HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics 1–70

Learnability and Competition in High-Dimensional Multi-Component ICA

Eser İlke Genç[1] EGENC25@KU.EDU.TR

Samet Demir[1] SDEMIR20@KU.EDU.TR

Zafer Doğan[1,2] ZDOGAN@KU.EDU.TR

[1] *Machine Learning and Information Processing Group, KUIS AI Center*
[2] *Department of Electrical and Electronics Engineering*
Koç University
İstanbul, Turkey

Abstract

Independent Component Analysis (ICA) is a foundational tool for unsupervised representation learning, yet its high-dimensional theory remains largely limited to single-component recovery. We develop an asymptotically exact mean-field theory for multi-component online ICA, capturing the coupling from simultaneous learning and orthogonalization. In the high-dimensional limit, the overlap matrix between estimates and true components obeys a closed ODE system. This reveals an initialization-driven phase structure: a *decoupled regime*, where estimates align with distinct components and evolve nearly independently, and a *competition regime*, where overlapping initializations induce orthogonality-driven conflicts and delayed convergence. Explicit learnability and competition boundaries show larger higher-order moments and competition shrink the stable learning-rate window and induce a staircase in component recovery. Experiments on synthetic data and hyperspectral remote sensing data validate the predicted trajectories and phase behavior.

1. Introduction

ICA is a foundational *feature learning* framework for recovering latent independent components from linear mixtures [1–4]. After whitening, the remaining structure is necessarily *non-Gaussian*, so ICA relies on nonlinear score functions extracting information from *higher-order* moments [5–7]. In streaming settings, ICA is naturally implemented via stochastic gradient updates [8, 9], placing it within the broader theory of high-dimensional SGD [10–16] and making it a natural testbed for exact theories of high-dimensional online learning.

Despite substantial progress, a rigorous high-dimensional theory of online ICA remains largely limited to the single-component setting [9, 17–19]; however, practical ICA estimates many directions at once. This is not a straightforward extension: orthogonalization couples the updates, so the dynamics depend not only on the marginal evolution of each component, but also on competition between estimates, their initialization geometry, the step size, and the higher-order moments of the data.

This paper develops an asymptotically exact high-dimensional theory of multi-component online ICA. Our contributions are:

© E. Genç, S. Demir & Z. Doğan.

1. We give an exact mean-field theory for the *joint training dynamics of multiple ICA components*: the joint empirical measure converges to a deterministic weak-form PDE, and the component overlaps obey a closed ODE system.
2. We reveal an initialization-dependent phase structure: a *decoupled regime*, where estimates align with distinct components and evolve nearly independently, and a *competition regime*, where overlapping initializations induce orthogonality-driven conflicts and slow learning.
3. We derive explicit *learnability boundaries* and *competition conditions* in terms of data moments, learning rate, and initialization, showing when larger higher-order moments or competition force smaller learning rates and longer convergence times. Furthermore, we demonstrate that our theory remains valid in a practical ICA experiment: hyperspectral remote sensing.

Notation. We denote conditional expectation with respect to the filtration $\{\mathcal{F}_k\}_{k\geq 0}$ by $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_k]$, where $\mathcal{F}_k$ is the σ-algebra generated by all randomness up to iteration k, including the initialization and the samples $\{\boldsymbol{y}_s\}_{s<k}$. For a matrix $\boldsymbol{A}_t \in \mathbb{R}^{m\times n}$ indexed by iteration t, we denote its (i,j)-th entry by $A_{t,i,j}$, and its i-th row and j-th column by $\boldsymbol{A}_{t,i,:}$ and $\boldsymbol{A}_{t,:,j}$, respectively. $\mathrm{sgn}(x)$ is sign function and $\mathrm{Tr}(\cdot)$ is the trace operator for matrices.

2. Problem setting

2.1. Data model

We consider a high-dimensional online ICA model with ambient dimension n and a fixed number p of latent components, with p finite as $n \to \infty$. At each iteration k, we observe

$$\boldsymbol{y}_k = \frac{1}{\sqrt{n}}\boldsymbol{U}\boldsymbol{c}_k + \boldsymbol{a}_k, \tag{1}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p] \in \mathbb{R}^{n\times p}$ contains the unknown components to be recovered, and $\boldsymbol{c}_k = (c_{k,1}, \ldots, c_{k,p})^\top \in \mathbb{R}^p$ contains the latent source coefficients. The component vectors are deterministic, mutually orthogonal, and normalized as $\|\boldsymbol{u}_i\| = \sqrt{n}$. The coefficients $\{c_{k,i}\}$ are independent across both i and k, have zero mean and unit variance, and are non-Gaussian; their distributions may differ across components and are distinguished by higher-order moments. The Gaussian term $\boldsymbol{a}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I} - \frac{1}{n}\boldsymbol{U}\boldsymbol{U}^\top)$ serves only to whiten the observations, $\mathbb{E}_k[\boldsymbol{y}_k\boldsymbol{y}_k^\top] = \boldsymbol{I}$, so second-order statistics are uninformative. Consequently, the components are identifiable only through the higher-order non-Gaussian structure.

2.2. Online ICA via projected SGD

We estimate the components with a strictly online projected-SGD algorithm, processing one sample at a time without mini-batching. Let $\boldsymbol{X}_k \in \mathbb{R}^{p\times n}$ collect the estimate vectors $\boldsymbol{x}_{k,i}$, row-wise, satisfying $\|\boldsymbol{x}_{k,i}\| = \sqrt{n}$ for all $i \in \{1, \ldots, p\}$. Given $\boldsymbol{y}_k$, the SGD step is

$$\tilde{\boldsymbol{X}}_k = \boldsymbol{X}_k + \frac{\tau}{\sqrt{n}} f\Big(\frac{1}{\sqrt{n}}\boldsymbol{X}_k\boldsymbol{y}_k\Big)\boldsymbol{y}_k^\top - \frac{\tau}{n}\phi(\boldsymbol{X}_k), \tag{2}$$

where f is applied entrywise to the normalized projections $\frac{1}{\sqrt{n}}\boldsymbol{X}_k\boldsymbol{y}_k$ and extracts higher-order non-Gaussian structure. The map ϕ is applied entrywise and is the gradient of

the regularizer $\Phi(x) = \int \phi(x)\,dx$, allowing structural priors such as sparsity to enter the dynamics; τ is the learning rate. We then project back onto the row-orthogonal constraint,

$$\boldsymbol{X}_{k+1} = \text{Orthogonalize}(\tilde{\boldsymbol{X}}_k), \tag{3}$$

where Orthogonalize$(\cdot)$ returns mutually orthogonal rows, each with norm $\sqrt{n}$. For concreteness, we analyze Gram–Schmidt orthogonalization [20]. The resulting decoupling and competition regimes, however, are not artifacts of this choice: they persist under both sequential and symmetric orthogonalization schemes, as shown in Appendix H.

3. Main results

We begin by defining the central object of our analysis: the joint empirical measure of the estimates and the ground-truth feature vectors. For each coordinate index $\alpha \in \{1, \dots, n\}$, let $\bar{\boldsymbol{x}}_k^{(\alpha)} := \boldsymbol{X}_{k,:,\alpha}$ and $\bar{\boldsymbol{u}}^{(\alpha)} := \boldsymbol{U}_{\alpha,:}$. We define μ_k^n as a probability measure on $\mathbb{R}^p \times \mathbb{R}^p$ by

$$\mu_k^n(\boldsymbol{x}, \boldsymbol{u}) := \frac{1}{n} \sum_{\alpha=1}^{n} \delta\Big(\boldsymbol{x} - \bar{\boldsymbol{x}}_k^{(\alpha)},\, \boldsymbol{u} - \bar{\boldsymbol{u}}^{(\alpha)}\Big), \tag{4}$$

where $\delta(\boldsymbol{x} - \bar{\boldsymbol{x}}_k^{(\alpha)}, \boldsymbol{u} - \bar{\boldsymbol{u}}^{(\alpha)})$ denotes the Dirac measure at $(\bar{\boldsymbol{x}}_k^{(\alpha)}, \bar{\boldsymbol{u}}^{(\alpha)})$. The following theorem characterizes the time evolution of μ_k^n, which is our main theoretical result.

Theorem 1 (Informal) *Define $k = \lfloor tn \rfloor$ with $t \geq 0$. As $n \to \infty$, the empirical measure process $\{\mu_{\lfloor tn \rfloor}^n\}_{t \geq 0}$ converges weakly to a deterministic measure-valued process $\{\mu_t\}_{t \geq 0}$. For any bounded test function $\varphi \in C^3(\mathbb{R}^{2p})$, the limiting dynamics satisfy*

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle + \int_0^t \Big\langle \nabla_{\boldsymbol{x}} \varphi^\top \boldsymbol{\omega}_s, \mu_s \Big\rangle \mathrm{d}s + \frac{1}{2} \int_0^t \Big\langle \mathrm{Tr}\big(\boldsymbol{\Lambda}_s \nabla_{\boldsymbol{x}}^2 \varphi\big), \mu_s \Big\rangle \mathrm{d}s, \tag{5}$$

where the coefficients $\boldsymbol{\omega}_t$ and $\boldsymbol{\Lambda}_t$ depend on the macroscopic variables

$$\boldsymbol{Q}_t := \int_{\mathbb{R}^{2p}} \boldsymbol{x}\, \boldsymbol{u}^\top \mu_t \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{u}, \qquad \boldsymbol{R}_t := \int_{\mathbb{R}^{2p}} \boldsymbol{x} \phi(\boldsymbol{x})^\top \mu_t \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{u}. \tag{6}$$

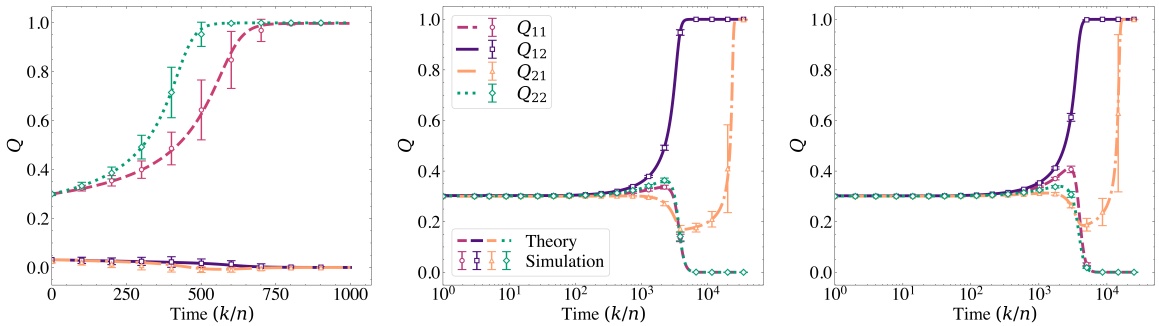
Appendix B derives the explicit coefficients $\boldsymbol{\omega}_t$ and $\boldsymbol{\Lambda}_t$, while Appendix C presents the formal theorem statement together with the assumptions and complete proof.

Theorem 1 yields the following two corollaries, which are useful to track the time-evolutions of the joint density P_t and overlap matrix $\boldsymbol{Q}_t$, respectively.

Corollary 1 *If μ_t admits a density $P_t(\boldsymbol{x}, \boldsymbol{u})$, then P_t satisfies the strong form of the PDE:*

$$\frac{\partial P_t}{\partial t} = -\nabla_{\boldsymbol{x}}\big(\boldsymbol{\omega}_t\, P_t\big) + \frac{1}{2}\, \mathrm{Tr}\big(\boldsymbol{\Lambda}_t\, \nabla_{\boldsymbol{x}}^2 P_t\big)\,. \tag{7}$$

In Appendix D, Figure 3 and 5 demonstrate that the PDE in Corollary 1 accurately tracks the empirical evolution in simulations.



(a) Decoupled alignment dynamics. (b) Competition between different estimates. (c) An estimate targets distinct components.

Figure 1: Learning dynamics in decoupled and competition regimes. Lines denote ODE predictions, and markers denote Monte Carlo simulations with error bars indicating two standard deviations across 20 runs for the overlap matrix $\boldsymbol{Q}_t \in \mathbb{R}^{2\times 2}$. In (a), initialization with $Q_{0,i,i} = 0.3$ yields decoupled dynamics at $\tau = 0.01$, $\beta_1 = 0.95$, $\beta_2 = 1$. In (b), competitive regime with $Q_{0,i,j} = 0.3$; $Q_{1,2}$ and $Q_{2,2}$ compete for the second component, $Q_{1,2}$ wins, at $\tau = 0.001$, $\beta_1 = 0.2, \beta_2 = 1$. In (c), we use the same initialization as (b) but different higher-order moments: first estimate competes across both components, $Q_{1,2}$ wins, at $\tau = 0.001$, $\beta_1 = 0.95, \beta_2 = 1$.

Corollary 2 *For $\phi(x) = 0$, choosing appropriate test functions in the weak formulation immediately yields a closed system of ODE's for the limiting overlap matrix $\boldsymbol{Q}_t$:*

$$\frac{d\boldsymbol{Q}_t}{dt} = -\frac{\tau^2}{2}\mathcal{T}(\boldsymbol{C}_t)\boldsymbol{Q}_t - \tau\mathcal{T}(\boldsymbol{M}_t)\boldsymbol{Q}_t + \tau\,\boldsymbol{\Psi}_t^\top, \tag{8}$$

where $\boldsymbol{C}_t$, $\boldsymbol{M}_t$, $\boldsymbol{\Psi}_t$ are defined in Theorem 1. Above, we define the matrix operator $\mathcal{T}(\boldsymbol{A}_t) := \mathrm{tril}(\boldsymbol{A}_t + \boldsymbol{A}_t^\top) - \mathrm{diag}(\boldsymbol{A}_t)$, where $\mathrm{tril}(\cdot)$ extracts the lower-triangular matrix by zeroing out all strictly upper-triangular entries of $\boldsymbol{A}_t$, and $\mathrm{diag}(\boldsymbol{A}_t)$ is the diagonal part of $\boldsymbol{A}_t$.

The overlap matrix $\boldsymbol{Q}_t$ captures all the learning dynamics of interest. Therefore, in the rest of this paper, we analyze the learning dynamics through the lens of $\mathbf{Q}_t$.

Example 1 *While Theorem 1 holds for any nonlinearity satisfying the stated assumptions, we focus on $f(x) = \pm x^3$ and $p = 2$ as an analytically tractable example. We adopt a data model used in the literature [21] that allows control of higher-order statistics through a single parameter. Specifically, the data samples $\boldsymbol{y}_k$ are generated using coefficients of the form: $c_{k,i} = \beta_i \varepsilon_{k,1} + \sqrt{1-\beta_i^2}\,\varepsilon_{k,2}$, where $\varepsilon_{k,1} \sim$ Rademacher, $\varepsilon_{k,2} \sim \mathcal{U}(-\sqrt{3}, \sqrt{3})$ for $\beta_1, \beta_2 \in [0, 1]$.*

Remark 1 *In the setting of Example 1, the ODE (8) for $\boldsymbol{Q}_t$ can be written explicitly in terms of the moments of $\boldsymbol{c}_t$. The resulting system is derived in Appendix E.*

Figure 1 confirms a strong agreement between our theoretical ODE (8) in Corollary 2 and empirical simulations across all evaluated initial conditions. Furthermore, this figure highlights two distinct regimes, defined by the initial overlap matrix $\boldsymbol{Q}_0$ as follows.

1. Decoupled dynamics. If each row and column of the initial overlap matrix $\boldsymbol{Q}_0$ has a single dominant $\mathcal{O}(1)$ entry, with all others negligible, then each estimator locks onto one ground-truth component. Consequently, $\boldsymbol{Q}_t$ becomes permutationally diagonal, and the coupled $p \times p$ dynamics reduce, to leading order, to p independent 1-D ODEs.

2. Competition dynamics. If some row or column of $\boldsymbol{Q}_0$ contains multiple $\mathcal{O}(1)$ entries, Gram–Schmidt coupling remains active at leading order and the $p \times p$ ODE system does not decouple. Multiple estimators may chase the same component, or one estimator may chase multiple components, producing competition, reorientation, and slower transients.

A diagonal initial overlap matrix (Figure 1(a)) yields a decoupled phase, where each estimate locks onto a distinct true feature without interference. In contrast, uniform initialization induces competition (Figure 1(b)): both estimates initially target the same component, and after one aligns strongly, orthogonalization forces the other to unlearn it before recovering the remaining feature. This coupling requires a smaller learning rate τ and slows convergence. Changing the data moments under the same initialization changes the competition pattern (Figure 1(c)): one estimate first aligns with *both* components before specializing, while the other captures the remaining feature asymptotically.

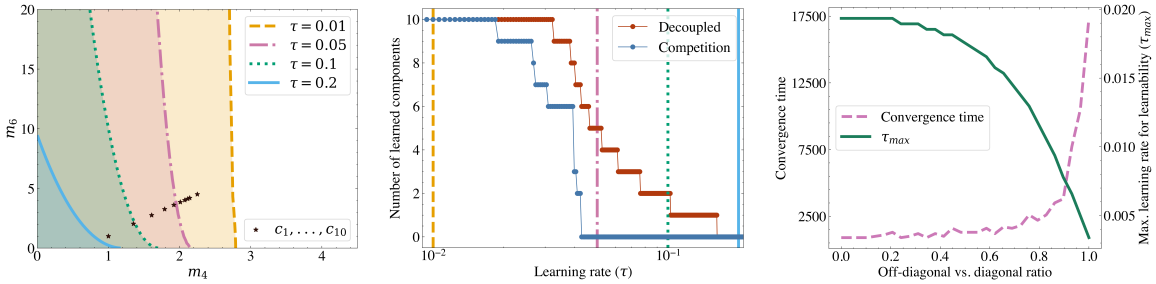
In Appendix G, we demonstrate that our ODE predictions accurately characterize the learning dynamics for a practical ICA experiment, hyperspectral remote sensing, even at a moderate spectral dimension of $n = 200$.

4. Steady state analysis

We can derive insights from the ODEs describing the learning dynamics; steady-state analysis of (8) reveals a nontrivial interplay between τ, initialization, and higher-order moments. We derive explicit learnability boundaries, show staircase behavior in the number of recoverable components, characterize the competition boundary, and quantify the instability it induces; details are deferred to Appendix F.

Learnability boundary and staircase behavior. Figure 2(a) shows the theoretically predicted learnability boundaries in the decoupled regime for different values of τ in the (m_4,m_6) plane, where m_4 and m_6 denote the fourth and sixth moment of the component. As τ increases, the stable recovery region contracts, while in the limit $\tau \to 0$ the boundary approaches the vertical asymptote $m_4 = 3$, indicating that the transition is controlled solely by the sign of the fourth cumulant $\kappa_4 = m_4 - 3$. The explicit boundary derivations and statements are in Appendix F.1. Figure 2(b) demonstrates that crossing these learnability boundaries produces discrete transitions in the number of recoverable components as a function of τ. Competition shifts the staircase toward smaller learning rates, implying that recovering more components requires more conservative learning rates.

Competition boundary and instability. The theoretical competition condition we derive and give in Appendix F.3 defines a separatrix in the overlap matrix's phase plane, visualized in Figure 7(a) by the red dashed lines, which partition the space into distinct basins of attraction. This boundary governs the dynamics in Figure 1(b): initialization determines which component first estimator learns. For the second estimate, Figure 7(b) initially points toward the second component, but as the first estimate reshapes the vector field, the second estimate is redirected toward the first component (Figure 7(c)), forcing it to unlearn before recovering the remaining feature (Figure 7(d)). In Figure 2(c), as the offdiagonal-to-digonal ratio of overlaps increases, the max stable learning rate required for



(a) Theoretical learnability boundaries for different learning rates.

(b) Number of learned components as a function of the learning rate.

(c) Instability of competition: effect of initial off-diagonal/diagonal ratio.

Figure 2: Steady-state analysis reveals learnability boundaries, staircase behavior, and competition-induced instability. In (a)–(b), component moments m_4, m_6 decrease with the component index. (a) Learnability regions for different learning rates, with scattered points showing the moments of $c_1, \ldots, c_{10}$; curves denote phase boundaries. (b) Number of recovered components versus τ for decoupled and competition regimes with $p = 10$; vertical lines mark the theoretical boundaries from Figure 2(a). In (c), as the off-diagonal/diagonal initialization ratio increases, the largest learning rate recovering all components, $\tau_{\max}$, decreases, while convergence time increases.

full recovery decreases significantly. This reduction in admissible τ, together with stronger competition between components, leads to a sharp increase in convergence time.

5. Conclusion

This work presents an asymptotically exact high-dimensional theory for multi-component online independent component analysis. We showed that the joint empirical measure converges to a deterministic nonlinear PDE, yielding a closed ODE system for the overlap dynamics. The resulting theory reveals two regimes,*decoupling* and *competition*, and yields explicit learnability boundaries linking learning rate, initialization, and higher-order moments. These boundaries predict staircase-like recovery behavior and quantify how competition reduces the stable learning-rate region and slows convergence. Synthetic simulations and hyperspectral remote sensing experiments validate our theoretical predictions regarding learning dynamics and phase transitions.

Acknowledgements

We acknowledge that this work is supported partially by TUBİTAK under project 124E063 in the ARDEB 1001 program. E.İ.G. is supported by an MS Scholarship (BIDEB 2210) from TUBİTAK and a Graduate Fellowship provided by Koç University. S.D. is supported by an AI Fellowship provided by KUIS AI Research Center and a PhD Scholarship (BIDEB 2211) from TUBİTAK.

