A Gentle Introduction to Over-Smoothing

Andreas Roth

TU Dortmund University, 44227 Dortmund, Germany
andreas.roth@tu-dortmund.de

Abstract. Graph convolutions have gained popularity due to their ability to efficiently operate on data with an irregular geometric structure. However, graph convolutions cause over-smoothing, which refers to representations becoming more similar with increased depth. However, many different definitions and intuitions currently coexist, leading to research efforts focusing on incompatible directions. This paper attempts to align these directions by showing that over-smoothing is merely a special case of power iteration. This greatly simplifies the existing theory on over-smoothing, making it more accessible. Based on the theory, we provide a novel comprehensive definition of over-smoothing and show that over-smoothing is a solvable phenomenon.

Keywords: graph neural networks · message-passing neural networks · graph convolutions.

1 Introduction

When operating with message-passing neural networks on graph-structured data, over-smoothing describes a phenomenon in which node representations become more similar when the number of convolutional layers increases. Many research efforts provide theoretical insights on over-smoothing and methods to mitigate its effects [24, 23, 13, 2, 17, 14, 18]. However, due to the multitude of different theoretical insights and their complexity, different research efforts often use distinct definitions for over-smoothing, which are partly incompatible. In particular, some works study normalized representations [4, 15, 9] while others consider unnormalized representations [22, 20, 16]. Some define over-smoothing as the convergence to a constant state [17, 18, 22, 16, 20], others claim different limit distributions depending on the spectrum of the aggregation function [8, 1, 12, 25, 4, 15, 9].

To combine these strands, we show that the theory behind over-smoothing can be greatly simplified and reduced by connecting it to the classical power iteration method [7, 10, 11]. While our resulting insights are not novel, our novel proofs aim to make the theory more accessible to a broader part of the community. We first recap power iteration with its in-depth proof. We show that graph convolutions are a special case for which the dominant eigenvector takes a special form, namely a Kronecker product. Its properties lead to over-smoothing, for which we provide a novel theoretically founded definition.

2 Power Iteration

As the proof for over-smoothing of graph convolutions will be a special case, we first provide the detailed proof for the well-known power iteration [7, 10, 11]. Power iteration refers to the process where a vector, when repeatedly multiplied by a matrix, gets dominated by an eigenvector of the matrix that corresponds to the eigenvalue with the largest magnitude. The proof we provide mostly follows [6], but similar proofs are available in many textbooks. For any square matrix $\mathbf{M}$, its eigenvalues are denoted by $\lambda_1^{\mathbf{M}}, \dots, \lambda_n^{\mathbf{M}}$ and are sorted descending by their magnitude, i.e., $|\lambda_i^{\mathbf{M}}| \geq |\lambda_{i+1}^{\mathbf{M}}|$.

Proposition 1. *(Power Iteration [6]) Let* $\mathbf{S} \in \mathbb{R}^{p\times p}$ *be a matrix with* $|\lambda_1^{\mathbf{S}}| > |\lambda_2^{\mathbf{S}}|$ *and* $\mathbf{v}_1^{\mathbf{S}} \in \mathbb{R}^p$ *be an eigenvector corresponding to* $\lambda_1^{\mathbf{S}}$*. Further, let* $\mathbf{x}_0 \in \mathbb{R}^q$ *be a vector that has a non-zero component* c_1 *in direction* $\mathbf{v}_1^{\mathbf{S}}$*. Then,*

$$\frac{\mathbf{S}^k\mathbf{x}_0}{\|\mathbf{S}^k\mathbf{x}_0\|} = \beta_k \mathbf{v}_1^{\mathbf{S}} + \mathbf{r}_k \tag{1}$$

for some $\mathbf{r}_k \in \mathbb{R}^p$ *with* $\lim_{k\to\infty} \|\mathbf{r}_k\| = 0$ *and* $\beta_k = \frac{c_1}{|c_1|}\left(\frac{\lambda_1^{\mathbf{S}}}{|\lambda_1^{\mathbf{S}}|}\right)^k \frac{1}{\|\mathbf{v}_1^{\mathbf{S}}\|} \in \mathbb{R}$.

