\mathbf{J}^n \mathbf{x} = \lambda_m^n \left[c_1 \left(\frac{\lambda_1}{\lambda_m} \right)^n \mathbf{v}_1 + c_2 \left(\frac{\lambda_2}{\lambda_m} \right)^n \mathbf{v}_2 + \dots + c_d \left(\frac{\lambda_d}{\lambda_m} \right)^n \mathbf{v}_d \right]$$

となる。ここで、 λ_m は \mathbf{J} の固有値の中で最大の絶対値を持つものである。i.e. $\lambda_m = \rho(\mathbf{J})$ である。

したがって、 $\forall i > 1$ かつ $i \neq m$ に対し、 $\left| \frac{\lambda_i}{\lambda_m} \right| < 1$ であるため、 $\left(\frac{\lambda_i}{\lambda_m} \right)^n \rightarrow 0$ ($n \rightarrow \infty$) となる。

次に、ノルムを取ると、

$$\begin{aligned}
 \|\mathbf{J}^n \mathbf{x}\| &= \left\| \lambda_m^n \left[c_1 \left(\frac{\lambda_1}{\lambda_m} \right)^n \mathbf{v}_1 + c_2 \left(\frac{\lambda_2}{\lambda_m} \right)^n \mathbf{v}_2 + c_m \mathbf{v}_m + \cdots + c_d \left(\frac{\lambda_d}{\lambda_m} \right)^n \mathbf{v}_d \right] \right\| \\
 &= |\lambda_m|^n \left\| c_1 \left(\frac{\lambda_1}{\lambda_m} \right)^n \mathbf{v}_1 + c_2 \left(\frac{\lambda_2}{\lambda_m} \right)^n \mathbf{v}_2 + c_m \mathbf{v}_m + \cdots + c_d \left(\frac{\lambda_d}{\lambda_m} \right)^n \mathbf{v}_d \right\| \\
 &\approx |\lambda_m|^n \|c_m \mathbf{v}_m\|
 \end{aligned} \tag{18}$$

よって、十分大きな n に対して $\sup_{\|x\|=1} \|\mathbf{J}^n \mathbf{x}\| = \|\mathbf{J}^n\| \approx |\lambda_m|^n = \rho(\mathbf{J})^n$ が示された。 ■