References

[1] Bernard Ans, Jeanny Hérault, and Christian Jutten. Architectures neuromimétiques adaptatives : Détection de primitives. In *Cognitiva 85*, volume 2, pages 593–597, 1985.

[2] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36 (3):287–314, 1994.

[3] Shun-ichi Amari, Andrzej Cichocki, and Howard Hua Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, 1995.

[4] A Hyvärinen and E Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[5] Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[6] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[7] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[8] Pierre Ablin, Alexandre Gramfort, Jean-François Cardoso, and Francis Bach. Stochastic algorithms with descent guarantees for ICA. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2019.

[9] Chris Junchi Li and Michael I. Jordan. Stochastic approximation for online tensorial independent component analysis. In *Proceedings of Conference on Learning Theory*, 2021.

[10] David Saad and Sara A. Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. In *Advances in Neural Information Processing Systems*, 1995.

[11] Chuang Wang, Jonathan Mattingly, and Yue M Lu. Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and pca. *arXiv preprint arXiv:1712.04332*, 2017.

[12] Yazhen Wang and Shang Wu. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *Journal of Machine Learning Research*, 21(199):1–103, 2020.

[13] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Proceedings of Conference on Learning Theory*, 2021.

[14] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In *Advances in Neural Information Processing Systems*, 2022.

[15] Aukosh Jagannath, Taj Jones-McCormick, and Varnan Sarangian. High-dimensional limit theorems for SGD: Momentum and adaptive step-sizes. In *Proceedings of International Conference on Learning Representations*, 2026.

[16] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: an ODE for SGD learning dynamics on GLMs and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4), 2024.

[17] Chris Junchi Li, Zhaoran Wang, and Han Liu. Online ICA: Understanding global dynamics of nonconvex optimization via diffusion processes. In *Advances in Neural Information Processing Systems*, 2016.

[18] Chuang Wang and Yue M. Lu. The scaling limit of high-dimensional online independent component analysis. In *Advances in Neural Information Processing Systems*, 2017.

[19] Fabiola Ricci, Lorenzo Bardone, and Sebastian Goldt. Feature learning from non-Gaussian inputs: the case of Independent Component Analysis in high dimensions. In *Proceedings of International Conference on Machine Learning*, 2025.

[20] Erhard Schmidt. Theorie der linearen und nichtlinearen integralgleichungen i. teil: Entwicklung willkürlicherfunktionen nach systemen vorgeschriebener. *Mathematische Annalen*, 63:433–476, 1907.

[21] M. Oğuzhan Gültekin, Samet Demir, and Zafer Doğan. Learning Rate Should Scale Inversely with High-Order Data Moments in High-Dimensional Online Independent Component Analysis. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, 2025.

[22] Alan Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *Proceedings of Conference on Foundations of Computer Science*, 1996.

[23] Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems*, 1988.

[24] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[25] Nathalie Delfosse and Philippe Loubaton. Adaptive blind separation of independent sources: A deflation approach. *Signal Processing*, 45(1):59–83, 1995.

[26] J.-F. Cardoso. High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192, 1999.

[27] Aapo Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, 1997.

[28] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[29] Aapo Hyvärinen and Erkki Oja. Independent component analysis by general nonlinear hebbian-like learning rules. *Signal Processing*, 64(3):301–313, 1998.

[30] A. Hyvärinen. A family of fixed-point algorithms for independent component analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997.

[31] Arnab Auddy and Ming Yuan. Large-dimensional independent component analysis: Statistical optimality and computational tractability. *The Annals of Statistics*, 53(2): 477 – 505, 2025.

[32] Gleb Basalyga and Magnus Rattray. Statistical dynamics of on-line independent component analysis. *Journal of Machine Learning Research*, 4:1393–1410, 2003.

[33] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, 214:1–90, 2025.

[34] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[35] Henry P. McKean. Propagation of chaos for a class of non-linear parabolic equations. In *Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7)*, pages 41–57. Catholic University, 1967.

[36] A. Sznitman. Topics in propagation of chaos. *Lecture Notes in Mathematics*, pages 165–251, 1991.

[37] Krishnakumar Balasubramanian, Promit Ghosal, and Ye He. High-dimensional scaling limits and fluctuations of online least-squares SGD with smooth covariance. *The Annals of Applied Probability*, 35(5):2983–3045, 2025.

[38] Jiaqi Li, Zhipeng Lou, Johannes Schmidt-Hieber, and Wei Biao Wu. Statistical guarantees for high-dimensional stochastic gradient descent. In *Advances in Neural Information Processing Systems*, 2025.

[39] Chuang Wang and Yue M. Lu. Online learning for sparse PCA in high dimensions: Exact dynamics and phase transitions. In *2016 IEEE Information Theory Workshop*, 2016.

[40] Samet Demir and Zafer Dogan. Implicitly normalized online PCA: A regularized algorithm with exact high-dimensional dynamics. *arXiv preprint arXiv:2512.01231*, 2025.

[41] Laura Balzano, Yuejie Chi, and Yue M. Lu. Streaming PCA and subspace tracking: The missing data case. *Proceedings of the IEEE*, 106(8):1293–1310, 2018.

[42] Chuang Wang, Yonina C. Eldar, and Yue M. Lu. Subspace estimation from incomplete observations: A high-dimensional analysis. *IEEE Journal of Selected Topics in Signal Processing*, 12(6):1240–1252, 2018.

[43] Linghuan Meng and Chuang Wang. Training dynamics of nonlinear contrastive learning model in the high dimensional limit. *IEEE Signal Processing Letters*, 31:2535–2539, 2024.

[44] Chuang Wang, Hong Hu, and Yue Lu. A solvable high-dimensional model of GAN. In *Advances in Neural Information Processing Systems*, 2019.

[45] Andrew Bond and Zafer Dogan. Exploring the precise dynamics of single-layer GAN models. In *Advances in Neural Information Processing Systems*, 2024.

[46] Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher–student setup. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124010, 2020.

[47] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences U.S.A*, 115(33):E7665–E7671, 2018.

[48] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[49] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. In *Proceedings of Conference on Learning Theory*, 2023.

[50] H.P. Rosenthal. On the subspaces of $\mathrm{L}^p(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8(3):273–303, 1970.

[51] Pingyan Chen and Soo Hak Sung. Rosenthal type inequalities for random variables. *Journal of Mathematical Inequalities*, 14(2):305–318, 2020. doi: 10.7153/jmi-2020-14-20.

[52] Jonathan Novak. Three lectures on free probability. In *Random Matrices*, volume 65 of *MSRI Publications*, pages 309–383. 2014.

[53] Ronald A Fisher and John Wishart. The derivation of the pattern formulae of two-way partitions from those of simpler patterns. *Proceedings of the London Mathematical Society*, pages 195–208, 1931.

[54] Daniela Lupu, Ion Necoara, Joseph L. Garrett, and Tor Arne Johansen. Stochastic higher-order independent component analysis for hyperspectral dimensionality reduction. *IEEE Transactions on Computational Imaging*, 8:1184–1194, 2022.

[55] Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 220 band AVIRIS hyperspectral image data set: June 12, 1992 indian pine test site 3, 2015.

[56] Ramesh Naidu Annavarapu. Singular value decomposition and the centrality of Löwdin orthogonalizations. *American Journal of Computational and Applied Mathematics*, 3 (1):33–35, 2013.

[57] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[58] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Math. Program.*, 142(1–2):397–434, 2013.

[59] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Appendix Contents

A Related Work **13**

B Explicit derivations of drift and diffusion coefficients **13**

C Formal statement and proof of Theorem 1 **37**

D Evolution of the limiting density **55**

E Derivation of the ODEs for cubic nonlinearity: $f(x) = \pm x^3$ **58**

F Steady state analysis **62**
F.1 Learnability boundary 62
F.2 Staircase behavior 65
F.3 Competition boundary 65
F.4 Instability induced by competition 66
F.5 Further analysis of the decoupled regime 67

G Multi-component online ICA for hyperspectral remote sensing **68**
G.1 Pre-processing 68
G.2 Post-processing 69

H Competition in other orthogonalization schemes and nonlinearities **69**

Additional notation used in the Appendix

For vectors $\boldsymbol{x}_{k,i}$, the first index k denotes the iteration (time), and the second index $i \in \{1, \ldots, p\}$ denotes the component. For any vector $\boldsymbol{a}$, we write $a^{(\alpha)}$ for its α-th entry. Throughout the Appendix, $\|\cdot\|$ denotes the standard Euclidean (ℓ_2) norm for vectors, while $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the spectral and Frobenius norms for matrices, respectively.

We let $\mathrm{diag}(\boldsymbol{A}_t)$ be the diagonal matrix whose diagonal entries coincide with those of the matrix $\boldsymbol{A}_t$ or $\mathrm{diag}(\boldsymbol{v}_t)$ for the diagonal matrix formed by the entries of the vector $\boldsymbol{v}_t$. We define the matrix operator $\mathcal{T}(\boldsymbol{A}) \coloneqq \mathrm{tril}(\boldsymbol{A} + \boldsymbol{A}^\top) - \mathrm{diag}(\boldsymbol{A})$, where $\mathrm{tril}(\cdot)$ extracts the lower-triangular matrix by zeroing out all strictly upper-triangular entries of $\boldsymbol{A}$. The first derivative of a function f is denoted f'.

To simplify notation in sections B and E, the expectation $\mathbb{E}_{c,e}[\cdot]$ over the random variables c_i and $e_i \sim \mathcal{N}(0,1)$, is abbreviated as $\langle\cdot\rangle$.

Finally, we use standard asymptotic notation to describe scaling. We use Big O notation, $O(g(n))$, to denote that a term scales at most at the same order as $g(n)$. Conversely, we use Little o notation, $o(g(n))$, to denote terms that are asymptotically smaller than $g(n)$.

Appendix A. Related Work

Independent component analysis. ICA originated in blind source separation [1] and was formalized through independence, mutual information, and cumulant-based contrast functions [2], including fourth-order cumulant-tensor methods [22]. This line of work led to Infomax and neural learning rules [5, 23], online mutual-information and natural-gradient algorithms [3, 24], higher-order algebraic and cumulant-based contrasts [25, 26], and negentropy-based fixed-point methods, most notably FastICA [4, 6, 7, 27–30]. Despite this extensive literature, the high-dimensional streaming theory of ICA remains largely restricted to the single-component setting: prior works establish diffusion limits [17], finite-sample guarantees [9], scaling limits without orthogonalization-induced coupling [18], and recent analyses of FastICA and online SGD for recovering one non-Gaussian direction [19]. Complementary batch results characterize computational–statistical tradeoffs in large-dimensional ICA [31], but do not describe online learning trajectories. The closest multi-component dynamical analysis [32] studies an online Hebbian ICA algorithm [29] locally near metastable fixed points. In contrast, we provide a full-trajectory high-dimensional limit for orthogonality-coupled multi-component online ICA, yielding an empirical-measure PDE and closed overlap ODEs that capture both transient competition and steady-state learnability.

High-dimensional dynamics of SGD. Our work is also connected to recent advances in deriving closed dynamical descriptions of stochastic gradient methods in high dimensions. Depending on the model and scaling, these descriptions take the form of deterministic ODE or PDE systems for order parameters, homogenized stochastic processes, or empirical-measure limits [11, 14, 16, 33, 34]. Empirical-measure approaches trace back to the mean-field theory of interacting particle systems [35, 36], and have recently been used to analyze SGD and online-learning dynamics in linear regression [37, 38], streaming PCA and subspace tracking [39–42], contrastive learning [43], high-dimensional GANs [44, 45], and neural-network training [46–49]. We bring this perspective to multi-component ICA, where the central challenge is resolving the orthogonalization-induced interaction between learned directions. This interaction produces an initialization-dependent transition between decoupled learning and competition, a phenomenon that cannot arise in the single-component setting.

Appendix B. Explicit derivations of drift and diffusion coefficients

This section provides a detailed derivation of the drift and diffusion terms, which characterize the conditional mean and variance of the increments $\boldsymbol{x}_{k+1,i} - \boldsymbol{x}_{k,i}$. To facilitate the derivation, we transition from the compact matrix-form update to a row-wise vector representation. For clarity, we first derive the explicit expressions for the two-component case ($p = 2$), when $\phi(x) = 0$ before generalizing to an arbitrary number of components p and an arbitrary $\phi(x)$.

The complete online update for $p = 2$ is defined by first computing the stochastic gradient increments:

$$\tilde{\boldsymbol{x}}_{k,1} = \boldsymbol{x}_{k,1} + \frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}} \boldsymbol{y}_k^\top \boldsymbol{x}_{k,1}\right) \boldsymbol{y}_k, \tag{9}$$

$$\tilde{\boldsymbol{x}}_{k,2} = \boldsymbol{x}_{k,2} + \frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}} \boldsymbol{y}_k^\top \boldsymbol{x}_{k,2}\right) \boldsymbol{y}_k. \tag{10}$$

Using these increments, we immediately apply the Gram–Schmidt procedure:

$$\tilde{\tilde{\boldsymbol{x}}}_{k,1} = \tilde{\boldsymbol{x}}_{k,1}, \tag{11}$$

$$\tilde{\tilde{\boldsymbol{x}}}_{k,2} = \tilde{\boldsymbol{x}}_{k,2} - \frac{\tilde{\boldsymbol{x}}_{k,2}^\top \tilde{\boldsymbol{x}}_{k,1}}{\|\tilde{\boldsymbol{x}}_{k,1}\|^2} \tilde{\boldsymbol{x}}_{k,1}. \tag{12}$$

Finally, we normalize both vectors to obtain the final updated iterates:

$$\boldsymbol{x}_{k+1,1} = \frac{\tilde{\boldsymbol{x}}_{k,1}}{\|\tilde{\boldsymbol{x}}_{k,1}\|} \sqrt{n}, \tag{13}$$

$$\boldsymbol{x}_{k+1,2} = \frac{\tilde{\tilde{\boldsymbol{x}}}_{k,2}}{\|\tilde{\tilde{\boldsymbol{x}}}_{k,2}\|} \sqrt{n}. \tag{14}$$

We proceed by explicitly analyzing the second component first, as it exposes the key cross-interaction terms that come from the Gram-Schmidt orthogonalization step in (12), which are necessary for the general inductive step.

The remainder of this section is organized as follows: In Appendix B.1, we derive the dynamics for the second component. We then provide the derivations for the first component in Appendix B.2, followed by the analysis of the cross terms in Appendix B.3. Finally, in Appendix B.4 we derive the additional regularization terms and in Appendix B.5, we generalize these results to the case of p components and provide the final explicit expressions for the drift and diffusion terms in Equations (101), and (102).

B.1. Dynamics of the second estimate

First, for notational convenience, we define the stochastic gradient increments as

$$\boldsymbol{g}_{k,1} \coloneqq \frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}} \boldsymbol{y}_k^\top \boldsymbol{x}_{k,1}\right) \boldsymbol{y}_k, \tag{15}$$

$$\boldsymbol{g}_{k,2} \coloneqq \frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}} \boldsymbol{y}_k^\top \boldsymbol{x}_{k,2}\right) \boldsymbol{y}_k. \tag{16}$$

Then, substituting the update expressions into the orthogonalization step (12) yields the explicit form below for $\boldsymbol{x}_{k,2}$:

$$
\begin{aligned}
\tilde{\tilde{\boldsymbol{x}}}_{k,2} &= (\boldsymbol{x}_{k,2}+\boldsymbol{g}_{k,2}) - \frac{(\boldsymbol{x}_{k,2}+\boldsymbol{g}_{k,2})^\top(\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,1})}{\|\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,1}\|^2}(\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,1}),\\
&= (\boldsymbol{x}_{k,2}+\boldsymbol{g}_{k,2}) - \frac{(\overbrace{\boldsymbol{x}_{k,2}^\top\boldsymbol{x}_{k,1}}^{=0}+\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{g}_{k,1})}{\underbrace{\|\boldsymbol{x}_{k,1}\|^2}_{=n}+\|\boldsymbol{g}_{k,1}\|^2+2\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}}(\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,1}),\\
&= (\boldsymbol{x}_{k,2}+\boldsymbol{g}_{k,2}) - \frac{(\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{g}_{k,1})}{n(1+\|\boldsymbol{g}_{k,1}\|^2/n+2\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}/n)}(\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,1}).
\end{aligned}
\tag{17}
$$

We used the orthogonality condition $\boldsymbol{x}_{k,2}^\top\boldsymbol{x}_{k,1}=0$, which holds by construction from the previous update step, as well as the normalization $\|\boldsymbol{x}_{k,1}\|^2=n$. Next, we expand the denominator to first order in $1/n$. Higher-order terms involving products of three or more gradient increments contribute only at order $o(\frac{1}{n})$ and are omitted from the expansion:

$$
\begin{aligned}
\tilde{\tilde{\boldsymbol{x}}}_{k,2} = \boldsymbol{x}_{k,2}+\boldsymbol{g}_{k,2} &- \frac{\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{g}_{k,1}}{n}\,\boldsymbol{x}_{k,1}\\
&- \frac{\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}}{n}\Big(-2\,\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}/n\Big)\boldsymbol{x}_{k,1}\\
&- \frac{\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}}{n}\,\boldsymbol{g}_{k,1}+o\Big(\frac{1}{n}\Big).
\end{aligned}
\tag{18}
$$

We define the orthogonalized increment as $\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2} \coloneqq \tilde{\tilde{\boldsymbol{x}}}_{k,2}-\boldsymbol{x}_{k,2}$. Thus, we arrive at

$$
\begin{aligned}
\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2} = \boldsymbol{g}_{k,2} &- \frac{\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{g}_{k,1}}{n}\,\boldsymbol{x}_{k,1}\\
&- \frac{\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}}{n}\Big(-2\,\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}/n\Big)\boldsymbol{x}_{k,1}\\
&- \frac{\boldsymbol{x}_{k,2}^\top\boldsymbol{g}_{k,1}+\boldsymbol{g}_{k,2}^\top\boldsymbol{x}_{k,1}}{n}\,\boldsymbol{g}_{k,1}+o\Big(\frac{1}{n}\Big).
\end{aligned}
\tag{19}
$$

Following orthogonalization, the update is rescaled to norm $\sqrt{n}$, and the corresponding normalization factor is expanded around $\boldsymbol{x}_{k,2}$:

$$\begin{aligned}
\boldsymbol{x}_{k+1,2} &= \frac{\sqrt{n}}{\|\tilde{\tilde{\boldsymbol{x}}}_{k,2}\|}\tilde{\tilde{\boldsymbol{x}}}_{k,2} \ , \\
&= \frac{\sqrt{n}}{\|\boldsymbol{x}_{k,2} + \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|}(\boldsymbol{x}_{k,2} + \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}) \ , \\
&= (1 - \|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2/2n - \boldsymbol{x}_{k,2}^\top \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}/n)(\boldsymbol{x}_{k,2} + \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}) + o\left(\frac{1}{n}\right) , \qquad (20)
\end{aligned}$$

$$\boldsymbol{x}_{k+1,2} - \boldsymbol{x}_{k,2} = \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2} + (-\|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2/2n - \boldsymbol{x}_{k,2}^\top \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}/n)\boldsymbol{x}_{k,2} + o\left(\frac{1}{n}\right). \qquad (21)$$

Recall that $x_{k,i}^{(j)}$ denotes the $j-th$ entry/element of $\boldsymbol{x}_{k,i}$. With this convention, we obtain

$$\begin{aligned}
\tilde{\tilde{\Delta}}_{k,2}^{(j)} = g_{k,2}^{(j)} &- \frac{\boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,1} + \boldsymbol{g}_{k,2}^\top \boldsymbol{x}_{k,1} + \boldsymbol{g}_{k,2}^\top \boldsymbol{g}_{k,1}}{n} \ x_{k,1}^{(j)} \\
&+ \frac{\boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,1} + \boldsymbol{g}_{k,2}^\top \boldsymbol{x}_{k,1}}{n} \left(2\, \boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1}/n\right) x_{k,1}^{(j)} \\
&- \frac{\boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,1} + \boldsymbol{g}_{k,2}^\top \boldsymbol{x}_{k,1}}{n} \ g_{k,1}^{(j)}. \qquad (22)
\end{aligned}$$

After the normalization step, the entry-wise update satisfies

$$x_{k+1,2}^{(j)} - x_{k,2}^{(j)} = \tilde{\tilde{\Delta}}_{k,2}^{(j)} - \|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2 x_{k,2}^{(j)}/2n - x_{k,2}^{(j)}(\boldsymbol{x}_{k,2}^\top \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2})/n. \qquad (23)$$

Taking expectations yields the first moment of the entry-wise update for the second component:

$$\begin{aligned}
\mathbb{E}_k[x_{k+1,2}^{(j)} - x_{k,2}^{(j)}] &= \mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}] - \frac{\mathbb{E}_k[\|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2] x_{k,2}^{(j)}}{2n} - \frac{x_{k,2}^{(j)} \mathbb{E}_k[\boldsymbol{x}_{k,2}^\top \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}]}{n}, \\
&= \mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}] - \frac{\mathbb{E}_k[\|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2] x_{k,2}^{(j)}}{2n} - \frac{x_{k,2}^{(j)} \left(\sum_{\alpha=1}^n x_{k,2}^{(\alpha)} \mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(\alpha)}]\right)}{n}. \qquad (24)
\end{aligned}$$

Consequently, it remains to compute

$$\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}], \qquad \mathbb{E}_k[\|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2]. \qquad (25)$$

It follows from (22) that the analysis of the above terms require evaluating the following expectations:

$$\mathbb{E}_k[\boldsymbol{g}_{k,1}^{(j)}], \quad \mathbb{E}_k[\boldsymbol{g}_{k,2}^{(j)}], \quad \mathbb{E}_k[\boldsymbol{g}_{k,2}^\top \boldsymbol{g}_{k,1}], \quad \mathbb{E}_k[\boldsymbol{g}_{k,1} \boldsymbol{g}_{k,2}^\top]. \qquad (26)$$

In the following Subsections B.1.1 through B.1.3, we derive the expressions for these individual terms. Building on these results, we derive the expressions in (25) in Subsections B.1.4 and B.1.5. Finally in Subsection B.1.6, we state the final expression for (24).