Proof. Let $\mathbf{S} = \mathbf{V}\mathbf{J}\mathbf{V}^{-1}$ be its Jordan decomposition, where $\mathbf{J} \in \mathbb{C}^{p\times p}$ is a block diagonal matrix containing the eigenvalues on its diagonal and $\mathbf{V} \in \mathbb{C}^{n\times n}$ contains the generalized eigenvectors as columns. As the generalized eigenvectors form a basis of $\mathbf{R}^n$, $\mathbf{x}_0$ can be decomposed as $\mathbf{x}_0 = c_1\mathbf{v}_1^{\mathbf{S}} + \dots + c_n\mathbf{v}_n^{\mathbf{S}}$ into a linear combination. This allows the following equalities:

$$\begin{aligned}
\frac{\mathbf{S}^k\mathbf{x}_0}{\|\mathbf{S}^k\mathbf{x}_0\|} &= \frac{(\mathbf{V}\mathbf{J}\mathbf{V}^{-1})^k(c_1\mathbf{v}_1^{\mathbf{S}} + \dots c_n\mathbf{v}_n^{\mathbf{S}})}{\|(\mathbf{V}\mathbf{J}\mathbf{V}^{-1})^k(c_1\mathbf{v}_1^{\mathbf{S}} + \dots c_n\mathbf{v}_n^{\mathbf{S}})\|} \\
&= \frac{\mathbf{V}\mathbf{J}^k(c_1\mathbf{e}_1 + \dots c_n\mathbf{e}_n)}{\|\mathbf{V}\mathbf{J}^k(c_1\mathbf{e}_1 + \dots c_n\mathbf{e}_n)\|} \\
&= \frac{c_1}{|c_1|}\left(\frac{\lambda_1^{\mathbf{S}}}{|\lambda_1^{\mathbf{S}}|}\right)^k \frac{\mathbf{V}(\frac{1}{\lambda_1^{\mathbf{S}}}\mathbf{J})^k\frac{1}{c_1}(c_1\mathbf{e}_1 + \dots c_n\mathbf{e}_n)}{\|\mathbf{V}(\frac{1}{\lambda_1^{\mathbf{S}}}\mathbf{J})^k\frac{1}{c_1}(c_1\mathbf{e}_1 + \dots c_n\mathbf{e}_n)\|}
\end{aligned} \tag{2}$$

The second equation uses the fact $\mathbf{V}^{-1}\mathbf{v}_k^{\mathbf{S}} = \mathbf{e}_k$, i.e., the natural basis vector pointing in direction k. As $\mathbf{J}$ is normalized by its unique largest entry $\lambda_1^{\mathbf{S}}$, it converges to

$$\lim_{k\to\infty}\left(\frac{1}{\lambda_1^{\mathbf{S}}}\mathbf{J}\right)^k = \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}. \tag{3}$$

Equation 2 then simplifies to

$$\frac{c_1}{|c_1|}\left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\mathbf{V}(\frac{1}{\lambda_1}\mathbf{J})^k\frac{1}{c_1}(c_1\mathbf{e}_1 + \dots c_n\mathbf{e}_n)}{\|\mathbf{V}(\frac{1}{\lambda_1}\mathbf{J})^k\frac{1}{c_1}(c_1\mathbf{e}_1 + \dots c_n\mathbf{e}_n)\|} = \frac{c_1}{|c_1|}\left(\frac{\lambda_1^{\mathbf{S}}}{|\lambda_1^{\mathbf{S}}|}\right)^k \frac{\mathbf{v}_1^{\mathbf{S}}}{\|\mathbf{v}_1^{\mathbf{S}}\|} + \mathbf{r}_k \tag{4}$$

with $\lim_{k\to\infty}\|\mathbf{r}_k\| = 0$. It converges to $\frac{\mathbf{v}_1^{\mathbf{S}}}{\|\mathbf{v}_1^{\mathbf{S}}\|}$ iff $\lambda_1^{\mathbf{S}} > 0$. □