B.1.1. Derivation of $\mathbb{E}_k[g_{k,i}^{(j)}]$

For the gradient $g_{k,i}^{(j)}$ we have

$$\begin{aligned} g_{k,i}^{(j)} &= \frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}}\, \boldsymbol{y}_k^\top \boldsymbol{x}_{k,i}\right) y_k^{(j)}, \\ &= \frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}}\left(\tfrac{1}{\sqrt{n}}\left(c_{k,1}\boldsymbol{u}_1^\top + c_{k,2}\boldsymbol{u}_2^\top\right) + \boldsymbol{a}_k^\top\right)\boldsymbol{x}_{k,i}\right) y_k^{(j)}, \\ &= \frac{\tau}{\sqrt{n}} f\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + \frac{1}{\sqrt{n}}\, \boldsymbol{a}_k^\top \boldsymbol{x}_{k,i}\right) y_k^{(j)}. \end{aligned} \tag{27}$$

Recall that we defined the cosine similarities between the estimates $\boldsymbol{x}_{k,i}$ and the true components $\boldsymbol{u}_l$, for $i, l \in \{1, \dots, p\}$, by $Q_{k,i,l} := \frac{\boldsymbol{x}_{\boldsymbol{k},\boldsymbol{i}}^\top \boldsymbol{u}_{\boldsymbol{l}}}{n}$. We define $e_{k,i}^{\setminus j} := \frac{1}{\sqrt{n}}(\boldsymbol{a}_k^\top \boldsymbol{x}_{k,i} - a_k^{(j)} x_{k,i}^{(j)})$. Thus we arrive at

$$g_{k,i}^{(j)} = \frac{\tau}{\sqrt{n}} f\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_{k,i}^{\setminus j} + \frac{a_k^{(j)} x_{k,i}^{(j)}}{\sqrt{n}}\right) y_k^{(j)}. \tag{28}$$

In the limit $n \to \infty$, the contribution of any single coordinate is negligible compared to the contribution of all coordinates. We therefore apply a first order Taylor expansion of $f(\cdot)$ around the leading-order term $c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_{k,i}^{\setminus j}$:

$$\begin{aligned} g_{k,i}^{(j)} = &\frac{\tau}{\sqrt{n}} f\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_{k,i}^{\setminus j}\right)\left(\frac{c_{k,1}u_1^{(j)}}{\sqrt{n}} + \frac{c_{k,2}u_2^{(j)}}{\sqrt{n}} + a_k^{(j)}\right) \\ &+ \frac{\tau}{\sqrt{n}} f'\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_{k,i}^{\setminus j}\right)\frac{a_k^{(j)} x_{k,i}^{(j)}}{\sqrt{n}}\left(\frac{c_{k,1}u_1^{(j)}}{\sqrt{n}} + \frac{c_{k,2}u_2^{(j)}}{\sqrt{n}} + a_k^{(j)}\right) \\ &+ o\left(\frac{1}{n}\right). \end{aligned} \tag{29}$$

Lemma 1 *For the random vector* $\begin{pmatrix} e_{k,i}^{\setminus j} \\ a_k^{(j)} \end{pmatrix}$, *its covariance matrix takes the form:*

$$\mathrm{Cov}\left(\begin{pmatrix} e_{k,i}^{\setminus j} \\ a_k^{(j)} \end{pmatrix}\right) = \begin{bmatrix} 1 - Q_{k,i,1}^2 - Q_{k,i,2}^2 & -\frac{1}{\sqrt{n}}\left(u_1^{(j)}Q_{k,i,1} + u_2^{(j)}Q_{k,i,2}\right) \\ -\frac{1}{\sqrt{n}}\left(u_1^{(j)}Q_{k,i,1} + u_2^{(j)}Q_{k,i,2}\right) & 1 \end{bmatrix}. \tag{30}$$

Proof We first compute the cross-correlation between $e_{k,i}^{\setminus j}$ and $a_k^{(j)}$. Substituting the definition of $e_{k,i}^{\setminus j}$, we have

$$\mathbb{E}_k[e_{k,i}^{\setminus j} a_k^{(j)}] = \mathbb{E}_k\left[\frac{1}{\sqrt{n}}\left(\sum_{\alpha \neq j} a_k^{(\alpha)} x_{k,i}^{(\alpha)}\right) a_k^{(j)}\right]. \tag{31}$$

Recalling that the vector $\boldsymbol{a}_k$ follows the distribution $\boldsymbol{a}_k \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I} - \frac{1}{n}(\boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top)\right)$, we substitute the off-diagonal covariance terms into the expectation:

$$\mathbb{E}_k[e_{k,i}^{\setminus j} a_k^{(j)}] = -\frac{1}{\sqrt{n}}\left(\frac{u_1^{(j)}\sum_{\alpha\neq j} u_1^{(\alpha)} x_{k,i}^{(\alpha)}}{n} + \frac{u_2^{(j)}\sum_{\alpha\neq j} u_2^{(\alpha)} x_{k,i}^{(\alpha)}}{n}\right). \tag{32}$$

Recognizing the parameters $Q_{k,i,1} = \frac{1}{n}\boldsymbol{x}_{k,i}^\top \boldsymbol{u}_1$ and $Q_{k,i,2} = \frac{1}{n}\boldsymbol{x}_{k,i}^\top \boldsymbol{u}_2$, this simplifies to

$$\mathbb{E}_k[e_{k,i}^{\setminus j} a_k^{(j)}] = -\frac{1}{\sqrt{n}}\left(u_1^{(j)} Q_{k,i,1} + u_2^{(j)} Q_{k,i,2}\right) + o\left(\frac{1}{\sqrt{n}}\right). \tag{33}$$

Next, we determine the variance of $e_{k,i}^{\setminus j}$.For notational convenience, let $\boldsymbol{x}_{k,i}^{(-j)}$ denote the vector $\boldsymbol{x}_{k,i}$ with the j-th entry set to zero. This allows us to express the sum as

$$\mathbb{E}_k[(e_{k,i}^{\setminus j})^2] = \mathbb{E}_k\left[\frac{1}{n}\left(\sum_{\alpha\neq j} a_k^{(\alpha)} x_{k,i}^{(\alpha)}\right)^2\right] = \mathbb{E}_k\left[\frac{1}{n}\left((\boldsymbol{x}_{k,i}^{(-j)})^\top \boldsymbol{a}_k\right)^2\right]. \tag{34}$$

Evaluating the expectation using the covariance matrix of $\boldsymbol{a}_k$ yields

$$\begin{aligned}\mathbb{E}_k[(e_{k,i}^{\setminus j})^2] &= \frac{1}{n}\left((\boldsymbol{x}_{k,i}^{(-j)})^\top \left(\boldsymbol{I} - \frac{1}{n}(\boldsymbol{u}_1\boldsymbol{u}_1^\top + \boldsymbol{u}_2\boldsymbol{u}_2^\top)\right)\boldsymbol{x}_{k,i}^{(-j)}\right),\\ &= \frac{1}{n}\left(\sum_{\alpha\neq j}(x_{k,i}^{(\alpha)})^2 - \frac{(\boldsymbol{x}_{k,i}^{(-j)})^\top \boldsymbol{u}_1\boldsymbol{u}_1^\top \boldsymbol{x}_{k,i}^{(-j)}}{n} - \frac{(\boldsymbol{x}_{k,i}^{(-j)})^\top \boldsymbol{u}_2\boldsymbol{u}_2^\top \boldsymbol{x}_{k,i}^{(-j)}}{n}\right).\end{aligned} \tag{35}$$

Applying $\sum_{\alpha\neq j}(x_{k,i}^{(\alpha)})^2 = \|\boldsymbol{x}_{k,i}\|^2 - O(1) = n - O(1)$, and $\frac{1}{n}\boldsymbol{u}_m^\top \boldsymbol{x}_{k,i}^{(-j)} = Q_{k,i,m} - O(\frac{1}{n})$ for $m \in \{1,2\}$, we obtain

$$\mathbb{E}_k[(e_{k,i}^{\setminus j})^2] = \frac{1}{n}\left(n - nQ_{k,i,1}^2 - nQ_{k,i,2}^2\right) + O\left(\frac{1}{n}\right), \tag{36}$$

$$= 1 - Q_{k,i,1}^2 - Q_{k,i,2}^2 + O\left(\frac{1}{n}\right). \tag{37}$$

Finally, we note that the variance of $a_k^{(j)}$ itself is normalized such that:$\mathbb{E}_k[(a_k^{(j)})^2] = 1 - O\left(\frac{1}{n}\right)$. Thus, omitting the terms of $O\left(\frac{1}{n}\right)$ yields the covariance matrix:

$$\mathrm{Cov}\left(\begin{pmatrix} e_{k,i}^{\setminus j} \\ a_k^{(j)}\end{pmatrix}\right) = \begin{bmatrix} 1 - Q_{k,i,1}^2 - Q_{k,i,2}^2 & -\frac{1}{\sqrt{n}}\left(u_1^{(j)} Q_{k,i,1} + u_2^{(j)} Q_{k,i,2}\right) \\ -\frac{1}{\sqrt{n}}\left(u_1^{(j)} Q_{k,i,1} + u_2^{(j)} Q_{k,i,2}\right) & 1 \end{bmatrix}. \tag{38}$$

■

By Lemma 1, the cross-covariance between $e_{k,i}^{\setminus j}$ and $a_k^{(j)}$ is $O(n^{-1/2})$. Therefore $e_{k,i}^{\setminus j}$ and $a_k^{(j)}$ may be treated as independent up to $O(n^{-1/2})$ accuracy in when evaluating any

term whose overall contribution is already $O(n^{-1/2})$. Moreover, we introduce an auxiliary random variable $e_i \sim \mathcal{N}(0,1)$ and represent $e_{k,i}^{\setminus j}$ as

$$e_{k,i}^{\setminus j} = \sqrt{1 - Q_{k,i,1}^2 - Q_{k,i,2}^2}\, e_i. \tag{39}$$

It follows from (29) that $\mathbb{E}[g_{k,i}^{(j)}]$ can be written as

$$\begin{aligned}
\mathbb{E}_k[g_{k,i}^{(j)}] &= \frac{\tau}{\sqrt{n}}\mathbb{E}\left[f\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_{k,i}^{\setminus j}\right)\left(\frac{c_{k,1}u_1^{(j)}}{\sqrt{n}} + \frac{c_{k,2}u_2^{(j)}}{\sqrt{n}} + a_k^{(j)}\right)\right] \\
&\quad + \frac{\tau}{\sqrt{n}}\mathbb{E}\left[f'\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_{k,i}^{\setminus j}\right)\frac{a_k^{(j)}x_{k,i}^{(j)}}{\sqrt{n}}\right. \\
&\qquad\qquad \left.\times\left(\frac{c_{k,1}u_1^{(j)}}{\sqrt{n}} + \frac{c_{k,2}u_2^{(j)}}{\sqrt{n}} + a_k^{(j)}\right)\right] + o\left(\frac{1}{n}\right), \\
&= \frac{\tau}{\sqrt{n}}L_1 + \frac{\tau}{\sqrt{n}}L_2 + o\left(\frac{1}{n}\right),
\end{aligned} \tag{40}$$

where

$$\begin{aligned}
L_2 &\coloneqq \mathbb{E}_k\left[f'\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_{k,i}^{\setminus j}\right)\frac{a_k^{(j)}x_{k,i}^{(j)}}{\sqrt{n}}\left(\frac{c_{k,1}u_1^{(j)}}{\sqrt{n}} + \frac{c_{k,2}u_2^{(j)}}{\sqrt{n}} + a_k^{(j)}\right)\right], \\
&= \frac{x_{k,i}^{(j)}}{\sqrt{n}}\left\langle f\left(c_{k,1}Q_{k,i,1} + c_{k,2}Q_{k,i,2} + e_i\sqrt{1 - Q_{k,i,1}^2 - Q_{k,i,2}^2}\right)\right\rangle + o\left(\frac{1}{\sqrt{n}}\right),
\end{aligned} \tag{41}$$

$$
\begin{aligned}
L_1 &\coloneqq \mathbb{E}_k\left[f\left(c_{k,1}Q_{k,i,1}+c_{k,2}Q_{k,i,2}+e_{k,i}^{\setminus j}\right)\left(\frac{c_{k,1}u_1^{(j)}}{\sqrt{n}}+\frac{c_{k,2}u_2^{(j)}}{\sqrt{n}}+a_k^{(j)}\right)\right],\\
&= \mathbb{E}_k\Bigg[f\ \Bigg(c_{k,1}Q_{k,i,1}+c_{k,2}Q_{k,i,2}+\sqrt{1-Q_{k,i,1}^2-Q_{k,i,2}^2}\,e_i\\
&\qquad\qquad -\frac{1}{\sqrt{n}}\left(u_1^{(j)}Q_{k,i,1}+u_2^{(j)}Q_{k,i,2}\right)a_k^{(j)}\Bigg)\left(\frac{c_{k,1}u_1^{(j)}}{\sqrt{n}}+\frac{c_{k,2}u_2^{(j)}}{\sqrt{n}}+a_k^{(j)}\right)\Bigg],\\
&= \frac{u_1^{(j)}}{\sqrt{n}}\Bigg(\Bigg\langle c_{k,1}f\left(c_1Q_{k,i,1}+c_2Q_{k,i,2}+e_i\sqrt{1-Q_{k,i,1}^2-Q_{k,i,2}^2}\right)\Bigg\rangle\\
&\qquad\qquad -Q_{k,i,1}\Bigg\langle f'\left(c_1Q_{k,i,1}+c_2Q_{k,i,2}+e_i\sqrt{1-Q_{k,i,1}^2-Q_{k,i,2}^2}\right)\Bigg\rangle\Bigg)\\
&\quad +\frac{u_2^{(j)}}{\sqrt{n}}\Bigg(\Bigg\langle c_{k,2}f\left(c_1Q_{k,i,1}+c_2Q_{k,i,2}+e_i\sqrt{1-Q_{k,i,1}^2-Q_{k,i,2}^2}\right)\Bigg\rangle\\
&\qquad\qquad -Q_{k,i,2}\Bigg\langle f'\left(c_1Q_{k,i,1}+c_2Q_{k,i,2}+e_i\sqrt{1-Q_{k,i,1}^2-Q_{k,i,2}^2}\right)\Bigg\rangle\Bigg)\\
&\quad +o\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
\tag{42}
$$

Henceforth, to streamline the notation, we introduce

$$\gamma_1 \coloneqq f\Big(c_1Q_{k,1,1}+c_2Q_{k,1,2}+e_1\sqrt{1-Q_{k,1,1}^2-Q_{k,1,2}^2}\Big), \tag{43}$$

$$\gamma_2 \coloneqq f\Big(c_1Q_{k,2,1}+c_2Q_{k,2,2}+e_2\sqrt{1-Q_{k,2,1}^2-Q_{k,2,2}^2}\Big). \tag{44}$$

More generally, for $i \in \{1,2\}$ we write

$$\gamma_i \coloneqq f\Big(c_1Q_{k,i,1}+c_2Q_{k,i,2}+e_i\sqrt{1-Q_{k,i,1}^2-Q_{k,i,2}^2}\Big). \tag{45}$$

Throughout, γ_i' denotes the derivative of f evaluated at the same argument as in γ_i. Finally, we arrive at

$$\mathbb{E}_k[g_{k,i}^{(j)}] = \frac{\tau}{n}\left(x_{k,i}^{(j)}\langle\gamma_i'\rangle + u_1^{(j)}\Big(\langle c_1\gamma_i\rangle - Q_{k,i,1}\langle\gamma_i'\rangle\Big) + u_2^{(j)}\Big(\langle c_2\gamma_i\rangle - Q_{k,i,2}\langle\gamma_i'\rangle\Big)\right) + o\left(\frac{1}{n}\right). \tag{46}$$

B.1.2. Derivation of $\mathbb{E}_k[\boldsymbol{g}_{k,1}^\top \boldsymbol{g}_{k,2}]$

Now, we consider the cross term $\mathbb{E}_k[\boldsymbol{g}_{k,1}^\top \boldsymbol{g}_{k,2}]$. Substituting the definitions of $\boldsymbol{g}_{k,1}$ and $\boldsymbol{g}_{k,2}$ into the expectation yields

$$\begin{aligned}
\mathbb{E}_k\Big[\boldsymbol{g}_{k,1}^\top \boldsymbol{g}_{k,2}\Big] &= \mathbb{E}_k\left[\frac{\tau^2}{n} f\left(\frac{\boldsymbol{y}_k^\top \boldsymbol{x}_{k,1}}{\sqrt{n}}\right) f\left(\frac{\boldsymbol{y}_k^\top \boldsymbol{x}_{k,2}}{\sqrt{n}}\right) \boldsymbol{y}_k^\top \boldsymbol{y}_k\right] + o\left(\frac{1}{n}\right), \\
&= \frac{\tau^2}{n}\,\mathbb{E}_k\Bigg[\Big(\gamma_1 + \gamma_1' \frac{a_k^{(j)} x_{k,1}^{(j)}}{\sqrt{n}}\Big)\Big(\gamma_2 + \gamma_2' \frac{a_k^{(j)} x_{k,2}^{(j)}}{\sqrt{n}}\Big) \\
&\qquad\qquad \times \Big(c_{k,1}^2 + c_{k,2}^2 + \sum_{\alpha=1}^{n} a_k^{(\alpha)} a_k^{(\alpha)}\Big)\Bigg], \\
&= \frac{\tau^2}{n}\,\mathbb{E}_k\Bigg[\Big(\gamma_1\gamma_2 + \gamma_1'\gamma_2' \frac{(a_k^{(j)})^2 x_{k,1}^{(j)} x_{k,2}^{(j)}}{n} + \gamma_2\,\gamma_1' \frac{a_k^{(j)} x_{k,1}^{(j)}}{\sqrt{n}} \\
&\qquad + \gamma_1\,\gamma_2' \frac{a_k^{(j)} x_{k,2}^{(j)}}{\sqrt{n}}\Big)\Big(c_{k,1}^2 + c_{k,2}^2 + \sum_{\alpha=1}^{n} a_k^{(\alpha)} a_k^{(\alpha)}\Big)\Bigg]. \qquad (47)
\end{aligned}$$

We begin by noting the following expectations:

$$\mathbb{E}_k\left[(a_k^{(j)})^2\right] = 1 + o(1), \quad \mathbb{E}_k\left[a_k^{(j)} e_{k,i}^{\setminus j}\right] = o(1), \quad \sum_{\alpha=1}^{n} \mathbb{E}_k\left[(a_k^{(\alpha)})^2\right] = n + O(1).$$

Following these, we can expand the expectation as follows:

$$\begin{aligned}
\mathbb{E}_k\Big[\boldsymbol{g}_{k,1}^\top \boldsymbol{g}_{k,2}\Big] = \frac{\tau^2}{n}\mathbb{E}_k &\left[\left(\gamma_1\gamma_2 + \gamma_1'\gamma_2' \frac{(a_k^{(j)})^2 x_{k,1}^{(j)} x_{k,2}^{(j)}}{n} + \gamma_2\gamma_1' \frac{a_k^{(j)} x_{k,1}^{(j)}}{n}\right.\right. \\
&\qquad \left.\left. + \gamma_2'\gamma_1 \frac{a_k^{(j)} x_{k,2}^{(j)}}{n}\right)\left(c_{k,1}^2 + c_{k,2}^2\right)\right] \\
+ \frac{\tau^2}{n}\mathbb{E}_k &\left[\left(\gamma_1\gamma_2 + \gamma_1'\gamma_2' \frac{(a_k^{(j)})^2 x_{k,1}^{(j)} x_{k,2}^{(j)}}{n} + \gamma_2\gamma_1' \frac{a_k^{(j)} x_{k,1}^{(j)}}{n}\right.\right. \\
&\qquad \left.\left. + \gamma_2'\gamma_1 \frac{a_k^{(j)} x_{k,2}^{(j)}}{n}\right)\left(\sum_{\alpha=1}^{n} (a_k^{(\alpha)})^2\right)\right]. \qquad (48)
\end{aligned}$$

Since the contributions from the auxiliary terms are of order $o(1/n)$, we obtain the final expression:

$$\mathbb{E}_k[\boldsymbol{g}_{k,1}^\top \boldsymbol{g}_{k,2}] = \tau^2 \langle \gamma_1\gamma_2 \rangle + o(1). \qquad (49)$$