3 Graph Convolutions as Power Iteration

This proof applies to many graph convolutions as they can be expressed as a matrix $\mathbf{S}$ that takes a particular form. Given a state $\mathbf{X} \in \mathbb{R}^{n\times d}$, its rows indicate data samples or nodes, and its columns describe features. Most graph convolutions apply a node mixing function $\mathbf{A} \in \mathbb{R}^{n\times n}$ that represents the graph structure and its edge weights, and a feature transformation $\mathbf{W} \in \mathbb{R}^{d\times d}$. Let $\text{vec}(\cdot)$ describe the operation that stacks the columns of a matrix into a vector. A graph convolution be expressed in vector notation

$$\text{vec}(\mathbf{AXW}) = (\mathbf{W}^T \otimes \mathbf{A})\text{vec}(\mathbf{X}) = \mathbf{S}\mathbf{x}_0 \tag{5}$$

using the Kronecker product $\otimes$ that is defined as $\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$. This formulation is commonly used to study over-smoothing [4, 9, 15] and other properties of graph convolutions [5, 14, 3]. The Kronecker product has a key spectral property affecting power iteration: All eigenvectors $\mathbf{v}^{\mathbf{S}}_{ij} = \mathbf{v}^{(\mathbf{W}^\mathbf{T})}_i \otimes \mathbf{v}^{\mathbf{A}}_j$ of $\mathbf{W}^T \otimes \mathbf{A}$ are Kronecker products of the eigenvectors of $\mathbf{A}$ and $\mathbf{W}^T$ with corresponding eigenvalue $\lambda^{\mathbf{A}}_i \lambda^{\mathbf{W}}_j$ [19]. This lets us state the reason behind over-smoothing in a clearer way than in previous works by substituting $\mathbf{v}^{\mathbf{S}}_1$:

Proposition 2. *(Power Iteration with a Kronecker Product) Let* $\mathbf{S} = \mathbf{W} \otimes \mathbf{A} \in \mathbb{R}^{(n\cdot d)\times(n\cdot d)}$ *for any* $\mathbf{W} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *with* $|\lambda^{\mathbf{S}}_1| > |\lambda^{\mathbf{S}}_2|$. *Let* $\mathbf{v}^{\mathbf{A}}_1, \mathbf{v}^{\mathbf{W}}_1$ *be two eigenvectors corresponding to* $\lambda^{\mathbf{A}}_1$ *and* $\lambda^{\mathbf{W}}_1$, *respectively. Further, let* $\mathbf{x}_0 \in \mathbb{R}^{n\cdot d}$ *be a vector that has a non-zero component* c_1 *in direction* $\mathbf{v}^{\mathbf{S}}_1 = \mathbf{v}^{\mathbf{W}}_1 \otimes \mathbf{v}^{\mathbf{A}}_1$. *Then,*

$$\frac{(\mathbf{W} \otimes \mathbf{A})^k \mathbf{x}_0}{\|(\mathbf{W} \otimes \mathbf{A})^k \mathbf{x}_0\|} = \beta_k \cdot \mathbf{v}^{\mathbf{W}}_1 \otimes \mathbf{v}^{\mathbf{A}}_1 + \mathbf{r}_k \tag{6}$$

for some $\mathbf{r}_k \in \mathbb{R}^{n\cdot d}$ *with* $\lim_{k\to\infty} \|\mathbf{r}_k\| = 0$ *and* $\beta_k = \frac{c_1}{|c_1|} \frac{\left(\frac{\lambda^{\mathbf{A}}_1 \lambda^{\mathbf{W}}_1}{|\lambda^{\mathbf{A}}_1 \lambda^{\mathbf{W}}_1|}\right)^k}{\|\mathbf{v}^{\mathbf{W}}_1 \otimes \mathbf{v}^{\mathbf{A}}_1\|} \in \mathbb{R}$.

Proof. Given that $|\lambda^{\mathbf{S}}_1| > |\lambda^{\mathbf{S}}_2|$, and $\lambda^{\mathbf{S}}_{i\cdot j} = \lambda^{\mathbf{A}}_i \cdot \lambda^{\mathbf{W}}_j$ for all $0 < i < n$ and $0 < j < d$, we have $|\lambda^{\mathbf{A}}_1| > |\lambda^{\mathbf{A}}_2|$ and $|\lambda^{\mathbf{W}}_1| > |\lambda^{\mathbf{W}}_2|$. The corresponding eigenvector $\mathbf{v}^{\mathbf{S}}_1 = \mathbf{v}^{\mathbf{A}}_1 \otimes \mathbf{v}^{\mathbf{W}}_1$ is the Kronecker product of the corresponding eigenvectors of $\mathbf{A}$ and $\mathbf{W}$. Substituting these in Proposition 1 results in our statement. □

The statement for any $\mathbf{W}$ and possibly repeated $\lambda^{\mathbf{W}}_1$ is similar, as all generalized eigenvectors of $\mathbf{W} \otimes \mathbf{A}$ corresponding to $\lambda^{\mathbf{S}}_1$ are of the form $\mathbf{u} \otimes \mathbf{v}^{\mathbf{A}}_1$ for different $\mathbf{u}$. To simplify this work, we provide the statement in Appendix A. The implications of this statement become clearer when looking into its matrix form:

Remark 1. (Power Iteration with a Kronecker Product in Matrix Notation) Stating Proposition 3 in matrix notation leads to

$$\frac{\mathbf{A}^k \mathbf{X} \mathbf{W}^k}{\|\mathbf{A}^k \mathbf{X} \mathbf{W}^k\|} = \beta_k \mathbf{v}^{\mathbf{A}}_1 \left(\mathbf{v}^{\mathbf{W}}_1\right)^T + \mathbf{R}_k \tag{7}$$

for $\text{vec}(\mathbf{X}) = \mathbf{x}_0$ and some $\mathbf{R}_k$ with $\lim_{k\to\infty} \|\mathbf{R}_k\| = 0$

Any graph convolution of this form amplifies the same signal across all feature columns, and the state gets closer to a rank one matrix, with each column becoming a multiple of $\mathbf{v}_1^{\mathbf{A}}$. This phenomenon has also been termed rank collapse [15]. It is commonly referred to as over-smoothing as the eigenvector $\mathbf{v}_1^{\mathbf{A}}$ is a smooth vector for typical choices of $\mathbf{A}$, e.g., it is $\mathbf{v}_1^{\mathbf{A}} = \mathbb{1}$ for the (weighted) mean aggregation, and $\mathbf{v}_1^{\mathbf{A}} = \mathbf{D}^{\frac{1}{2}}\mathbb{1}$ for the symmetrically normalized adjacency matrix [21]. The Dirichlet energy

$$E\left(\frac{\mathbf{A}^k\mathbf{X}\mathbf{W}^k}{\|\mathbf{A}^k\mathbf{X}\mathbf{W}^k\|}\right) = \text{tr}\left(\frac{\mathbf{A}^k\mathbf{X}\mathbf{W}^k}{\|\mathbf{A}^k\mathbf{X}\mathbf{W}^k\|}\mathbf{\Delta}\frac{\mathbf{A}^k\mathbf{X}\mathbf{W}^k}{\|\mathbf{A}^k\mathbf{X}\mathbf{W}^k\|}\right) \tag{8}$$

is frequently used to quantify over-smoothing, as $\mathbf{v}_1$ is in the nullspace of $\mathbf{\Delta}$, i.e., $\mathbf{\Delta}\mathbf{v}_1 = \mathbf{0}$. However, this requires different $\mathbf{\Delta}$ for different aggregation functions, as $\mathbf{v}_1^{\mathbf{A}}$ may be different. We propose a more general definition as a consequence of the theory:

Definition 1. *(Over-Smoothing) A sequence of matrices $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(k)} \in \mathbb{R}^{n\times d}$ over-smoothes if there exists a sequence of rank-one matrices $\mathbf{Y}^{(k)}, \dots, \mathbf{Y}^{(k)}$ such that*

$$\lim_{k\to\infty}\left\|\frac{\mathbf{X}^{(k)}}{\|\mathbf{X}^{(k)}\|} - \mathbf{Y}^{(k)}\right\| = 0 \tag{9}$$

4 Conclusion

We have shown that over-smoothing is a special case of power iteration, with the dominant eigenvector of graph convolutions $\mathbf{W}\otimes\mathbf{A}$ taking the form $\mathbf{v}_1^{\mathbf{W}}\otimes\mathbf{v}_1^{\mathbf{A}}$. As given in power iteration, normalization is required, and the limit distribution is not always the constant vector, as it depends on the dominant eigenvector of $\mathbf{A}$. To solve the underlying problem, it needs to be ensured that the dominant eigenvector $\mathbf{v}_1^{\mathbf{S}}$ is not a simple Kronecker product so that it can amplify different signals across feature columns. As pointed out before [15], one direction is to operate on multiple computational graphs $\mathbf{A}_1, \dots, \mathbf{A}_l$ with distinct feature transformations $\mathbf{W}_1, \dots, \mathbf{W}_l$:

$$\begin{aligned}\mathbf{S}\text{vec}(\mathbf{X}) &= (\mathbf{W}_1\otimes\mathbf{A}_1 + \dots + \mathbf{W}_l\otimes\mathbf{A}_l)\text{vec}(\mathbf{X}) \\ &= \text{vec}(\mathbf{A}_1\mathbf{X}\mathbf{W}_1^T + \dots + \mathbf{A}_l\mathbf{X}\mathbf{W}_l^T)\,.\end{aligned} \tag{10}$$

Acknowledgments. This research has been funded by the Federal Ministry of Education and Research of Germany under grant no. 01IS22094E WEST-AI.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cai, C., Wang, Y.: A note on over-smoothing for graph neural networks. arXiv preprint arXiv:2006.13318 (2020)
2. Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., Sun, X.: Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3438–3445 (2020)
3. Di Giovanni, F., Rusch, T.K., Bronstein, M.M., Deac, A., Lackenby, M., Mishra, S., Veličković, P.: How does over-squashing affect the power of gnns? arXiv preprint arXiv:2306.03589 (2023)
4. Giovanni, F.D., Rowbottom, J., Chamberlain, B.P., Markovich, T., Bronstein, M.M.: Understanding convolution on graphs via energies. Transactions on Machine Learning Research (2023)
5. Gu, F., Chang, H., Zhu, W., Sojoudi, S., El Ghaoui, L.: Implicit graph neural networks. Advances in Neural Information Processing Systems **33**, 11984–11995 (2020)
6. Knabner, P., Barth, W.: Lineare Algebra. Springer (2017)
7. Kowalewski, G.: Einführung in die Determinantentheorie einschließlich der unendlichen und der Fredholmschen Determinanten. Veit & comp. (1909)
8. Li, Q., Han, Z., Wu, X.M.: Deeper insights into graph convolutional networks for semi-supervised learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
9. Maskey, S., Paolino, R., Bacho, A., Kutyniok, G.: A fractional graph laplacian approach to oversmoothing. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 13022–13063. Curran Associates, Inc. (2023)
10. Miintz, L.: Solution direct de 1'equation seculaire et de quelques problemes analogues transcendents. Comptes Rendus de I'Academie des Sciences, Paris **156**, 43–6 (1913)
11. Mises, R., Pollaczek-Geiringer, H.: Praktische verfahren der gleichungsauflösung. ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik **9**(1), 58–77 (1929)
12. Oono, K., Suzuki, T.: Graph neural networks exponentially lose expressive power for node classification. In: International Conference on Learning Representations (2020)
13. Rong, Y., Huang, W., Xu, T., Huang, J.: Dropedge: Towards deep graph convolutional networks on node classification. In: International Conference on Learning Representations (2020)
14. Roth, A., Liebig, T.: Transforming pagerank into an infinite-depth graph neural network. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 469–484. Springer (2022)
15. Roth, A., Liebig, T.: Rank collapse causes over-smoothing and over-correlation in graph neural networks. In: The Second Learning on Graphs Conference (2023)
16. Rusch, T.K., Bronstein, M.M., Mishra, S.: A survey on oversmoothing in graph neural networks. arXiv preprint arXiv:2303.10993 (2023)
17. Rusch, T.K., Chamberlain, B., Rowbottom, J., Mishra, S., Bronstein, M.: Graph-coupled oscillator networks. In: International Conference on Machine Learning. pp. 18888–18909. PMLR (2022)