B.1.3. Derivation of $\mathbb{E}_k[\boldsymbol{g}_{k,1}\boldsymbol{g}_{k,2}^\top]$

Substituting the definitions of $\boldsymbol{g}_{k,1}$ and $\boldsymbol{g}_{k,2}$ yields

$$\begin{aligned}\mathbb{E}_k[\boldsymbol{g}_{k,1}\boldsymbol{g}_{k,2}^\top] &= \mathbb{E}_k\left[\frac{\tau^2}{n} f\left(\frac{\boldsymbol{y}_k^\top \boldsymbol{x}_{k,1}}{\sqrt{n}}\right) f\left(\frac{\boldsymbol{y}_k^\top \boldsymbol{x}_{k,2}}{\sqrt{n}}\right)\boldsymbol{y}_k\boldsymbol{y}_k^\top\right] + o\left(\frac{1}{n}\right),\\ &= \frac{\tau^2}{n}\mathbb{E}_k\Bigg[\Bigg(\gamma_1\gamma_2 + \gamma_1'\gamma_2'\frac{(a_k^{(j)})^2 x_{k,1}^{(j)} x_{k,2}^{(j)}}{n}\\ &\qquad + \gamma_2\gamma_1'\frac{a_k^{(j)}x_{k,1}^{(j)}}{n} + \gamma_2'\gamma_1\frac{a_k^{(j)}x_{k,2}^{(j)}}{n}\Bigg)\boldsymbol{y}_k\boldsymbol{y}_k^\top\Bigg].\end{aligned} \tag{50}$$

We define the matrices $\boldsymbol{K} \coloneqq \boldsymbol{c}_k\boldsymbol{c}_k^\top$ and $\boldsymbol{A} \coloneqq \boldsymbol{a}_k\boldsymbol{a}_k^\top$. Their expectations are given by

$$\mathbb{E}_k[\boldsymbol{A}] = \boldsymbol{I} - \frac{\boldsymbol{U}\boldsymbol{U}^\top}{n}, \quad \text{and} \quad \mathbb{E}_k[\boldsymbol{K}] = \boldsymbol{I}. \tag{51}$$

Consequently, we arrive at the following for $\boldsymbol{y}_k\boldsymbol{y}_k^\top$:

$$\boldsymbol{y}_k\boldsymbol{y}_k^\top = \frac{\boldsymbol{U}\boldsymbol{K}\boldsymbol{U}^\top}{n} + \boldsymbol{A} + \frac{\boldsymbol{U}\boldsymbol{c}\boldsymbol{a}_k^\top}{\sqrt{n}} + \frac{\boldsymbol{a}_k\boldsymbol{c}^\top\boldsymbol{U}}{\sqrt{n}}. \tag{52}$$

Substituting this into Equation (50) yields the final form of the desired expectation:

$$\begin{aligned}\mathbb{E}_k[\boldsymbol{g}_{k,1}\boldsymbol{g}_{k,2}^\top] &= \frac{\tau^2}{n}\mathbb{E}_k[\gamma_1\gamma_2\boldsymbol{A}] + o\left(\frac{1}{n}\right),\\ &= \frac{\tau^2}{n}\langle\gamma_1\gamma_2\rangle\left(\boldsymbol{I} - \frac{\boldsymbol{U}\boldsymbol{U}^\top}{n}\right) + o\left(\frac{1}{n}\right),\\ &= \frac{\tau^2}{n}\langle\gamma_1\gamma_2\rangle\boldsymbol{I} + o\left(\frac{1}{n}\right).\end{aligned} \tag{53}$$

B.1.4. Derivation of $\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}]$

Taking the expectation of (22), we arrive at the following:

$$\begin{aligned}\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}] = \mathbb{E}_k[g_{k,2}^{(j)}] &- x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,2}^{(\alpha)}\mathbb{E}_k[g_{k,1}^{(\alpha)}]}{n} - x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}\mathbb{E}_k[g_{k,2}^{(\alpha)}]}{n}\\ &+ 2x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,2}^{(\alpha)}x_{k,1}^{(\alpha)}\mathbb{E}_k[(g_{k,1}^{(\alpha)})^2]}{n^2} + 2x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n (x_{k,1}^{(\alpha)})^2\mathbb{E}_k[g_{k,1}^{(\alpha)}g_{k,2}^{(\alpha)}]}{n^2}\\ &- \frac{x_{k,2}^{(j)}\mathbb{E}_k[(g_{k,1}^{(j)})^2]}{n} - \frac{x_{k,1}^{(j)}\mathbb{E}_k[g_{k,2}^{(j)}g_{k,1}^{(j)}]}{n} - x_{k,1}^{(j)}\frac{\mathbb{E}_k[\boldsymbol{g}_{k,2}^\top\boldsymbol{g}_{k,1}]}{n}.\end{aligned} \tag{54}$$

Moreover the expectations $\mathbb{E}_k[(g_{k,i}^{(j)})^2]$, $\mathbb{E}_k[g_{k,1}^{(j)}g_{k,2}^{(j)}]$ and $\mathbb{E}_k[\boldsymbol{g}_{k,1}^\top \boldsymbol{g}_{k,2}]$ were found to be independent of index (j), they can be taken outside the summations, yielding

$$\begin{aligned}
\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}] &= \mathbb{E}_k[g_{k,2}^{(j)}] - x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,2}^{(\alpha)}\mathbb{E}_k[g_{k,1}^{(\alpha)}]}{n} - x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}\mathbb{E}_k[g_{k,2}^{(\alpha)}]}{n} \\
&\quad + 2x_{k,1}^{(j)}\frac{\tau^2\langle\gamma_1^2\rangle\overbrace{\boldsymbol{x}_{k,2}^\top\boldsymbol{x}_{k,1}}^{=0}}{n^3} + 2x_{k,1}^{(j)}\frac{\tau^2\langle\gamma_1\gamma_2\rangle\, n}{n^3} \\
&\quad - \frac{x_{k,2}^{(j)}\tau^2\langle\gamma_1^2\rangle}{n^2} - \frac{x_{k,1}^{(j)}\tau^2\langle\gamma_1\gamma_2\rangle}{n^2} - x_{k,1}^{(j)}\frac{\tau^2\langle\gamma_1\gamma_2\rangle}{n}, \\
&= \mathbb{E}_k[g_{k,2}^{(j)}] - \underbrace{x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,2}^{(\alpha)}\mathbb{E}_k[g_{k,1}^{(\alpha)}]}{n}}_{T_1} - \underbrace{x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}\mathbb{E}_k[g_{k,2}^{(\alpha)}]}{n}}_{T_2} \\
&\quad + x_{k,1}^{(j)}\tau^2\langle\gamma_1\gamma_2\rangle(-\frac{1}{n}+\frac{1}{n^2}) - \frac{x_{k,2}^{(j)}\tau^2\langle\gamma_1^2\rangle}{n^2}.
\end{aligned} \tag{55}$$

We evaluate the α-summation terms labeled T_1 and T_2 separately:

$$\begin{aligned}
T_1 &= x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,2}^{(\alpha)}\mathbb{E}_k[g_{k,1}^{(\alpha)}]}{n}, \\
&= \frac{x_{k,1}^{(j)}\tau}{n\sqrt{n}}\sum_{\alpha=1}^n x_{k,2}^{(\alpha)}\Bigg(\frac{1}{\sqrt{n}}x_{k,1}^{(\alpha)}\langle\gamma_1^{'}\rangle + \frac{u_1^{(\alpha)}}{\sqrt{n}}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) \\
&\qquad\qquad + \frac{u_2^{(\alpha)}}{\sqrt{n}}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Bigg), \\
&= \frac{x_{k,1}^{(j)}\tau}{n^{3/2}}\Bigg(\frac{Q_{k,2,1}n}{\sqrt{n}}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) \\
&\qquad\qquad + \frac{Q_{k,2,2}n}{\sqrt{n}}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Bigg), \\
&= \frac{x_{k,1}^{(j)}\tau}{n}\Bigg(Q_{k,2,1}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) \\
&\qquad\qquad + Q_{k,2,2}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Bigg),
\end{aligned} \tag{56}$$

$$
\begin{aligned}
T_2 &= x_{k,1}^{(j)} \frac{\sum_{\alpha=1}^{n} x_{k,1}^{(\alpha)} \mathbb{E}_k[g_{k,2}^{(\alpha)}]}{n}, \\
&= \frac{x_{k,1}^{(j)} \tau}{n} \Bigg(Q_{k,1,1}(\langle c_1 \gamma_2 \rangle - Q_{k,2,1} \langle \gamma_2^{'} \rangle) \\
&\qquad\qquad + Q_{k,1,2}((\langle c_2 \gamma_2 \rangle - Q_{k,2,2} \langle \gamma_2^{'} \rangle)) \Bigg).
\end{aligned} \tag{57}
$$

Finally, we arrive at the following for the final expression of $\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}]$:

$$
\begin{aligned}
\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}] = &- \frac{x_{k,1}^{(j)} \tau^2 \langle \gamma_1 \gamma_2 \rangle}{n} \\
&+ \frac{\tau}{\sqrt{n}} \Bigg(\frac{1}{\sqrt{n}} x_{k,2}^{(j)} \langle \gamma_2^{'} \rangle + \frac{u_1^{(j)}}{\sqrt{n}} \left(\langle c_1 \gamma_2 \rangle - Q_{k,2,1} \langle \gamma_2^{'} \rangle \right) \\
&\qquad\qquad + \frac{u_2^{(j)}}{\sqrt{n}} \left(\langle c_2 \gamma_2 \rangle - Q_{k,2,2} \langle \gamma_2^{'} \rangle \right) \Bigg) \\
&- \frac{x_{k,1}^{(j)} \tau}{n} \left(Q_{k,1,1} \left(\langle c_1 \gamma_2 \rangle - Q_{k,2,1} \langle \gamma_2^{'} \rangle \right) + Q_{k,1,2} \left(\langle c_2 \gamma_2 \rangle - Q_{k,2,2} \langle \gamma_2^{'} \rangle \right) \right) \\
&- \frac{x_{k,1}^{(j)} \tau}{n} \left(Q_{k,2,1} \left(\langle c_1 \gamma_1 \rangle - Q_{k,1,1} \langle \gamma_1^{'} \rangle \right) + Q_{k,2,2} \left(\langle c_2 \gamma_1 \rangle - Q_{k,1,2} \langle \gamma_1^{'} \rangle \right) \right) \\
&+ o\left(\frac{1}{n} \right).
\end{aligned} \tag{58}
$$

B.1.5. Derivation of $\mathbb{E}_k[\|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2]$

Following Equation (22), we can obtain the squared form as follows:

$$
\begin{aligned}
\tilde{\tilde{\Delta}}_{k,2}^{(j)2} = &(g_{k,2}^{(j)})^2 + (x_{k,1}^{(j)})^2 \left(\frac{\boldsymbol{x}_{k,2}^{\top} \boldsymbol{g}_{k,1} + \boldsymbol{g}_{k,2}^{\top} \boldsymbol{x}_{k,1}}{n} \right)^2 \\
&- 2 x_{k,1}^{(j)} g_{k,2}^{(j)} \left(\frac{\boldsymbol{x}_{k,2}^{\top} \boldsymbol{g}_{k,1} + \boldsymbol{g}_{k,2}^{\top} \boldsymbol{x}_{k,1}}{n} \right) + o\left(\frac{1}{n} \right).
\end{aligned} \tag{59}
$$

Recalling that $\mathbb{E}_k[\boldsymbol{g_{k,1}} \boldsymbol{g_{k,2}}^{\top}]$ is diagonal (as derived in subsection B.1.3) taking the expectation yields

$$\mathbb{E}_k[(\tilde{\tilde{\Delta}}_{k,2}^{(j)})^2] = \mathbb{E}_k[(g_{k,2}^{(j)})^2] - 2\frac{x_{k,1}^{(j)}(x_{k,2}^{(j)}\mathbb{E}_k[g_{k,2}^{(j)}g_{k,1}^{(j)}] + x_{k,1}^{(j)}\mathbb{E}_k[(g_{k,2}^{(j)})^2])}{n} + x_{k,1}^{(j)2}\frac{1}{n^2}\Big(\mathbb{E}_k[(g_{k,1}^{(j)})^2]\underbrace{\sum_\alpha (x_{k,2}^{(\alpha)})^2}_{=\,n} + \mathbb{E}_k[g_{k,2}^{(j)2}]\underbrace{\sum_\alpha (x_{k,1}^{(\alpha)})^2}_{=\,n} + 2\mathbb{E}_k[g_{k,1}^{(j)}g_{k,2}^{(j)}]\underbrace{\sum_\alpha x_{k,1}^{(\alpha)}x_{k,2}^{(\alpha)}}_{=\,0}\Big) . \tag{60}$$

Substituting our previous results (49) and (53) into the expression, we arrive at

$$\begin{aligned}\mathbb{E}_k[(\tilde{\tilde{\Delta}}_{k,2}^{(j)})^2] &= \frac{\tau^2\langle\gamma_2^2\rangle}{n} - 2\frac{x_{k,1}^{(j)}(x_{k,2}^{(j)}\tau^2\langle\gamma_2\gamma_1\rangle/n + x_{k,1}^{(j)}\tau^2\langle\gamma_2^2\rangle/n)}{n} + \frac{\tau^2(x_{k,1}^{(j)})^2}{n^2}(\langle\gamma_1^2\rangle + \langle\gamma_2^2\rangle),\\ &= \frac{\tau^2\langle\gamma_2^2\rangle}{n} - 2\frac{x_{k,1}^{(j)}(x_{k,2}^{(j)}\tau^2\langle\gamma_2\gamma_1\rangle)}{n^2} + (x_{k,1}^{(j)})^2\frac{\tau^2}{n^2}(\langle\gamma_1^2\rangle - \langle\gamma_2^2\rangle) .\end{aligned} \tag{61}$$

Finally, this allows us to directly compute the expectation $\mathbb{E}_k[||\tilde{\tilde{\mathbf{\Delta}}}_{k,2}||^2]$ as follows:

$$\begin{aligned}\mathbb{E}_k[||\tilde{\tilde{\mathbf{\Delta}}}_{k,2}||^2] = \mathbb{E}_k[\sum_{\alpha=1}^n(\tilde{\tilde{\Delta}}_{k,2}^{(\alpha)})^2] &= \tau^2\langle\gamma_2^2\rangle - 2\frac{\tau^2\langle\gamma_2\gamma_1\rangle\overbrace{\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}x_{k,2}^{(\alpha)}}^{=\,0}}{n^2},\\ &\quad + \frac{\tau^2}{n^2}(\langle\gamma_1^2\rangle - \langle\gamma_2^2\rangle)\overbrace{\sum_{\alpha=1}^n(x_{k,1}^{(\alpha)})^2}^{=\,n},\\ &= \tau^2\langle\gamma_2^2\rangle + \frac{\tau^2}{n}(\langle\gamma_1^2\rangle - \langle\gamma_2^2\rangle),\\ &= \tau^2\langle\gamma_2^2\rangle + o\left(\frac{1}{n}\right) .\end{aligned} \tag{62}$$

B.1.6. Final expressions for the second component

Recall that we obtained the following in (24); we restate the equation here for convenience:

$$\mathbb{E}_k[x_{k+1,2}^{(j)} - x_{k,2}^{(j)}] = \mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}] - \mathbb{E}_k[\|\tilde{\tilde{\mathbf{\Delta}}}_{k,2}\|^2]x_{k,2}^{(j)}/2n - x_{k,2}^{(j)}\Big(\sum_{\alpha=1}^n x_{k,2}^{(\alpha)}\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(\alpha)}]\Big)/n. \tag{63}$$

Having determined the first two terms involving the expectation of $\tilde{\tilde{\mathbf{\Delta}}}_{k,2}$ in Subsections B.1.4 and B.1.5, we now proceed to compute the last term:

$$
\begin{aligned}
&x_{k,2}^{(j)}\frac{1}{n}\sum_{\alpha=1}^{n}x_{k,2}^{(\alpha)}\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(\alpha)}] = \\
&x_{k,2}^{(j)}\frac{1}{n}\sum_{\alpha=1}^{n}x_{k,2}^{(\alpha)}\Bigg(-\frac{x_{k,1}^{(\alpha)}\tau^2\langle\gamma_1\gamma_2\rangle}{n} + \frac{\tau}{\sqrt{n}}\Big(\frac{1}{\sqrt{n}}x_{k,2}^{(\alpha)}\langle\gamma_2^{'}\rangle + \frac{u_1^{(\alpha)}}{\sqrt{n}}(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle) \\
&\qquad\qquad + \frac{u_2^{(\alpha)}}{\sqrt{n}}(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle)\Big) \\
&\qquad - \frac{x_{k,1}^{(\alpha)}\tau}{n}\Big(Q_{k,1,1}(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle) + Q_{k,1,2}(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle)\Big) \\
&\qquad - \frac{x_{k,1}^{(\alpha)}\tau}{n}\Big(Q_{k,2,1}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) + Q_{k,2,2}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Big)\Bigg). \quad (64)
\end{aligned}
$$

Realize that the terms involving $\sum_{\alpha=1}^{n} x_{k,2}^{(\alpha)}x_{k,1}^{(\alpha)} = 0$ because of the orthogonality constraint of the previous iteration step, thus we arrive at the following for the corresponding term:

$$
\begin{aligned}
x_{k,2}^{(j)}\frac{1}{n}\sum_{\alpha=1}^{n}x_{k,2}^{(\alpha)}\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(\alpha)}] &= x_{k,2}^{(j)}\frac{1}{n}\frac{\tau}{\sqrt{n}}\Bigg(\frac{1}{\sqrt{n}}\langle\gamma_2^{'}\rangle\sum_{\alpha=1}^{n}x_{k,2}^{(\alpha)}x_{k,2}^{(\alpha)} \\
&\qquad + \frac{\sum_{\alpha=1}^{n}x_{k,2}^{(\alpha)}u_1^{(\alpha)}}{\sqrt{n}}(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle) \\
&\qquad + \frac{\sum_{\alpha=1}^{n}x_{k,2}^{(\alpha)}u_2^{(\alpha)}}{\sqrt{n}}(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle)\Bigg), \\
&= x_{k,2}^{(j)}\frac{1}{n}\frac{\tau}{\sqrt{n}}\Bigg(\frac{1}{\sqrt{n}}n\langle\gamma_2^{'}\rangle + \frac{Q_{2,1}n}{\sqrt{n}}(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle) \\
&\qquad + \frac{Q_{2,2}n}{\sqrt{n}}(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle)\Bigg), \\
&= x_{k,2}^{(j)}\frac{\tau}{\sqrt{n}}\Bigg(\frac{1}{\sqrt{n}}\langle\gamma_2^{'}\rangle + \frac{Q_{2,1}}{\sqrt{n}}(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle) \\
&\qquad + \frac{Q_{2,2}}{\sqrt{n}}(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle)\Bigg). \quad (65)
\end{aligned}
$$

Having derived the expression above, we finally substitute all our previous calculations (58),(62) and (65) into (63) to get

$$\begin{aligned}
\mathbb{E}_k[x_{k+1,2}^{(j)} - x_{k,2}^{(j)}] = &-\frac{x_{k,1}^{(j)}\tau^2\langle\gamma_1\gamma_2\rangle}{n} \\
&+ \frac{\tau}{\sqrt{n}}\Bigg(\frac{1}{\sqrt{n}}x_{k,2}^{(j)}\langle\gamma_2^{'}\rangle + \frac{u_1^{(j)}}{\sqrt{n}}\Big(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle\Big) \\
&\qquad + \frac{u_2^{(j)}}{\sqrt{n}}\Big(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle\Big)\Bigg) \\
&- \frac{x_{k,1}^{(j)}\tau}{n}\Bigg(Q_{k,1,1}\Big(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle\Big) + Q_{k,1,2}\Big(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle\Big)\Bigg) \\
&- \frac{x_{k,1}^{(j)}\tau}{n}\Bigg(Q_{k,2,1}\Big(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle\Big) + Q_{k,2,2}\Big(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle\Big)\Bigg) \\
&- \frac{x_{k,2}^{(j)}}{2n}\tau^2\langle\gamma_2^2\rangle \\
&- x_{k,2}^{(j)}\frac{\tau}{\sqrt{n}}\Bigg(\frac{1}{\sqrt{n}}\langle\gamma_2^{'}\rangle + \frac{Q_{2,1}}{\sqrt{n}}\Big(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle\Big) \\
&\qquad + \frac{Q_{2,2}}{\sqrt{n}}\Big(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle\Big)\Bigg) + o\left(\frac{1}{n}\right). \qquad (66)
\end{aligned}$$