18. Rusch, T.K., Chamberlain, B.P., Mahoney, M.W., Bronstein, M.M., Mishra, S.: Gradient gating for deep multi-rate learning on graphs. In: The Eleventh International Conference on Learning Representations (2023)
19. Schacke, K.: On the kronecker product. Master's thesis, University of Waterloo (2004)
20. Scholkemper, M., Wu, X., Jadbabaie, A., Schaub, M.: Residual connections and normalization can provably prevent oversmoothing in gnns. arXiv preprint arXiv:2406.02997 (2024)
21. Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing **17**, 395–416 (2007)
22. Wu, X., Ajorlou, A., Wu, Z., Jadbabaie, A.: Demystifying oversmoothing in attention-based graph neural networks. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 35084–35106. Curran Associates, Inc. (2023)
23. Zhao, L., Akoglu, L.: Pairnorm: Tackling oversmoothing in gnns. In: International Conference on Learning Representations (2020)
24. Zhou, K., Huang, X., Li, Y., Zha, D., Chen, R., Hu, X.: Towards deeper graph neural networks with differentiable group normalization. Advances in neural information processing systems **33**, 4917–4928 (2020)
25. Zhou, K., Huang, X., Zha, D., Chen, R., Li, L., Choi, S.H., Hu, X.: Dirichlet energy constrained learning for deep graph neural networks. Advances in Neural Information Processing Systems **34**, 21834–21846 (2021)

A Appendix

Proposition 3. *(Power Iteration with a Kronecker Product) Let* $\mathbf{S} = \mathbf{W} \otimes \mathbf{A}$ *for* $\mathbf{W} \in \mathbb{R}^{d \times d}$ *and* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *with* $|\lambda_1^{\mathbf{A}}| > |\lambda_2^{\mathbf{A}}|$*. Let* $\mathbf{v}_1^{\mathbf{A}}$ *be an eigenvector corresponding to* $\lambda_1^{\mathbf{A}}$*. Further, let* $\mathbf{x}_0 \in \mathbb{R}^{n \cdot d}$ *be any vector that has a non-zero component in the direction of a generalized eigenvector* $\mathbf{v}_1^{\mathbf{S}}$ *corresponding to* $\lambda_1^{\mathbf{S}}$*. Then,*

$$\frac{(\mathbf{W} \otimes \mathbf{A})^k \mathbf{x}_0}{\|(\mathbf{W} \otimes \mathbf{A})^k \mathbf{x}_0\|} = \beta_k \cdot \mathbf{u} \otimes \mathbf{v}_1^{\mathbf{A}} + \mathbf{r}_k \tag{11}$$

for some $\mathbf{r}_k \in \mathbb{R}^{n \cdot d}$ *with* $\lim_{k \to \infty} \|\mathbf{r}_k\| = 0$*, bounded* β_k*, and some* $\mathbf{u} \in \mathbb{R}^d$*.*