The diffusion term, the second moment of the incremental update for the $j-th$ element, can be stated as follows:

$$\begin{aligned}
\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})^2] &= \mathbb{E}_k[(g_{k,2}^{(j)})^2] - 2\frac{x_{k,1}^{(j)}\Big(x_{k,2}^{(j)}\mathbb{E}_k[g_{k,2}^{(j)}g_{k,1}^{(j)}] + x_{k,1}^{(j)}\mathbb{E}_k[(g_{k,2}^{(j)})^2]\Big)}{n} \\
&\quad + x_{k,1}^{(j)2}\frac{1}{n^2}\Big(x_{k,2}^{(j)2}\mathbb{E}_k[(g_{k,2}^{(j)})^2] + x_{k,1}^{(j)2}\mathbb{E}_k[(g_{k,2}^{(j)})^2] \\
&\qquad + 2x_{k,1}^{(j)}x_{k,2}^{(j)}\mathbb{E}_k[g_{k,1}^{(j)}g_{k,2}^{(j)}]\Big), \\
&= \frac{\tau^2\langle\gamma_2^2\rangle}{n} - 2\frac{x_{k,1}^{(j)}(x_{k,2}^{(j)}\frac{\tau^2\langle\gamma_1\gamma_2\rangle}{n} + x_{k,1}^{(j)}\frac{\tau^2\langle\gamma_2^2\rangle}{n})}{n} \\
&\quad + x_{k,1}^{(j)2}\frac{1}{n^2}\Big(x_{k,2}^{(j)2}\frac{\tau^2\langle\gamma_1^2\rangle}{n} + x_{k,1}^{(j)2}\frac{\tau^2\langle\gamma_2^2\rangle}{n} + 2x_{k,1}^{(j)}x_{k,2}^{(j)}\frac{\tau^2\langle\gamma_1\gamma_2\rangle}{n}\Big), \\
&= \frac{\tau^2\langle\gamma_2^2\rangle}{n} + o\left(\frac{1}{n}\right). \qquad (67)
\end{aligned}$$

Combining the preceding results, we arrive at the final expressions for the second estimate $\boldsymbol{x}_{k,2}$:

$$\begin{aligned}\mathbb{E}_k\big[x_{k+1,2}^{(j)} - x_{k,2}^{(j)}\big] = &-\frac{x_{k,1}^{(j)}\tau^2\langle\gamma_1\gamma_2\rangle}{n} + \frac{\tau}{\sqrt{n}}\Big(\frac{1}{\sqrt{n}}x_{k,2}^{(j)}\langle\gamma_2'\rangle + \frac{u_1^{(j)}}{\sqrt{n}}\big(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2'\rangle\big)\\ &+ \frac{u_2^{(j)}}{\sqrt{n}}\big(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2'\rangle\big)\Big) - \frac{x_{k,2}^{(j)}}{2n}\tau^2\langle\gamma_2^2\rangle\\ &- \frac{x_{k,1}^{(j)}\tau}{n}\Big(Q_{k,1,1}\big(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2'\rangle\big) + Q_{k,1,2}\big(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2'\rangle\big)\Big)\\ &- \frac{x_{k,1}^{(j)}\tau}{n}\Big(Q_{k,2,1}\big(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1'\rangle\big) + Q_{k,2,2}\big(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1'\rangle\big)\Big)\\ &- x_{k,2}^{(j)}\frac{\tau}{\sqrt{n}}\Big(\frac{1}{\sqrt{n}}\langle\gamma_2^{'}\rangle + \frac{Q_{2,1}}{\sqrt{n}}(\langle c_1\gamma_2\rangle - Q_{k,2,1}\langle\gamma_2^{'}\rangle)\\ &\qquad + \frac{Q_{2,2}}{\sqrt{n}}(\langle c_2\gamma_2\rangle - Q_{k,2,2}\langle\gamma_2^{'}\rangle)\Big) + o\left(\frac{1}{n}\right), \end{aligned} \tag{68}$$

$$\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})^2] = \frac{\tau^2\langle\gamma_2^2\rangle}{n} + o\left(\frac{1}{n}\right). \tag{69}$$

B.2. Dynamics of the first estimate

For the initial estimate, where Gram-Schmidt orthogonalization is not applicable ($\tilde{\tilde{\boldsymbol{x}}}_{k,1} = \tilde{\boldsymbol{x}}_{k,1}$), we substitute the update expression directly into the normalization step. Applying a first-order Taylor expansion to the resulting normalization factor, we obtain

$$\begin{aligned}\boldsymbol{x}_{k+1,1} &= \frac{\boldsymbol{x}_{k,1} + \boldsymbol{g}_{k,1}}{\sqrt{1 + \|\boldsymbol{g}_{k,1}\|^2/n + 2\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}/n}},\\ &= \boldsymbol{x}_{k,1} - \boldsymbol{x}_{k,1}\frac{\|\boldsymbol{g}_{k,1}\|^2}{2n} - \boldsymbol{x}_{k,1}\frac{\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}}{n} + \boldsymbol{g}_{k,1}\\ &\quad - \boldsymbol{g}_{k,1}\frac{\|\boldsymbol{g}_{k,1}\|^2}{2n} - \boldsymbol{g}_{k,1}\frac{\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}}{n} + o\left(\frac{1}{n}\right). \end{aligned} \tag{70}$$

Consistent with the analysis of the second estimate, the expectation of the entry-wise update of the first estimate satisfies

$$\begin{aligned}&\mathbb{E}_k[x_{k+1,1}^{(j)} - x_{k,1}^{(j)}]\\ &\quad= \underbrace{\mathbb{E}_k[g_{k,1}^{(j)}]}_{P_1} - \underbrace{x_{k,1}^{(j)}\frac{\mathbb{E}_k[||\boldsymbol{g}_{k,1}||^2]}{2n}}_{P_2} - \underbrace{x_{k,1}^{(j)}\frac{\mathbb{E}_k[\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}]}{n}}_{P_3} - \underbrace{\mathbb{E}_k[g_{k,1}^{(j)}\frac{\boldsymbol{x}_{k,1}^\top\boldsymbol{g}_{k,1}}{n}]}_{P_4} + o\Big(\frac{1}{n}\Big), \end{aligned} \tag{71}$$

where the terms labeled P_1, P_2, P_3 and P_4 evaluated separately. Leveraging the results established in Section B.1 for the expected gradients, it follows that:

$$P_1 = \mathbb{E}_k[g_{k,1}^{(j)}] = \frac{\tau}{n}\Big(x_{k,1}^{(j)}\langle\gamma_1^{'}\rangle + u_1^{(j)}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) + u_2^{(j)}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Big), \quad (72)$$

$$P_2 = x_{k,1}^{(j)}\frac{\mathbb{E}_k[\|\boldsymbol{g}_{k,1}\|^2]}{2n} = x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n \mathbb{E}_k[(g_{k,1}^{(\alpha)})^2]}{2n} = \frac{x_{k,1}^{(j)}}{2n}\tau^2\langle\gamma_1^2\rangle, \quad (73)$$

$$\begin{aligned}
P_3 &= x_{k,1}^{(j)}\frac{\mathbb{E}_k[\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1}]}{n}, \\
&= x_{k,1}^{(j)}\frac{\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}\mathbb{E}_k[g_{k,1}^{(\alpha)}]}{n}, \\
&= \frac{x_{k,1}^{(j)}}{n}\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}\frac{\tau}{n}\left(x_{k,1}^{(\alpha)}\langle\gamma_1^{'}\rangle + u_1^{(\alpha)}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) + u_2^{(\alpha)}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\right), \\
&= \frac{x_{k,1}^{(j)}}{n}\tau\Bigg(\frac{1}{n}\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}x_{k,1}^{(\alpha)}\langle\gamma_1^{'}\rangle \\
&\qquad + \frac{\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}u_1^{(\alpha)}}{n}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) \\
&\qquad + \frac{\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}u_2^{(\alpha)}}{n}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Bigg), \\
&= \frac{x_{k,1}^{(j)}}{n}\tau\Big(\frac{1}{n}n\langle\gamma_1^{'}\rangle + \frac{Q_{k,1,1}n}{n}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) + \frac{Q_{k,1,2}n}{n}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Big), \\
&= \frac{x_{k,1}^{(j)}}{n}\tau\Big(\langle\gamma_1^{'}\rangle + Q_{k,1,1}(\langle c_1\gamma_1\rangle - Q_{k,1,1}\langle\gamma_1^{'}\rangle) + Q_{k,1,2}(\langle c_2\gamma_1\rangle - Q_{k,1,2}\langle\gamma_1^{'}\rangle)\Big), \quad (74)
\end{aligned}$$

$$\begin{aligned}
P_4 &= \mathbb{E}_k[g_{k,1}^{(j)}\frac{\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1}}{n}], \\
&= \frac{1}{n}\mathbb{E}_k[\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}g_{k,1}^{(\alpha)}g_{k,1}^{(j)}], \\
&= \frac{1}{n}\sum_{\alpha=1}^n x_{k,1}^{(\alpha)}\mathbb{E}_k[g_{k,1}^{(\alpha)}g_{k,1}^{(j)}], \\
&= \frac{1}{n}x_{k,1}^{(j)}\mathbb{E}_k[(g_{k,1}^{(j)})^2], \\
&= \frac{1}{n^2}x_{k,1}^{(j)}\tau^2\langle\gamma_1^2\rangle. \quad (75)
\end{aligned}$$

Combining these results yields the final formulation for the first estimate in the following:

$$
\begin{aligned}
&\mathbb{E}_k[x_{k+1,1}^{(j)} - x_{k,1}^{(j)}] \\
&\quad = \frac{\tau}{\sqrt{n}}\Big(\frac{1}{\sqrt{n}} x_{k,1}^{(j)} \langle \gamma_1^{'} \rangle + \frac{u_1^{(j)}}{\sqrt{n}}(\langle c_1 \gamma_1 \rangle - Q_{k,1,1}\langle \gamma_1^{'} \rangle) + \frac{u_2^{(j)}}{\sqrt{n}}(\langle c_2 \gamma_1 \rangle - Q_{k,1,2}\langle \gamma_1^{'} \rangle)\Big) \\
&\qquad - \frac{x_{k,1}^{(j)}}{n}\tau\Big(\langle \gamma_1^{'} \rangle + Q_{k,1,1}(\langle c_1 \gamma_1 \rangle - Q_{k,1,1}\langle \gamma_1^{'} \rangle) + Q_{k,1,2}(\langle c_2 \gamma_1 \rangle - Q_{k,1,2}\langle \gamma_1^{'} \rangle)\Big) \\
&\qquad - \frac{x_{k,1}^{(j)}}{2n}\tau^2 \langle \gamma_1^2 \rangle + o\left(\frac{1}{n}\right), \\
&\quad = \frac{\tau}{\sqrt{n}}\Big(\frac{u_1^{(j)}}{\sqrt{n}}(\langle c_1 \gamma_1 \rangle - Q_{k,1,1}\langle \gamma_1^{'} \rangle) + \frac{u_2^{(j)}}{\sqrt{n}}(\langle c_2 \gamma_1 \rangle - Q_{k,1,2}\langle \gamma_1^{'} \rangle)\Big) - \frac{x_{k,1}^{(j)}}{2n}\tau^2 \langle \gamma_1^2 \rangle \\
&\qquad - \frac{x_{k,1}^{(j)}}{n}\tau\Big(Q_{k,1,1}(\langle c_1 \gamma_1 \rangle - Q_{k,1,1}\langle \gamma_1^{'} \rangle) + Q_{k,1,2}(\langle c_2 \gamma_1 \rangle - Q_{k,1,2}\langle \gamma_1^{'} \rangle)\Big) + o\left(\frac{1}{n}\right).
\end{aligned} \tag{76}
$$

Proceeding analogously to the derivation in Section B.1, only the leading-order term survives, consistent with the previous analysis we arrive at

$$
\mathbb{E}_k[(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})^2] = \frac{\tau^2 \langle \gamma_1^2 \rangle}{n} + o\left(\frac{1}{n}\right). \tag{77}
$$

B.3. Derivation of the $\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})]$

Following the expressions in (23) and (70), expectation of the cross term satisfies

$$
\begin{aligned}
&\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})] \\
&\quad = \mathbb{E}_k\Bigg[\Big(\tilde{\tilde{\Delta}}_{k,2}^{(j)} - \frac{\|\tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}\|^2 x_{k,2}^{(j)}}{2n} - x_{k,2}^{(j)}\frac{\boldsymbol{x}_{k,2}^\top \tilde{\tilde{\boldsymbol{\Delta}}}_{k,2}}{n}\Big) \\
&\qquad \times \Big(g_{k,1}^{(j)} - x_{k,1}^{(j)}\frac{\|\boldsymbol{g}_{k,1}\|^2}{2n} - x_{k,1}^{(j)}\frac{\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1}}{n} - g_{k,1}\frac{\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1}}{n}\Big)\Bigg] + o\left(\frac{1}{n}\right).
\end{aligned} \tag{78}
$$

Furthermore, substituting Equation (22) into the expression above yields

$$
\begin{aligned}
&\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})] \\
&\quad = \mathbb{E}_k\Bigg[\Big(g_{k,2}^{(j)} - x_{k,1}^{(j)}\frac{\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,2} + \boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,1}}{n} - \frac{x_{k,2}^{(j)}}{n}(\boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,2})\Big) \\
&\qquad \times \Big(g_{k,1}^{(j)} - x_{k,1}^{(j)}\frac{\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1}}{n}\Big)\Bigg] + o\left(\frac{1}{n}\right),
\end{aligned} \tag{79}
$$

$$
\begin{aligned}
&\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})] \\
&\quad = \mathbb{E}_k[g_{k,1}^{(j)} g_{k,2}^{(j)}] - \frac{x_{k,1}^{(j)}}{n}\mathbb{E}[g_{k,2}^{(j)}(\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1})] - \frac{x_{k,1}^{(j)}}{n}\mathbb{E}[g_{k,1}^{(j)}(\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,2} + \boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,1})] \\
&\qquad - \frac{x_{k,2}^{(j)}}{n}\mathbb{E}[g_{k,1}^{(j)}(\boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,2})] + \frac{x_{k,1}^{(j)} x_{k,2}^{(j)}}{n^2}\mathbb{E}[(\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1})(\boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,2})] \\
&\qquad + \frac{x_{k,1}^{(j)2}}{n^2}\mathbb{E}[(\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,2} + \boldsymbol{x}_{k,2}^\top \boldsymbol{g}_{k,1})(\boldsymbol{x}_{k,1}^\top \boldsymbol{g}_{k,1})].
\end{aligned} \tag{80}
$$

We leverage the results we obtained for the expectations over the gradients in the previous sections, thus we can simplify the expression as

$$
\begin{aligned}
&\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})] \\
&\quad = \mathbb{E}_k[g_{k,1}^{(j)} g_{k,2}^{(j)}] - \frac{x_{k,1}^{(j)2}}{n}\mathbb{E}[g_{k,2}^{(j)} g_{k,1}^{(j)}] - \frac{x_{k,1}^{(j)}}{n}(x_{k,1}^{(j)}\mathbb{E}[g_{k,1}^{(j)} g_{k,2}^{(j)}] + x_{k,2}^{(j)}\mathbb{E}[g_{k,1}^{(j)2}]) \\
&\qquad - \frac{x_{k,2}^{(j)2}}{n}\mathbb{E}[g_{k,1}^{(j)} g_{k,2}^{(j)}] + \frac{x_{k,1}^{(j)} x_{k,2}^{(j)}}{n^2}\sum_{\alpha=1}^{n} x_{k,1}^{(\alpha)} x_{k,2}^{(\alpha)}\mathbb{E}[g_{k,1}^{(\alpha)} g_{k,2}^{(\alpha)}] \\
&\qquad + \frac{x_{k,1}^{(j)2}}{n^2}(\sum_{\alpha=1}^{n} x_{k,1}^{(\alpha)2}\mathbb{E}[g_{k,1}^{(\alpha)} g_{k,2}^{(\alpha)}] + \sum_{\alpha=1}^{n} x_{k,1}^{(\alpha)} x_{k,2}^{(\alpha)}\mathbb{E}[g_{k,1}^{(\alpha)2}]).
\end{aligned} \tag{81}
$$

Recall that $\mathbb{E}[g_{k,1}^{(j)2}]$,$\mathbb{E}[g_{k,2}^{(j)2}]$ and $\mathbb{E}[g_{k,1}^{(j)} g_{k,2}^{(j)}]$ were found to be independent of j, thus these expectations get out of the sum operator. We also recognize that $\sum_{\alpha=1}^{n} x_{k,1}^{(\alpha)2} = n$ and $\sum_{\alpha=1}^{n} x_{k,1}^{(\alpha)} x_{k,2}^{(\alpha)} = 0$, and arrive at the following:

$$
\begin{aligned}
&\mathbb{E}_k[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})] \\
&\quad = \frac{\tau^2\langle\gamma_1\gamma_2\rangle}{n} - \frac{x_{k,1}^{(j)2}}{n^2}\tau^2\langle\gamma_1\gamma_2\rangle - \frac{x_{k,1}^{(j)}}{n^2}(x_{k,1}^{(j)}\tau^2\langle\gamma_1\gamma_2\rangle + x_{k,2}^{(j)}\tau^2\langle\gamma_1^2\rangle) \\
&\qquad - \frac{x_{k,2}^{(j)2}}{n^2}\tau^2\langle\gamma_1\gamma_2\rangle + \frac{x_{k,1}^{(j)2}}{n^2}\tau^2\langle\gamma_1\gamma_2\rangle.
\end{aligned} \tag{82}
$$

Observe that all the other terms other than the first, is of order $O\left(\frac{1}{n^2}\right)$. Consequently, we arrive at the final expression:

$$
\mathbb{E}_k\left[(x_{k+1,2}^{(j)} - x_{k,2}^{(j)})(x_{k+1,1}^{(j)} - x_{k,1}^{(j)})\right] = \tau^2\frac{\langle\gamma_1\gamma_2\rangle}{n} + o\left(\frac{1}{n}\right). \tag{83}
$$

B.4. Regularization Term

Regarding the regularization term $(-\frac{\tau}{n}\phi(x))$, we first define the macroscopic variable $R_{k,i,j}$ as

$$
R_{k,i,j} = \frac{\boldsymbol{x}_{k,i}^\top \phi(\boldsymbol{x}_{k,j})}{n}. \tag{84}
$$

We derive the additional regularization terms by redefining the gradient expressions in (15). The resulting updates are given by

$$\tilde{\boldsymbol{x}}_{k,1} = \boldsymbol{x}_{k,1} + \underbrace{\frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}} \boldsymbol{y}_k^\top \boldsymbol{x}_{k,1}\right) \boldsymbol{y}_k - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1})}_{:=\boldsymbol{g}_{k,1} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1})}, \tag{85}$$

$$\tilde{\boldsymbol{x}}_{k,2} = \boldsymbol{x}_{k,2} + \underbrace{\frac{\tau}{\sqrt{n}} f\left(\frac{1}{\sqrt{n}} \boldsymbol{y}_k^\top \boldsymbol{x}_{k,2}\right) \boldsymbol{y}_k - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,2})}_{:=\boldsymbol{g}_{k,2} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,2})}. \tag{86}$$

Using our result (71) from Section B.2, we obtain the following expression for the first estimate:

$$\begin{aligned}
\mathbb{E}_k[x_{k+1,1}^{(j)} - x_{k,1}^{(j)}] = & \underbrace{\mathbb{E}_k[g_{k,1}^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1})^{(j)}]}_{P_1} - \underbrace{x_{k,1}^{(j)} \frac{\mathbb{E}_k[||\boldsymbol{g}_{k,1} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1})||^2]}{2n}}_{P_2} \\
& - \underbrace{x_{k,1}^{(j)} \frac{\mathbb{E}_k[\boldsymbol{x}_{k,1}^\top(\boldsymbol{g}_{k,1} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1}))]}{n}}_{P_3} \\
& - \underbrace{\mathbb{E}_k[(g_{k,1}^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1})^{(j)}) \frac{\boldsymbol{x}_{k,1}^\top(\boldsymbol{g}_{k,1} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1}))}{n}]}_{P_4} + o\left(\frac{1}{n}\right).
\end{aligned} \tag{87}$$

Terms P_1, P_2, P_3, P_4 will be calculated separately in the following. For notational clarity and to avoid the repetition of lengthy baseline expressions, we use $(\mathbb{E}_k)_{\text{base}}$ as a generic placeholder for the terms present in the unregularized dynamics. It should be understood that the specific content of $(\mathbb{E}_k)_{\text{base}}$ varies depending on the quantity being evaluated, representing the respective unregularized expectations derived in Section B.2 and B.1. This allows us to isolate and highlight the new contributions arising from the regularization term:

$$
\begin{aligned}
P_1 &= (\mathbb{E}_k)_{base} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1})^{(j)}, && (88)\\
P_2 &= (\mathbb{E}_k)_{base} - \frac{x_{k,1}^{(j)}}{2n^3}\tau^2\|\phi(\boldsymbol{x}_{k,1})\|^2 + \frac{x_{k,1}^{(j)}}{n^2}\tau\mathbb{E}[\boldsymbol{g}_{k,1}^\top\phi(\boldsymbol{x}_{k,1})],\\
&= (\mathbb{E}_k)_{base} + \frac{x_{k,1}^{(j)}}{n^2}\tau\sum_{\alpha=1}^{n}\frac{\tau}{n}\Bigg(\langle\gamma_1'\rangle x_{k,1}^{(\alpha)}\phi(\boldsymbol{x}_{k,1})^{(\alpha)}\\
&\qquad + u_1^{(\alpha)}\phi(\boldsymbol{x}_{k,1})^{(\alpha)}\Big(\langle c_1\gamma_i\rangle - Q_{k,i,1}\langle\gamma_i'\rangle\Big)\\
&\qquad + u_2^{(\alpha)}\phi(\boldsymbol{x}_{k,1})^{(\alpha)}\Big(\langle c_2\gamma_i\rangle - Q_{k,i,2}\langle\gamma_i'\rangle\Big)\Bigg) + o\left(\frac{1}{n}\right),\\
&= (\mathbb{E}_k)_{base} + \frac{x_{k,1}^{(j)}}{n^2}\tau^2\langle\gamma_1'\rangle R_{k,1,1} + o\left(\frac{1}{n}\right),\\
&= (\mathbb{E}_k)_{base} + o\left(\frac{1}{n}\right), && (89)\\
P_3 &= (\mathbb{E}_k)_{base} + \frac{x_{k,1}^{(j)}}{n^2}\tau\boldsymbol{x}_{k,1}^\top\phi(\boldsymbol{x}_{k,1}),\\
&= (\mathbb{E}_k)_{base} + \frac{\tau}{n}x_{k,1}^{(j)}R_{k,1,1}, && (90)\\
P_4 &= (\mathbb{E}_k)_{base} + o\left(\frac{1}{n}\right). && (91)
\end{aligned}
$$

Finally, we arrive at the following result for the first estimate:

$$
\mathbb{E}_k[x_{k+1,1}^{(j)} - x_{k,1}^{(j)}] = (\mathbb{E}_k)_{base} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,1})^{(j)} + \frac{\tau}{n}x_{k,1}^{(j)}R_{k,1,1} + o\left(\frac{1}{n}\right). \tag{92}
$$

Following Equation (24), we compute the regularization induced contributions for each term separately for the second estimate as follows:

$$
\begin{aligned}
\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}] = & \underbrace{\mathbb{E}_k[g_{k,2}^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,2}})^{(j)}]}_{(\mathbb{E}_k)_{base} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,2}})^{(j)}} - \underbrace{x_{k,1}^{(j)} \frac{\sum_{\alpha=1}^{n} x_{k,2}^{(\alpha)} \mathbb{E}_k[g_{k,1}^{(\alpha)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,1}})^{(\alpha)}]}{n}}_{(\mathbb{E}_k)_{base} + \frac{1}{n} x_{k,1}^{(j)} \tau R_{k,2,1} + o(1/n)} \\
& - \underbrace{x_{k,1}^{(j)} \frac{\sum_{\alpha=1}^{n} x_{k,1}^{(\alpha)} \mathbb{E}_k[g_{k,2}^{(\alpha)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,2}})^{(\alpha)}]}{n}}_{(\mathbb{E}_k)_{base} + \frac{1}{n} x_{k,1}^{(j)} \tau R_{k,1,2} + o(1/n)} \\
& + \underbrace{2x_{k,1}^{(j)} \frac{\sum_{\alpha=1}^{n} x_{k,2}^{(\alpha)} x_{k,1}^{(\alpha)} \mathbb{E}_k[(g_{k,1}^{(\alpha)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,1}})^{(\alpha)})^2]}{n^2}}_{(\mathbb{E}_k)_{base} + o(1/n)} \\
& + \underbrace{2x_{k,1}^{(j)} \frac{\sum_{\alpha=1}^{n} (x_{k,1}^{(\alpha)})^2 \mathbb{E}_k[(g_{k,1}^{(\alpha)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,1}})^{(\alpha)})(g_{k,2}^{(\alpha)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,2}})^{(\alpha)})]}{n^2}}_{(\mathbb{E}_k)_{base} + o(1/n)} \\
& - \underbrace{\frac{x_{k,2}^{(j)} \mathbb{E}_k[(g_{k,1}^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,1}})^{(j)})^2]}{n}}_{(\mathbb{E}_k)_{base} + o(1/n)} \\
& - \underbrace{\frac{x_{k,1}^{(j)} \mathbb{E}_k[(g_{k,2}^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,2}})^{(j)})(g_{k,1}^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,1}})^{(j)})]}{n}}_{(\mathbb{E}_k)_{base} + o(1/n)} \\
& - \underbrace{x_{k,1}^{(j)} \frac{\mathbb{E}_k[(\boldsymbol{g}_{k,2} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,2}}))^\top (\boldsymbol{g}_{k,1} - \frac{\tau}{n}\phi(\boldsymbol{x_{k,1}})^{(j)})]}{n}}_{(\mathbb{E}_k)_{base} + o(1/n)}, \\
= & (\mathbb{E}_k)_{base} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,2})^{(j)} + \frac{1}{n} x_{k,1}^{(j)} \tau R_{k,2,1} + \frac{1}{n} x_{k,1}^{(j)} \tau R_{k,1,2} + o\left(\frac{1}{n}\right),
\end{aligned}
\tag{93}
$$

$$
\begin{aligned}
\mathbb{E}_k[(\tilde{\tilde{\Delta}}_{k,2}^{(j)})^2] = & \mathbb{E}_k[(g_{k,2}^{(j)})^2] - 2\frac{x_{k,1}^{(j)}(x_{k,2}^{(j)} \mathbb{E}_k[g_{k,2}^{(j)} g_{k,1}^{(j)}] + x_{k,1}^{(j)} \mathbb{E}_k[(g_{k,2}^{(j)})^2])}{n} \\
& + x_{k,1}^{(j)2} \frac{1}{n^2} \Big(\mathbb{E}_k[(g_{k,1}^{(j)})^2] \underbrace{\sum_{\alpha} (x_{k,2}^{(\alpha)})^2}_{=\,n} + \mathbb{E}_k[g_{k,2}^{(j)2}] \underbrace{\sum_{\alpha} (x_{k,1}^{(\alpha)})^2}_{=\,n} \\
& + 2\mathbb{E}_k[g_{k,1}^{(j)} g_{k,2}^{(j)}] \underbrace{\sum_{\alpha} x_{k,1}^{(\alpha)} x_{k,2}^{(\alpha)}}_{=\,0} \Big).
\end{aligned}
\tag{94}
$$

As established in the evaluation of $\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(j)}]$, the second-order gradients exclusively contribute to the baseline dynamics $(\mathbb{E}_k)_{\text{base}}$ and higher-order terms. Consequently, the term $\mathbb{E}_k[(\tilde{\tilde{\Delta}}_{k,2}^{(j)})^2]$ does not introduce additional regularization components at order $O(1/n)$, leading to

$$\mathbb{E}_k[(\tilde{\tilde{\Delta}}_{k,2}^{(j)})^2] = (\mathbb{E}_k)_{base} + o\left(\frac{1}{n}\right). \tag{95}$$

Lastly, the final term is given by

$$\begin{aligned} x_{k,2}^{(j)}\Big(\sum_{\alpha=1}^{n} x_{k,2}^{(\alpha)}\mathbb{E}_k[\tilde{\tilde{\Delta}}_{k,2}^{(\alpha)}]\Big)/n &= \frac{x_{k,2}^{(j)}}{n}\sum_{\alpha=1}^{n} x_{k,2}^{(\alpha)}\Bigg((\mathbb{E}_k)_{base} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,2})^{(\alpha)} + \frac{1}{n}x_{k,1}^{(\alpha)}\tau R_{k,2,1} \\ &\qquad + \frac{1}{n}x_{k,1}^{(\alpha)}\tau R_{k,1,2} + o\left(\frac{1}{n}\right)\Bigg), \\ &= (\mathbb{E}_k)_{base} - \frac{x_{k,2}^{(j)}}{n}\sum_{\alpha=1}^{n} x_{k,2}^{(\alpha)}\frac{\tau}{n}\phi(\boldsymbol{x}_{k,2})^{(\alpha)} + o\left(\frac{1}{n}\right), \\ &= (\mathbb{E}_k)_{base} - \frac{x_{k,2}^{(j)}}{n}\tau R_{k,2,2} + o\left(\frac{1}{n}\right). \end{aligned} \tag{96}$$

Finally, for the second estimate's incremental update we arrive at

$$\begin{aligned} \mathbb{E}_k[x_{k+1,2}^{(j)} - x_{k,2}^{(j)}] = (\mathbb{E}_k)_{base} - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,2})^{(j)} + \frac{\tau}{n}x_{k,2}^{(j)}R_{k,2,2} + \frac{\tau}{n}x_{k,1}^{(j)}(R_{k,1,2} + R_{k,2,1}) \\ + o\left(\frac{1}{n}\right). \end{aligned} \tag{97}$$

B.5. Generalization to multi-component case ($p \geq 2$)

To generalize our results to p component vectors, we express the terms in vector-matrix notation, furthermore we define the following vectors first in $\mathbb{R}^p$:

$$\boldsymbol{c}_k := [c_{k,1}, \ldots, c_{k,p}]^\top, \tag{98a}$$

$$\boldsymbol{u}^{(j)} := [u_1^{(j)}, \ldots, u_p^{(j)}]^\top, \tag{98b}$$

$$\boldsymbol{x}_k^{(j)} := [x_{k,1}^{(j)}, \ldots, x_{k,p}^{(j)}]^\top, \tag{98c}$$

$$\boldsymbol{q}_i := [Q_{k,i,1}, \ldots, Q_{k,i,p}]^\top, \tag{98d}$$

$$\boldsymbol{\psi}_i := \begin{bmatrix} \langle c_{k,1}\gamma_i\rangle - Q_{k,i,1}\langle\gamma_i'\rangle \\ \vdots \\ \langle c_{k,p}\gamma_i\rangle - Q_{k,i,p}\langle\gamma_i'\rangle \end{bmatrix}, \quad \text{where} \quad \gamma_i := f\Big(\boldsymbol{c}_k^\top\boldsymbol{q}_i + e_i\sqrt{1 - \boldsymbol{q}_i^\top\boldsymbol{q}_i}\Big). \tag{98e}$$

Having derived the explicit dynamics for the initial two components, we observe a consistent pattern in the interaction terms. By proceeding inductively, we generalize the formulation to an arbitrary $l - th$ component, obtaining the following expression for the drift term:

$$\mathbb{E}_k[x_{k+1,l}^{(j)} - x_{k,l}^{(j)}] = -\frac{\tau^2}{n}\left(\frac{1}{2}x_{k,l}^{(j)}\langle\gamma_l^2\rangle + \sum_{i=1}^{l-1} x_{k,i}^{(j)}\langle\gamma_i\gamma_l\rangle\right) + \frac{\tau}{n}\,\boldsymbol{u}^\top\boldsymbol{\psi}_l$$
$$- \frac{\tau}{n}x_{k,l}^{(j)}\mathbf{q}_l^\top\boldsymbol{\psi}_l - \frac{\tau}{n}\sum_{i=1}^{l-1} x_{k,i}^{(j)}\Big(\mathbf{q}_i^\top\boldsymbol{\psi}_l + \mathbf{q}_l^\top\boldsymbol{\psi}_i\Big)$$
$$+ \frac{\tau}{n}x_{k,l}^{(j)}R_{k,l,l} + \frac{\tau}{n}\sum_{i=1}^{l-1} x_{k,i}^{(j)}(R_{k,i,l} + R_{k,l,i}) - \frac{\tau}{n}\phi(\boldsymbol{x}_{k,l})^{(j)}$$
$$+ o\left(\frac{1}{n}\right). \tag{99}$$

Furthermore, the diffusion term satisfies the following:

$$\mathbb{E}_k[(x_{l,k+1}^{(j)} - x_{l,k}^{(j)})(x_{m,k+1}^{(j)} - x_{m,k}^{(j)})] = \frac{\tau^2}{n}\langle\gamma_l\gamma_m\rangle + o\left(\frac{1}{n}\right). \tag{100}$$

For more convenient notation, we express everything in matrix form as in the main text. Define the matrices $\boldsymbol{\Psi} \coloneqq [\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_p]$, $\boldsymbol{Q} \coloneqq [\boldsymbol{q}_1, ..., \boldsymbol{q}_p]^\top$, $(M)_{i,j} = \boldsymbol{q}_i^\top\boldsymbol{\psi}_j$ and lastly $(C)_{i,j} = \langle\gamma_i\gamma_j\rangle$. Finally, we arrive at the following expectations for $\boldsymbol{\Delta}_k^{(j)} \coloneqq \boldsymbol{x}_{k+1}^{(j)} - \boldsymbol{x}_k^{(j)}$:

$$\mathbb{E}[\boldsymbol{\Delta}_k^{(j)}] = -\frac{\tau^2}{2n}\Big(\mathrm{tril}(\boldsymbol{C}_k + \boldsymbol{C}_k^\top) - \mathrm{diag}(\boldsymbol{C}_k)\Big)\boldsymbol{x}_k^{(j)}$$
$$- \frac{\tau}{n}\Big(\mathrm{tril}(\boldsymbol{M}_k + \boldsymbol{M}_k^\top) - \mathrm{diag}(\boldsymbol{M}_k)\Big)\boldsymbol{x}_k^{(j)}$$
$$+ \frac{\tau}{n}\Big(\mathrm{tril}(\boldsymbol{R}_k + \boldsymbol{R}_k^\top) - \mathrm{diag}(\boldsymbol{R}_k)\Big)\boldsymbol{x}_k^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x}_k^{(j)})$$
$$+ \frac{\tau}{n}\,\boldsymbol{\Psi}_k^\top\,\boldsymbol{u}^{(j)} + o\left(\frac{1}{n}\right), \tag{101}$$

$$\mathbb{E}[\boldsymbol{\Delta}_k^{(j)}\boldsymbol{\Delta}_k^{(j)\top}] = \frac{\tau^2}{n}\boldsymbol{C}_k + o\left(\frac{1}{n}\right). \tag{102}$$

Appendix C. Formal statement and proof of Theorem 1

Theorem 1 *Suppose that the initial empirical measure μ_0^n converges weakly to a deterministic measure μ_0 as $n \to \infty$, the function f and its first derivative are at most of polynomial growth, and ϕ is Lipschitz. We further assume that all moments of $\boldsymbol{u}_i$ and $\boldsymbol{c}_k$ are bounded, as well as the moments of the initial estimates $\boldsymbol{x}_{0,i}$. Then, as $n \to \infty$, the empirical measure process $\{\mu^n_{\lfloor tn \rfloor}\}_{t\geq 0}$ converges weakly to a deterministic limiting measure-valued process $\{\mu_t\}_{t\geq 0}$. Using the limiting process, we define*

$$\boldsymbol{Q}_t \coloneqq \int_{\mathbb{R}^{2p}} \boldsymbol{x}\,\boldsymbol{u}^\top \,\mu_t \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{u}, \qquad \boldsymbol{R}_t \coloneqq \int_{\mathbb{R}^{2p}} \boldsymbol{x}\phi(\boldsymbol{x})^\top \mu_t \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{u}. \tag{103}$$

Then, for any bounded test function φ that is three-times continuously differentiable on $\mathbb{R}^{2p}$, the limiting dynamics can be written in the weak integral form

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle + \int_0^t \left\langle \nabla_{\boldsymbol{x}} \varphi^\top \,\boldsymbol{\omega}_s, \mu_s \right\rangle \mathrm{d}s + \int_0^t \frac{1}{2} \left\langle \mathrm{Tr}\big(\boldsymbol{\Lambda}_s \,\nabla^2_{\boldsymbol{x}} \varphi\big)\,, \mu_s \right\rangle \mathrm{d}s, \tag{104}$$

where $\langle \varphi, \mu_t \rangle \coloneqq \int \varphi(\boldsymbol{x}, \boldsymbol{u})\,\mu_t\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{u}$. We define the coefficients $\boldsymbol{\omega}_t$ and $\boldsymbol{\Lambda}_t$ as

$$\boldsymbol{\omega}_t = -\frac{\tau^2}{2}\,\mathcal{T}(\boldsymbol{C}_t)\boldsymbol{x} - \tau\mathcal{T}(\boldsymbol{M_t})\boldsymbol{x} + \tau\mathcal{T}(\boldsymbol{R}_t)\boldsymbol{x} + \tau\ \boldsymbol{\Psi}_t^\top\ \boldsymbol{u} - \tau\phi(\boldsymbol{x}), \tag{105}$$

$$\boldsymbol{\Lambda}_t = \tau^2 \boldsymbol{C}_t, \tag{106}$$

where

$$\boldsymbol{\Psi}_t \coloneqq \mathbb{E}_{\boldsymbol{c}_t,\boldsymbol{e}}\Big[\boldsymbol{c}_t f(\boldsymbol{v}_t)^\top - \boldsymbol{Q}_t^\top \mathrm{diag}(f'(\boldsymbol{v}_t))\Big], \qquad \boldsymbol{M}_t \coloneqq \boldsymbol{Q}_t\boldsymbol{\Psi}_t, \qquad \boldsymbol{C}_t \coloneqq \mathbb{E}_{\boldsymbol{c}_t,\boldsymbol{e}}\Big[f(\boldsymbol{v}_t)f(\boldsymbol{v}_t)^\top\Big],$$

$$\boldsymbol{v}_t \coloneqq \boldsymbol{Q}_t\boldsymbol{c}_t + \Big(\mathrm{diag}(\boldsymbol{I} - \boldsymbol{Q}_t\boldsymbol{Q}_t^\top)\Big)^{1/2}\boldsymbol{e}, \quad \textit{with} \quad \boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$$

and $\boldsymbol{c}_t$ denotes the mean-field scaled $\boldsymbol{c}_k$ in (1)($k = \lfloor tn \rfloor$ with $t \geq 0$).

The formal proof of Theorem 1 utilizes the meta-theorem by [11]. This theorem provides sufficient conditions for a sequence of measure-valued processes $\{(\mu_t^n)_{0\leq k\leq T}\}_n$ to converge to a limiting partial differential equation (PDE) . Below, we restate these assumptions C.1-C.10, generalizing the original scalar formulations to our multi-component vector and matrix notation. We subsequently verify that each condition is satisfied within our setting.

C.1. Assumption 1

The Markov Chain $\{(\boldsymbol{X}_k, \boldsymbol{U}_k)\}_{k\geq 0}$ is exchangeable.

In our multi-component online ICA setting, the state $(\boldsymbol{X}_k, \boldsymbol{U}_k)$ forms a Markov chain, where the iterates are $\boldsymbol{X}_k = [\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{x}_{k,p}]^\top \in \mathbb{R}^{p\times n}$ and the fixed true components are $\boldsymbol{U}_k = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p] \in \mathbb{R}^{n\times p}$. We formally include the trivial update $\boldsymbol{U}_{k+1} = \boldsymbol{U}_k$ for completeness, even though the true components are fixed. We show that the complete update step; comprising the online update, orthogonalization, and normalization preserves joint exchangeability over the n spatial coordinates (the columns of $\boldsymbol{X}_k$ and the rows of $\boldsymbol{U}_k$).

The complete algorithm at each iteration k operates as a sequence of distinct, deterministic transformations: the online update, Gram-Schmidt orthogonalization, and normalization. Because each procedure takes the output of the preceding step as its direct input, the overall update can be formally expressed as a composite mapping.