Proof. This proof is similar to the proof of Proposition 1. However, $\pm\lambda_1^{\mathbf{S}}$ may occur multiple times, so there can be multiple Jordan blocks corresponding to $\pm\lambda_1^{\mathbf{S}}$, and they can have a size larger than one. Let p be the size of the largest Jordan block corresponding to $\lambda_1^{\mathbf{S}}$. Then, $\mathbf{J}^k$ will be dominated by $q_k = \binom{k}{p-1} \lambda_1^{S^{k-(p-1)}}$:

$$\lim_{k \to \infty} \left(\frac{1}{q_k} \mathbf{J} \right)^k = \begin{bmatrix} \begin{matrix} 0 & \dots & 0 & 1 \\ & \ddots & & 0 \\ \vdots & & & \vdots \\ 0 & \dots & & 0 \end{matrix} & & & & \\ & \ddots & & & \\ & & \begin{matrix} 0 & \dots & 0 & 1 \\ & \ddots & & 0 \\ \vdots & & & \vdots \\ 0 & \dots & & 0 \end{matrix} & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} . \tag{12}$$

The number of blocks containing a 1 is determined by the number of Jordan blocks corresponding to $\pm\lambda_1^{\mathbf{S}}$ with size p. Let there be i such blocks. We further know that all corresponding generalized eigenvectors are of the form $\mathbf{v}_{i \cdot p}^{\mathbf{S}} = \mathbf{v}_{i \cdot p}^{\mathbf{W}} \otimes \mathbf{v}_1^{\mathbf{W}}$. For eigenvalues constructed with $\lambda_2^{\mathbf{A}}$ it holds that $\lambda_2^{\mathbf{A}} \lambda_{i \cdot p} < \lambda_1^{\mathbf{A}} \lambda_{i \cdot p}$. This lets us simplify the statement:

$$\begin{aligned} \left(\frac{q_k}{|q_k|} \right)^k \frac{\mathbf{V}(\frac{1}{q_k}\mathbf{J})^k (c_1 \mathbf{e}_1 + \dots c_n \mathbf{e}_n)}{\|\mathbf{V}(\frac{1}{q_k}\mathbf{J})^k (c_1 \mathbf{e}_1 + \dots c_n \mathbf{e}_n)\|} &= \left(\frac{q_k}{|q_k|} \right)^k \frac{c_{1 \cdot p} \mathbf{v}_{1 \cdot p}^{\mathbf{S}} + \dots + c_{i \cdot p} \mathbf{v}_{i \cdot p}^{\mathbf{S}}}{\|c_{1 \cdot p} \mathbf{v}_{1 \cdot p}^{\mathbf{S}} + \dots + c_{i \cdot p} \mathbf{v}_{i \cdot p}^{\mathbf{S}}\|} + \mathbf{r}_k \\ &= \left(\frac{q_k}{|q_k|} \right)^k \frac{b \mathbf{u} \otimes \mathbf{v_1}^{\mathbf{A}}}{\|b \mathbf{u} \otimes \mathbf{v_1}^{\mathbf{A}}\|} + \mathbf{r}_k \end{aligned} \tag{13}$$

for $b = c_{1 \cdot p} \cdot \dots \cdot c_{i \cdot p}$, $\mathbf{u} = \mathbf{v}_{1 \cdot p} + \dots + \mathbf{v}_{i \cdot p}$, and $\lim_{k \to \infty} \|\mathbf{r}_k\| = 0$ which converges to $\frac{\mathbf{v_1}}{\|\mathbf{v_1}\|}$ iff $\lambda_1 > 0$. Setting $\beta_k = \left(\frac{q_k}{|q_k|}\right)^k \frac{1}{\|b\mathbf{u} \otimes \mathbf{v_1}^{\mathbf{A}}\|}$ leads to the desired statement. □