Without loss of generality, we can assume that the initial state $(\boldsymbol{X}_0, \boldsymbol{U}_0)$ forms an exchangeable Markov chain (see Remark 3 of [11]). To establish exchangeability, we must demonstrate that the online update rule is equivariant to permutations of the n spatial coordinates. Let $\boldsymbol{\pi} \in \mathbb{R}^{n\times n}$ be an arbitrary permutation matrix, $\boldsymbol{\pi}\boldsymbol{\pi}^\top = \boldsymbol{I}$. Thus, permuted matrices satisfy

$$\boldsymbol{U}^\pi = [\boldsymbol{\pi}\boldsymbol{u}_1, \ldots, \boldsymbol{\pi}\boldsymbol{u}_p] = \boldsymbol{\pi}\boldsymbol{U}, \qquad \boldsymbol{X}_k^\pi = [\boldsymbol{\pi}\boldsymbol{x}_{k,1}, \ldots, \boldsymbol{\pi}\boldsymbol{x}_{k,p}]^\top = \boldsymbol{X}_k\boldsymbol{\pi}^\top. \tag{107}$$

We first demonstrate that the online update preserves exchangeability by showing equivariance to spatial permutations:

$$\begin{aligned}
\tilde{\boldsymbol{X}}_k(\boldsymbol{\pi}^\top)^\top &= \tilde{\boldsymbol{X}}_k\boldsymbol{\pi}, \\
&= \Bigg(\boldsymbol{X}_k\boldsymbol{\pi}^\top + \frac{\tau}{\sqrt{n}} f\Big(\frac{1}{\sqrt{n}}\boldsymbol{X}_k\boldsymbol{\pi}^\top(\frac{1}{\sqrt{n}}\boldsymbol{\pi}\boldsymbol{U}\boldsymbol{c}_k + \boldsymbol{\pi}\boldsymbol{a}_{\boldsymbol{k}})\Big)(\frac{1}{\sqrt{n}}\boldsymbol{\pi}\boldsymbol{U}\boldsymbol{c}_k + \boldsymbol{\pi}\boldsymbol{a}_{\boldsymbol{k}})^\top \\
&\qquad - \frac{\tau}{n}\phi(\boldsymbol{X}_k\boldsymbol{\pi}^\top)\Bigg)\boldsymbol{\pi}, \\
&= \boldsymbol{X}_k\boldsymbol{\pi}^\top\boldsymbol{\pi} + \frac{\tau}{\sqrt{n}}\, f\Big(\frac{1}{n}\boldsymbol{X}_k\boldsymbol{\pi}^\top\boldsymbol{\pi}\boldsymbol{U}\boldsymbol{c}_k + \frac{1}{\sqrt{n}}\boldsymbol{X}_k\boldsymbol{\pi}^\top\boldsymbol{\pi}\boldsymbol{a}_{\boldsymbol{k}}\Big)(\frac{1}{\sqrt{n}}\boldsymbol{c}_{\boldsymbol{k}}^\top\boldsymbol{U}^\top\boldsymbol{\pi}^\top + \boldsymbol{a}_{\boldsymbol{k}}^\top\boldsymbol{\pi}^\top)\boldsymbol{\pi} \\
&\qquad - \frac{\tau}{n}\phi(\boldsymbol{X}_k\boldsymbol{\pi}^\top)\boldsymbol{\pi}, \\
&= \boldsymbol{X}_k\boldsymbol{I} + \frac{\tau}{\sqrt{n}}\, f\Big(\frac{1}{n}\boldsymbol{X}_k\boldsymbol{I}\boldsymbol{U}\boldsymbol{c}_k + \frac{1}{\sqrt{n}}\boldsymbol{X}_k\boldsymbol{I}\boldsymbol{a}_{\boldsymbol{k}}\Big)(\frac{1}{\sqrt{n}}\boldsymbol{c}_{\boldsymbol{k}}^\top\boldsymbol{U}^\top + \boldsymbol{a}_{\boldsymbol{k}}^\top\boldsymbol{\pi}^\top\boldsymbol{\pi})\boldsymbol{\pi}^\top\boldsymbol{\pi} \\
&\qquad - \frac{\tau}{n}\phi(\boldsymbol{X}_k)\boldsymbol{\pi}^\top\boldsymbol{\pi}, \\
&= \boldsymbol{X}_k + \frac{\tau}{\sqrt{n}}\, f\Big(\frac{1}{n}\boldsymbol{X}_k\boldsymbol{U}\boldsymbol{c}_k + \frac{1}{\sqrt{n}}\boldsymbol{X}_k\boldsymbol{a}_{\boldsymbol{k}}\Big)(\frac{1}{\sqrt{n}}\boldsymbol{U}\boldsymbol{c}_{\boldsymbol{k}} + \boldsymbol{a}_{\boldsymbol{k}})^\top - \frac{\tau}{n}\phi(\boldsymbol{X}_k)\boldsymbol{I}, \\
&= \boldsymbol{X}_k + \frac{\tau}{\sqrt{n}}\, f\Big(\frac{1}{\sqrt{n}}\boldsymbol{X}_k\boldsymbol{y}_{\boldsymbol{k}}\Big)\boldsymbol{y}_{\boldsymbol{k}}^\top - \frac{\tau}{n}\phi(\boldsymbol{X}_k).
\end{aligned} \tag{108}$$

Above, we exploit the fact that the regularization $\phi(\cdot)$ acts entrywise on the matrix. Consequently, it is equivariant with respect to spatial permutations applied to the columns: $\phi(\boldsymbol{X}_k\boldsymbol{\pi}^\top) = \phi(\boldsymbol{X}_k)\boldsymbol{\pi}^\top$. Furthermore, for the Gram-Schmidt process we have

$$\begin{aligned}
\tilde{\tilde{\boldsymbol{X}}}_k &= [\tilde{\tilde{\boldsymbol{x}}}_{k,1}, \ldots, \tilde{\tilde{\boldsymbol{x}}}_{k,p}]^\top, \\
&= [\tilde{\boldsymbol{x}}_{k,1}, \ldots, \tilde{\tilde{\boldsymbol{x}}}_{k,i} = \tilde{\boldsymbol{x}}_{k,i} - \sum_{j=1}^{i-1}\frac{\tilde{\boldsymbol{x}}_{k,i}^\top\tilde{\boldsymbol{x}}_{k,j}}{\|\tilde{\boldsymbol{x}}_{k,j}\|^2}\,\tilde{\boldsymbol{x}}_{k,j}, \ldots, \tilde{\boldsymbol{x}}_{k,p} - \sum_{j=1}^{p-1}\frac{\tilde{\boldsymbol{x}}_{k,p}^\top\tilde{\boldsymbol{x}}_{k,j}}{\|\tilde{\boldsymbol{x}}_{k,j}\|^2}\,\tilde{\boldsymbol{x}}_{k,j}],
\end{aligned} \tag{109}$$

To demonstrate exchangeability, let us consider the $i-th$ row vector, yielding

$$\begin{aligned}
\implies \boldsymbol{\pi}^\top \tilde{\tilde{\boldsymbol{x}}}_{k,i} &= \boldsymbol{\pi}^\top \left(\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,i} - \sum_{j=1}^{i-1} \frac{(\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,i})^\top (\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j})}{\|\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j}\|^2} (\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j}) \right) \\
&= \boldsymbol{\pi}^\top \boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,i} - \sum_{j=1}^{i-1} \boldsymbol{\pi}^\top \frac{(\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,i})^\top (\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j})}{\|\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j}\|^2} (\boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j}) \\
&= \boldsymbol{I}\tilde{\boldsymbol{x}}_{k,i} - \sum_{j=1}^{i-1} \frac{(\tilde{\boldsymbol{x}}_{k,i}^\top \boldsymbol{\pi}^\top \boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j})}{\tilde{\boldsymbol{x}}_{k,j}^\top \boldsymbol{\pi}^\top \boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j}} \boldsymbol{\pi}^\top \boldsymbol{\pi}\tilde{\boldsymbol{x}}_{k,j} \\
&= \tilde{\boldsymbol{x}}_{k,i} - \sum_{j=1}^{i-1} \frac{(\tilde{\boldsymbol{x}}_{k,i}^\top \tilde{\boldsymbol{x}}_{k,j})}{\tilde{\boldsymbol{x}}_{k,j}^\top \tilde{\boldsymbol{x}}_{k,j}} \tilde{\boldsymbol{x}}_{k,j}. \qquad (110)
\end{aligned}$$

Thus, the Gram-Schmidt process used to obtain $\tilde{\tilde{\boldsymbol{X}}}_k$ from $\tilde{\boldsymbol{X}}_k$ also preserves exchangeability. Finally, to show that the normalization step is exchangeable we write

$$\begin{aligned}
\boldsymbol{\pi}^\top \boldsymbol{x}_{k+1,i} &= \boldsymbol{\pi}^\top \Big(\frac{\boldsymbol{\pi}\tilde{\tilde{\boldsymbol{x}}}_{k,i}}{\|\boldsymbol{\pi}\tilde{\tilde{\boldsymbol{x}}}_{k,i}\|} \sqrt{n} \Big), \\
&= \frac{\boldsymbol{\pi}^\top \boldsymbol{\pi}\tilde{\tilde{\boldsymbol{x}}}_{k,i}}{\sqrt{\tilde{\tilde{\boldsymbol{x}}}_{k,i}^\top \boldsymbol{\pi}^\top \boldsymbol{\pi}\tilde{\tilde{\boldsymbol{x}}}_{k,i}}} \sqrt{n}, \\
&= \frac{\tilde{\tilde{\boldsymbol{x}}}_{k,i}}{\|\tilde{\tilde{\boldsymbol{x}}}_{k,i}\|} \sqrt{n}. \qquad (111)
\end{aligned}$$

Because the full update is a composition of these sequentially applied steps, demonstrating that the online update, orthogonalization and normalization are individually exchangeable is sufficient to conclude that the overall process forms an exchangeable Markov chain.

C.2. Assumption 2

The initial empirical measure $\mu_0^n(\boldsymbol{x}, \boldsymbol{u})$ converges weakly to a deterministic measure μ_0.

Our Theorem 1 is explicitly conditioned on this assumption.

C.3. Assumption 3

There is some finite constant C such that

$$\sup_n \Big\langle \mu_0^n, \sum_{i=1}^p x_i^4 + \sum_{j=1}^p u_j^4 \Big\rangle \leq C.$$

Our theorem assumes that all moments of the *initial* ground truth and estimate vectors are bounded; consequently, Assumption 3 is satisfied directly.

C.4. Assumption 4

Let $\boldsymbol{\Delta}_k^{(j)} = \boldsymbol{x}_{k+1}^{(j)} - \boldsymbol{x}_k^{(j)}$. *There exists a deterministic function* $\boldsymbol{\mathcal{G}} : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^{r_1 \times r_2} \to \mathbb{R}^p$ *for some* $r_1, r_2 \geq 0$, *such that for each* $T > 0$:

$$\max_{k \leq nT} \mathbb{E}\left[\left\|\mathbb{E}_k[\boldsymbol{\Delta}_k^{(j)}] - \frac{1}{n}\boldsymbol{\mathcal{G}}_k^{(j)}\right\|\right] \leq \frac{C(T)}{n^{1+\gamma}} \ , \tag{112}$$

where $\gamma > 0$ *is some positive constant. In the above expression,*

$$\boldsymbol{\mathcal{G}}_k^{(j)} = \boldsymbol{\mathcal{G}}(\boldsymbol{x}_k^{(j)}, \boldsymbol{u}^{(j)}, \boldsymbol{\Theta}_k^n), \tag{113}$$

and $\boldsymbol{\Theta}_k^n$ *is a matrix in* $\mathbb{R}^{r_1 \times r_2}$. *The* l, m*-th entry of* $\boldsymbol{\Theta}_k^n$ *is defined as*

$$(\boldsymbol{\Theta}_k^n)_{l,m} = \langle \mu_k^n, \varphi_{l,m}(\boldsymbol{x}, \boldsymbol{u}) \rangle, \tag{114}$$

where $\varphi_{l,m}(\boldsymbol{x}, \boldsymbol{u})$ *is some deterministic function.*

In our setting, we use the overlap matrix $\boldsymbol{Q}_t$ and matrix $\boldsymbol{R}_t$ as our macroscopic variables. We define $\varphi(\boldsymbol{x}, \boldsymbol{u}) = x^{(l)} u^{(m)}$ to obtain $(Q_k^n)_{l,m} = \langle \mu_k^n, x^{(l)} u^{(m)} \rangle$ and $\varphi(\boldsymbol{x}, \boldsymbol{u}) = x^{(l)} \phi(x)^{(m)}$ to obtain $(R_k^n)_{l,m} = \langle \mu_k^n, x^{(l)} \phi(x)^{(m)} \rangle$. Thus, we define $\boldsymbol{\Theta}$ as follows:

$$\boldsymbol{\Theta}_k^n \coloneqq \left[\boldsymbol{Q}_k^n | \boldsymbol{R}_k^n\right]. \tag{115}$$

We refer to the derivation of the expected incremental update (drift) in Appendix B, restating Equation (101) for convenience:

$$\begin{aligned}
\mathbb{E}_k[\boldsymbol{\Delta}_k^{(j)}] = &-\frac{\tau^2}{2n}\Big(\mathrm{tril}(\boldsymbol{C}_k + \boldsymbol{C}_k^\top) - \mathrm{diag}(\boldsymbol{C}_k)\Big)\boldsymbol{x}_k^{(j)} \\
&- \frac{\tau}{n}\Big(\mathrm{tril}(\boldsymbol{M}_k + \boldsymbol{M}_k^\top) - \mathrm{diag}(\boldsymbol{M}_k)\Big)\boldsymbol{x}_k^{(j)} \\
&+ \frac{\tau}{n}\Big(\mathrm{tril}(\boldsymbol{R}_k + \boldsymbol{R}_k^\top) - \mathrm{diag}(\boldsymbol{R}_k)\Big)\boldsymbol{x}_k^{(j)} - \frac{\tau}{n}\phi(\boldsymbol{x}_k^{(j)}) \\
&+ \frac{\tau}{n}\,\boldsymbol{\Psi}_k^\top\,\boldsymbol{u}^{(j)} + o\left(\frac{1}{n}\right).
\end{aligned} \tag{116}$$

We define the coefficient $\boldsymbol{\mathcal{G}}$ as

$$\begin{aligned}
\boldsymbol{\mathcal{G}}^{(j)} \coloneqq &-\frac{\tau^2}{2}\Big(\mathrm{tril}(\boldsymbol{C}_k + \boldsymbol{C}_k^\top) - \mathrm{diag}(\boldsymbol{C}_k)\Big)\boldsymbol{x}_k^{(j)} \\
&- \tau\Big(\mathrm{tril}(\boldsymbol{M}_k + \boldsymbol{M}_k^\top) - \mathrm{diag}(\boldsymbol{M}_k)\Big)\boldsymbol{x}_k^{(j)} \\
&+ \tau\Big(\mathrm{tril}(\boldsymbol{R}_k + \boldsymbol{R}_k^\top) - \mathrm{diag}(\boldsymbol{R}_k)\Big)\boldsymbol{x}_k^{(j)} - \tau\phi(\boldsymbol{x}_k^{(j)}) \\
&+ \tau\,\boldsymbol{\Psi}_k^\top\,\boldsymbol{u}^{(j)}.
\end{aligned} \tag{117}$$

Consequently, Assumption 4 is satisfied.

C.5. Assumption 5

There exists some deterministic function $\mathbf{\Lambda} : \mathbb{R}^{r_1 \times r_2} \to \mathbb{R}^{p \times p}$ *such that, for each* $T > 0$*,*

$$\max_{k \le nT} \mathbb{E}\left[\left\| \mathbb{E}_k\left[\mathbf{\Delta}_k^{(j)} \mathbf{\Delta}_k^{(j)\top}\right] - \frac{1}{n}\mathbf{\Lambda}_k \right\|_F\right] \le \frac{C(T)}{n^{1+\gamma}}, \tag{118}$$

where $\gamma > 0$ *is some positive constant, and*

$$\mathbf{\Lambda}_k = \mathbf{\Lambda}(\mathbf{\Theta}_k^n). \tag{119}$$

We refer to the derivation of the second moment of the incremental update (diffusion) in Appendix B, restating Equation (102) for convenience:

$$\mathbb{E}_k\left[\mathbf{\Delta}_k^{(j)} \mathbf{\Delta}_k^{(j)\top}\right] = \frac{1}{n}\tau^2(\mathbf{C}_k) + o\Big(\frac{1}{n}\Big). \tag{120}$$

Then we define the coefficient as follows:

$$\mathbf{\Lambda}(\boldsymbol{Q}_k^n) \coloneqq \tau^2 \mathbf{C}_k(\boldsymbol{Q}_k^n). \tag{121}$$

The diffusion term is independent of the matrix $\boldsymbol{R}$; therefore, we define it solely in terms of the $\boldsymbol{Q}$ matrix. Consequently, Assumption 5 is satisfied.

C.6. Assumption 6

For any $T > 0$*, there exists a finite constant* $B(T)$ *such that*

$$\lim_{n \to \infty} \mathbb{P}\left(\max_{k \le nT} \|\mathbf{\Theta}_k^n\|_{\max} > B(T)\right) = 0, \tag{122}$$

where $\|\cdot\|_{max}$ *is the max-norm of the matrix.*

$\boldsymbol{Q}_k^n$'s every entry $(Q_k^n)_{j,l}$ is bounded by definition: $(Q_k^n)_{j,l} \in [-1, 1]$; $\boldsymbol{R}_k^n$'s every entry is bounded by the assumption on $\phi(x)$ being Lipschitz: $|(R_k^n)_{i,j}| \le \frac{1}{\sqrt{n}}(L\|\boldsymbol{x}_{k,j}\| + \sqrt{n}|\phi(0)|) = L + |\phi(0)|$. Consequently, Assumption 6 holds.

C.7. Assumption 7

Define $(\Theta_k^n)_{\ell,m}(h) = \langle \mu_k^n, \varphi_{\ell,m}(\boldsymbol{x}, \boldsymbol{u}) \sqcap h \rangle$*, where* $\sqcap$ *is the projection operation defined as the projection of* x *onto the interval* $[-b, b]$ *:* $x \sqcap b = \min\{|x|, b\}\operatorname{sgn}(x)$*. For any* $b > B(T)$ *and* $T > 0$*, we have*

$$\limsup_{h \to \infty} \sup_n \max_{k \le nT} \mathbf{E}\left\| \boldsymbol{\mathcal{G}}(\boldsymbol{x}_k^{(j)}, \boldsymbol{u}^{(j)}, \mathbf{\Theta}_k^n) - \boldsymbol{\mathcal{G}}(\boldsymbol{x}_k^{(j)}, \boldsymbol{u}^{(j)}, \mathbf{\Theta}_k^n(h) \sqcap b) \right\| = 0, \tag{123}$$

and

$$\limsup_{h \to \infty} \sup_n \max_{k \le nT} \mathbf{E}\left\| \mathbf{\Lambda}(\boldsymbol{Q}_k^n) - \mathbf{\Lambda}(\boldsymbol{Q}_k^n(h) \sqcap b) \right\|_F = 0. \tag{124}$$

Assumption 7 holds by the same reasoning applied to Assumption C.6.

C.8. Assumption 8

For each $T > 0$, there exists $C(T) < \infty$ such that

$$\max_{k \le nT} \mathbb{E}[\|\boldsymbol{\mathcal{G}}_k^{(j)}\|^2] \le C(T) \quad \text{and} \quad \max_{k \le nT} \mathbb{E}[\|\boldsymbol{\Lambda}_k\|_F^2] \le C(T). \tag{125}$$

We first establish several supplementary lemmas needed to prove these bounds.

Lemma 2 *For $q > 0$, the vector $\boldsymbol{\upsilon}_t$ satisfies the following bound on its q-th absolute moment, where H_q and L_q are positive constants depending on q:*

$$\mathbb{E}\left[|\upsilon_i|^q\right] \le H_q(1 + L_q). \tag{126}$$

Proof We begin by recalling the growth assumption on the nonlinearity: $|f(x)| \le B(1 + |x|^q)$ for constants $B \ge 0$ and $q \ge 0$. We first establish that υ_i is zero-mean. By the the zero-mean properties of the components c_j and e_i, we have

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[\upsilon_i] = \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[c_1 Q_{i,1} + ... + c_p Q_{i,p} + e_i\sqrt{1 - Q_{i,1}^2 ... - Q_{i,p}^2}] = 0. \tag{127}$$

Next, we compute the second moment. Conditioning on c_i and applying the tower property yields

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[\upsilon_i^2] &= \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[(c_1 Q_{i,1} + ... + c_p Q_{i,p} + e_i\sqrt{1 - Q_{i,1}^2 ... - Q_{i,p}^2})^2], \\
&= \mathbb{E}_{\boldsymbol{c}}[\mathbb{E}_{\boldsymbol{e}}[(c_1 Q_{i,1} + ... + c_p Q_{i,p} + e_i\sqrt{1 - Q_{i,1}^2 ... - Q_{i,p}^2})^2 | c_1, c_2 ..., c_p]], \\
&= \mathbb{E}_c[(c_1 Q_{i,1} + ... + c_p Q_{i,p})^2 + 1 - Q_{i,1}^2 ... - Q_{i,p}^2], \\
&= 1.
\end{aligned} \tag{128}$$

To bound higher-order moments, we appeal to a Rosenthal-type inequality [50] in the form given by [51], which states the following: for a sequence of independent random variables $X_1, \ldots, X_n$ with zero mean satisfying $\mathbb{E}[|X_j|^q] < \infty$ for some $q > 2$, there exists a positive constant H_q depending only on q such that

$$\mathbb{E}\left[\left|\sum_{j=1}^n X_j\right|^q\right] \le H_q \max\left\{\sum_{j=1}^n \mathbb{E}[|X_j|^q], \left(\sum_{j=1}^n \mathbb{E}[X^2]\right)^{q/2}\right\}. \tag{129}$$

We apply this inequality by decomposing υ_i into a sum of $n = p+1$ independent, zero-mean terms: $X_1 = c_1 Q_{i,1}, \ \ldots, \ X_p = c_p Q_{i,p}$, and $X_{p+1} = e_i\sqrt{1 - Q_{i,1}^2 - \cdots}$. Each summand satisfies $\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[X_j] = 0$, and finiteness of all q-th moments follows from the assumption that the moments of c_i are bounded and e_i is Gaussian. We observe that

$$\sum_{j=1}^{p+1} X_j = \upsilon_i, \qquad \left(\sum_{j=1}^{p+1} \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[X^2]\right)^{q/2} = 1. \tag{130}$$

Thus, we arrive at the following:

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\left[|\upsilon|_i^q\right] \leq H_q \max\left\{\sum_{j=1}^{p+1} \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|X_j|^q], 1\right\}. \tag{131}$$

It remains to control the sum of individual q-th moments. Expanding, we obtain

$$\begin{aligned}
\sum_{j=1}^{p+1} \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|X_j|^q] &= \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|X_1|^q] + ... + \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|X_{p+1}|^q], \\
&= \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|c_1 Q_{i,1}|^q] + ... + \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[\left|e_i\sqrt{1 - Q_{i,1}^2 \cdots - Q_{i,p}^2}\right|^q], \\
&\leq \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|c_1|^q |Q_{i,q}|^q] + ... + \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|e_i|^q \left|\sqrt{1 - Q_{i,1}^2 \cdots - Q_{i,p}^2}\right|^q], \\
&= |Q_{i,1}|^q \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|c_1|^q] + ... + \left|\sqrt{1 - Q_{i,1}^2 \cdots - Q_{i,p}^2}\right|^q \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|e_i|^q].
\end{aligned} \tag{132}$$

Since the overlaps satisfy $|Q_{i,j}| \leq 1$ and $\left|\sqrt{1 - Q_{i,1}^2 \cdots - Q_{i,p}^2}\right| = \left|\sqrt{1 - \sum_{\alpha=1}^{p} Q_{i,\alpha}^2}\right| \leq 1$, each factor involving Q can be bounded by unity, giving

$$\begin{aligned}
\sum_{j=1}^{p+1} \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|X_j|^q] &= |Q_{i,1}|^q \underbrace{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|c_1|^q]}_{=Z_{1,q}} + ... + \left|\sqrt{1 - \sum_{\alpha=1}^{p} Q_{i,\alpha}^2}\right|^q \underbrace{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|e|^q]}_{=Z_{e,q}}, \\
&\leq Z_{1,q} + Z_{2,q} + ... + Z_{e,q}.
\end{aligned} \tag{133}$$

Substituting this back into the Rosenthal bound yields

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\left[|\upsilon|_i^q\right] \leq H_q \max\left\{Z_{1,q} + Z_{2,q} + \cdots + Z_{e,q}, 1\right\}. \tag{134}$$

To obtain a cleaner expression, we define $L_q \coloneqq Z_{1,q} + Z_{2,q} + \cdots + Z_{e,q}$ and we relax the maximum using the inequality $\max\{a, 1\} \leq 1 + a$ for $a \geq 0$:

$$H_q \max\left\{L_q, 1\right\} \leq H_q(1 + L_q), \tag{135}$$

which establishes the claimed bound and finalizes the proof:

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\left[|\upsilon_i|^q\right] = \mathbb{E}\left[|\upsilon_i|^q\right] \leq H_q(1 + L_q). \tag{136}$$

■

Lemma 3 *For $q > 0$, the entries of the matrix $\boldsymbol{C}_t$ satisfy the following elementwise bound, where B is the growth constant of the nonlinearity f and R_q and R_{2q} are positive constants depending on q:*

$$|C_{i,j}| \leq B^2(1 + 2R_q + R_{2q}). \tag{137}$$

Proof We bound the absolute value of each entry by applying Jensen's inequality followed by the growth condition $|f(x)| \leq B(1 + |x|^q)$:

$$\begin{aligned}
|C_{i,j}| &= |\mathbb{E}_{c,e}[f(\upsilon_i)f(\upsilon_j)]|, \\
&\leq \mathbb{E}_{c,e}[|f(\upsilon_i)f(\upsilon_j)|], \\
&\leq \mathbb{E}_{c,e}[|f(\upsilon_i)||f(\upsilon_j)|], \\
&\leq \mathbb{E}_{c,e}[B^2(1 + |\upsilon_i|^q)(1 + |\upsilon_j|^q)], \\
&= B^2(1 + \mathbb{E}_{c,e}|\upsilon_i|^q + \mathbb{E}_{c,e}|\upsilon_j|^q + \mathbb{E}[|\upsilon_i|^q|\upsilon_j|^q]).
\end{aligned} \tag{138}$$

Expanding the product and taking expectations termwise, we invoke Lemma 2 to substitute $\mathbb{E}_{c,e}[|\upsilon_i|^q] \leq R_q$, where we define $R_q \coloneqq H_q(1 + L_q)$. This yields

$$|C_{i,j}| \leq B^2(1 + 2R_q + \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|\upsilon_i|^q|\upsilon_j|^q]). \tag{139}$$

To handle the remaining cross-moment, we apply the Cauchy-Schwarz inequality:

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\left[|\upsilon_i|^q|\upsilon_j|^q\right] \leq \left(\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}|\upsilon_i|^{2q}\right)^{1/2}\left(\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}|\upsilon_j|^{2q}\right)^{1/2}. \tag{140}$$

Since Lemma 2 applied at order $2q$ gives $\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|\upsilon_i|^{2q}] \leq R_{2q}$ for all i, both factors are bounded by $R_{2q}^{1/2}$. Consequently, we arrive at

$$\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\left[|\upsilon_i|^q|\upsilon_j|^q\right] \leq R_{2q}. \tag{141}$$

Substituting back completes the proof:

$$|C_{i,j}| \leq B^2(1 + 2R_q + R_{2q}). \tag{142}$$

■

Lemma 4 *For $q > 0$, the entries of the matrix $\boldsymbol{\Psi}_t$ satisfy the following elementwise bound:*

$$|\Psi_{i,j}| \leq B\sqrt{2(1 + R_{2q})} + K(1 + R_q^{\frac{q-1}{q}}), \tag{143}$$

where B and K are the growth constants of the nonlinearity and its derivative, R_q and R_{2q} are positive constants depending on q.

Proof In addition to the growth condition on f,we recall the assumed bound on its derivative: $|f'(x)| \leq K(1+|x|^{q-1})$ for some $q \geq 0$ and $K \geq 0$. We begin by applying the triangle inequality:

$$\begin{aligned} |\Psi_{i,j}| &= \left|\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[c_i f(\upsilon_j)] - Q_{j,i}\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[f'(\upsilon_j)]\right|, \\ &\leq |\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[c_i f(\upsilon_j)]| + \left|Q_{j,i}\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[f'(\upsilon_j)]\right|. \end{aligned} \tag{144}$$

For the first term, we apply the Cauchy–Schwarz together with $\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[c_i^2] = 1$. For the second term, we use $|Q_{j,i}| \leq 1$ and Jensen's inequality:

$$\begin{aligned} |\Psi_{i,j}| &\leq \sqrt{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[c_i^2]\,\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[f(\upsilon_j)^2]} + |Q_{j,i}|\left|\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[f'(\upsilon_j)]\right|, \\ &\leq \sqrt{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[f(\upsilon_j)^2]} + \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\Big[|f'(\upsilon_j)|\Big]. \end{aligned} \tag{145}$$

We now substitute the growth bounds $|f(x)| \leq B(1+|x|^q)$ and $|f'(x)| \leq K(1+|x|^{q-1})$, and use $(1+a)^2 \leq 2(1+a^2)$ for $a \geq 0$ to simplify

$$\begin{aligned} |\Psi_{i,j}| &\leq \sqrt{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[B^2(1+|\upsilon_j|^q)^2]} + \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\Big[\Big|K(1+|\upsilon_j|^{q-1})\Big|\Big], \\ &= \sqrt{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[B^2(1+|\upsilon_j|^q)^2]} + \mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}\Big[K(1+|\upsilon_j|^{q-1})\Big], \\ &\leq \sqrt{2B^2\big(1+\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|\upsilon_j|^{2q}]\big)} + K\big(1+\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[|\upsilon_j|^{q-1}]\big). \end{aligned} \tag{146}$$

For the second term, we apply Jensen's in the form $\mathbb{E}[|X|^{q-1}] \leq (\mathbb{E}[|X|^q])^{(q-1)/q}$. Substituting the moment bounds from Lemma 2 then yields the claimed result:

$$|\Psi_{i,j}| \leq B\sqrt{2(1+R_{2q})} + K\big(1+R_q^{\frac{q-1}{q}}\big). \tag{147}$$

■

Lemma 5 *For $q > 0$, the entries of the matrix $\boldsymbol{M}_t$ satisfy the following elementwise bound:*

$$|M_{i,j}| \leq p\left(B\sqrt{2(1+R_{2q})} + K(1+R_q^{\frac{q-1}{q}})\right), \tag{148}$$

where B and K are the growth constants of the nonlinearity and its derivative, R_q and R_{2q} are positive constants depending on q and p is the number of components (the dimension of the matrix $\boldsymbol{M}_t \in \mathbb{R}^{p\times p}$)

Proof Expanding the matrix product and applying the triangle inequality, we bound each entry as a sum of p terms. Using $|Q_{i,l}| \leq 1$ and substituting the elementwise bound on $\boldsymbol{\Psi}$ from Lemma 4, we obtain

$$
\begin{aligned}
|M_{i,j}| &\leq \sum_{l=1}^{p} |Q_{i,l}||\Psi_{l,j}|, \\
&\leq \sum_{\alpha=1}^{p} |\Psi_{l,j}|, \\
&\leq \sum_{\alpha=1}^{p} \left(B\sqrt{2(1+R_{2q})} + K(1+R_q^{\frac{q-1}{q}}) \right), \\
&= p \left(B\sqrt{2(1+R_{2q})} + K(1+R_q^{\frac{q-1}{q}}) \right),
\end{aligned} \tag{149}
$$

where the final equality follows since the summand is independent of the index l, which concludes the proof.

■

Lemma 6 *For each coordinate j, the following bound holds for $k \leq nT$:*

$$\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] \leq A_1(T), \tag{150}$$

$$\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^4] \leq A_2(T), \tag{151}$$

where $A_1(T)$ and $A_2(T)$ are finite constants depending on T.

Proof Recall that $\boldsymbol{\Delta}_k^{(j)} \coloneqq \boldsymbol{x}_{k+1}^{(j)} - \boldsymbol{x}_k^{(j)}$. Expanding the squared norm of the next iterate yields

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{x}_{k+1}^{(j)\top}\boldsymbol{x}_{k+1}^{(j)}] &= \mathbb{E}[(\boldsymbol{x}_k^{(j)} + \boldsymbol{\Delta}_k^{(j)})^\top(\boldsymbol{x}_k^{(j)} + \boldsymbol{\Delta}_k^{(j)})], \\
&= \mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] + \mathbb{E}[\mathbb{E}_k[\|\boldsymbol{\Delta}_k^{(j)}\|^2]] + 2\mathbb{E}[\boldsymbol{x}_k^{(j)\top}\mathbb{E}_k[\boldsymbol{\Delta}_k^{(j)}]], \\
&= \mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] + \mathbb{E}[\mathrm{Tr}\left(\mathbb{E}_k[\boldsymbol{\Delta}_{\boldsymbol{k}}^{(j)}\boldsymbol{\Delta}_{\boldsymbol{k}}^{(j)\top}]\right)] + 2\mathbb{E}[\boldsymbol{x}_k^{(j)\top}\frac{\boldsymbol{\mathcal{G}}^{(j)}}{n}] + o\left(\frac{1}{n}\right), \\
&= \mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] + \mathbb{E}[\mathrm{Tr}\left(\frac{\boldsymbol{\Lambda}_k^{(j)}}{n}\right)] + 2\mathbb{E}[\boldsymbol{x}_k^{(j)\top}\frac{\boldsymbol{\mathcal{G}}^{(j)}}{n}] + o\left(\frac{1}{n}\right).
\end{aligned} \tag{152}
$$

We now bound each of the terms on the right-hand side separately. For bounding the cross term we expand it by substituting the expression for $\boldsymbol{\mathcal{G}}^{(j)}$. We let $\tilde{\boldsymbol{C}} \coloneqq \mathrm{tril}(\boldsymbol{C} + \boldsymbol{C}^\top) -$

$\mathrm{diag}(\boldsymbol{C})$, and similarly for $\tilde{\boldsymbol{M}}$ and $\tilde{\boldsymbol{R}}$ for notational simplicity, then using (117) we arrive at

$$\begin{aligned}\left|\mathbb{E}\Big[\boldsymbol{x}_k^{(j)\top}\frac{\boldsymbol{\mathcal{G}}^{(j)}}{n}\Big]\right| &\le \frac{1}{n}\mathbb{E}\Bigg[\Bigg|\boldsymbol{x}_k^{(j)\top}\Big(-\frac{\tau^2}{2}\tilde{\boldsymbol{C}}\,\boldsymbol{x}_k^{(j)} - \tau\,\tilde{\boldsymbol{M}}\,\boldsymbol{x}_k^{(j)} \\ &\qquad + \tau\tilde{\boldsymbol{R}}\boldsymbol{x}_k^{(j)} + \tau\,\boldsymbol{\Psi}_t\,\boldsymbol{u}^{(j)} - \tau\phi(\boldsymbol{x}_k^{(j)})\Big)\Bigg|\Bigg], \\ &\le \frac{1}{n}\Bigg(\frac{\tau^2}{2}\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{C}}\boldsymbol{x}_k^{(j)}\|] + \tau\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{M}}\boldsymbol{x}_k^{(j)}\|] \\ &\qquad + \tau\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{R}}\boldsymbol{x}_k^{(j)}\| \\ &\qquad + \tau\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\boldsymbol{\Psi}\boldsymbol{u}^{(j)}\|] + \tau\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\phi(\boldsymbol{x}_k)^{(j)}\|\Bigg). \end{aligned} \tag{153}$$

Since ϕ is Lipschitz, we can establish the element-wise bound on $\boldsymbol{R}$ matrix as $|R_{i,j}| = |\frac{\boldsymbol{x}_{k,i}^\top\phi(\boldsymbol{x}_{k,j})}{n}| \le \frac{1}{n}\|\boldsymbol{x}_{k,i}\|\|\phi(\boldsymbol{x}_{k,j})\| = \frac{1}{\sqrt{n}}\|\phi(\boldsymbol{x}_{k,j})\|$ where $|\phi(x)| \le L|x| + |\phi(0)|$. Thus $|R_{i,j}| \le \frac{1}{\sqrt{n}}(L\|\boldsymbol{x}_{k,j}\| + \sqrt{n}|\phi(0)|) = L + |\phi(0)|$. Here, we emphasize the distinction between $\boldsymbol{x}_{k,i} \in \mathbb{R}^n$ and $\boldsymbol{x}_k^{(j)} \in \mathbb{R}^p$. The former represents the i-th estimate vector, which has a norm of $\sqrt{n}$, while the latter is a p-dimensional vector constructed from the j-th elements of all estimate vectors. Furthermore, we can obtain the following bounds:

$$\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{C}}\boldsymbol{x}_k^{(j)}\|] \le \mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{C}}\|_2\|\boldsymbol{x}_k^{(j)}\|] = \mathbb{E}[\|\tilde{\boldsymbol{C}}\|_2\|\boldsymbol{x}_k^{(j)}\|^2], \tag{154}$$

$$\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{M}}\boldsymbol{x}_k^{(j)}\|] \le \mathbb{E}[\|\tilde{\boldsymbol{M}}\|_2\|\boldsymbol{x}_k^{(j)}\|^2], \tag{155}$$

$$\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{R}}\boldsymbol{x}_k^{(j)}\|] \le \mathbb{E}[\|\tilde{\boldsymbol{R}}\|_2\|\boldsymbol{x}_k^{(j)}\|^2], \tag{156}$$

$$\mathbb{E}[\|\boldsymbol{x}_k^{(j)\top}\|\|\boldsymbol{\Psi}_t\,\boldsymbol{u}^{(j)}\|] \le \mathbb{E}\big[\|\boldsymbol{\Psi}_t\|_2\|\boldsymbol{x}_k^{(j)}\|\,\|\boldsymbol{u}^{(j)}\|\big], \tag{157}$$

For a $p \times p$ matrix with element-wise bound $|A_{ij}| \le \alpha$, the spectral norm satisfies $\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_F \le p\,\alpha$. First we note that since $\tilde{\boldsymbol{C}} = \mathrm{tril}(\boldsymbol{C} + \boldsymbol{C}^\top) - \mathrm{diag}(\boldsymbol{C})$, its entries take the form

$$\tilde{C}_{ij} = \begin{cases} C_{ii}, & i = j, \\ C_{ij} + C_{ji}, & i > j, \\ 0, & i < j. \end{cases} \tag{158}$$

Applying the triangle inequality together with the element-wise bound on $|C_{ij}|$ from Lemma 3, and analogously for $\tilde{\boldsymbol{M}}$ and $\tilde{\boldsymbol{R}}$ yields

$$|\tilde{C}_{ij}| \leq 2B^2(1 + 2R_q + R_{2q}), \tag{159}$$

$$|\tilde{M}_{ij}| \leq 2p\left(B\sqrt{2(1+R_{2q})} + K(1 + R_q^{\frac{q-1}{q}})\right), \tag{160}$$

$$|\tilde{R}_{ij}| \leq 2(L + |\phi(0)|). \tag{161}$$

Finally, substituting these results into the elementwise bounds leads to the following:

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{C}}\boldsymbol{x}_k^{(j)}\|] &\leq \mathbb{E}[\|\tilde{\boldsymbol{C}}\|_2\|\boldsymbol{x}_k^{(j)}\|^2], \\
&\leq \mathbb{E}[2pB^2(1 + 2R_q + R_{2q})\|\boldsymbol{x}_k^{(j)}\|^2], \\
&= 2pB^2(1 + 2R_q + R_{2q})\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2], \\
\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{M}}\boldsymbol{x}_k^{(j)}\|] &\leq 2p^2\left(B\sqrt{2(1+R_{2q})} + K(1 + R_q^{\frac{q-1}{q}})\right)\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2], \\
\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|\|\tilde{\boldsymbol{R}}\boldsymbol{x}_k^{(j)}\|] &\leq 2p(L + |\phi(0)|)\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2], \\
\mathbb{E}\big[\|\boldsymbol{\Psi}_t\|_2\|\boldsymbol{x}_k^{(j)}\|\,\|\boldsymbol{u}^{(j)}\|\big] &\leq p\left(B\sqrt{2(1+R_{2q})} + K\big(1 + R_q^{\frac{q-1}{q}}\big)\right)\mathbb{E}\big[\|\boldsymbol{x}_k^{(j)}\|\,\|\boldsymbol{u}^{(j)}\|\big], \\
&\leq p\left(B\sqrt{2(1+R_{2q})} + K\big(1 + R_q^{\frac{q-1}{q}}\big)\right)\sqrt{\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2]\,\mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2]}, \\
&\leq p\left(B\sqrt{2(1+R_{2q})} + K\big(1 + R_q^{\frac{q-1}{q}}\big)\right)\frac{\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] + \mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2]}{2}, \\
\big|\mathbb{E}[\boldsymbol{x}_k^{(j)\top}\phi(\boldsymbol{x}_k^{(j)})]\big| &\leq \mathbb{E}\big[\|\boldsymbol{x}_k^{(j)}\|\,\|\phi(\boldsymbol{x}_k^{(j)})\|\big], \\
&\leq \sqrt{\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2]\mathbb{E}[\|\phi(\boldsymbol{x}_k^{(j)})\|^2]}, \\
&\leq \frac{\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] + \mathbb{E}[\|\phi(\boldsymbol{x}_k^{(j)})\|^2]}{2},
\end{aligned}$$

where we apply the inequality $\sqrt{ab} \leq \frac{a+b}{2}$ to eliminate the square roots. Combining these bounds and substituting them back into (153) yields

$$
\begin{aligned}
\left|\mathbb{E}\left[\boldsymbol{x}_k^{(j)\top}\frac{\boldsymbol{\mathcal{G}}^{(j)}}{n}\right]\right| \leq \frac{p}{n}\Bigg(&\frac{\tau^2}{2}(2B^2(1+2R_q+R_{2q})) \\
&+\tau 2p\Big(B\sqrt{2(1+R_{2q})}+K(1+R_q^{\frac{q-1}{q}})\Big) \\
&+2\tau(L+|\phi(0)|)\Bigg)\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] \\
&+\frac{\tau p}{n}\Big(B\sqrt{2(1+R_{2q})}+K\big(1+R_q^{\frac{q-1}{q}}\big)\Big)\frac{\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2]+\mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2]}{2} \\
&+\frac{\tau}{n}\frac{\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2]+\mathbb{E}[\|\phi(\boldsymbol{x}_k^{(j)})\|^2]}{2}, \\
\leq \frac{p}{n}\Bigg(&\frac{\tau^2}{2}(2B^2(1+2R_q+R_{2q})) \\
&+\tau(2p+\frac{1}{2})\left(B\sqrt{2(1+R_{2q})}+K(1+R_q^{\frac{q-1}{q}})\right) \\
&+2\tau(L+|\phi(0)|)+\frac{\tau}{2p}\Bigg)\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] \\
&+\frac{\tau p}{2n}(B\sqrt{2(1+R_{2q})}+K\big(1+R_q^{\frac{q-1}{q}}\big))\mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2] \\
&+\frac{\tau}{2n}\mathbb{E}[\|\phi(\boldsymbol{x}_k^{(j)})\|^2].
\end{aligned}
\tag{162}
$$

Furthermore, the Lipschitz assumption on ϕ yields $\mathbb{E}[\|\phi(\boldsymbol{x}_k^{(j)})\|^2] \leq 2L^2\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] + 2p\phi(0)^2$ Consequently, we arrive at the following bound:

$$\begin{aligned}
\left|\mathbb{E}\Big[\boldsymbol{x}_k^{(j)\top}\frac{\boldsymbol{\mathcal{G}}^{(j)}}{n}\Big]\right| \leq \frac{p}{n}\Bigg(&\frac{\tau^2}{2}(2B^2(1+2R_q+R_{2q})) \\
&+\tau(2p+\frac{1}{2})\left(B\sqrt{2(1+R_{2q})}+K(1+R_q^{\frac{q-1}{q}})\right) \\
&+2\tau(L+|\phi(0)|)+\frac{\tau}{2p}\Bigg)\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] \\
&+\frac{\tau p}{2n}(B\sqrt{2(1+R_{2q})}+K\big(1+R_q^{\frac{q-1}{q}}\big))\mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2] \\
&+\frac{\tau}{2n}(2L^2\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2]+2p\phi(0)^2), \\
=\frac{p}{n}\Bigg(&\frac{\tau^2}{2}(2B^2(1+2R_q+R_{2q})) \\
&+\tau(2p+\frac{1}{2})\left(B\sqrt{2(1+R_{2q})}+K(1+R_q^{\frac{q-1}{q}})\right) \\
&+2\tau(L+|\phi(0)|)+\frac{\tau}{2p}+\frac{\tau}{2p}2L^2\Bigg)\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] \\
&+\frac{\tau p}{2n}(B\sqrt{2(1+R_{2q})}+K\big(1+R_q^{\frac{q-1}{q}}\big))\mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2]+\frac{\tau}{n}p\phi(0)^2. \qquad (163)
\end{aligned}$$

The second term in Equation (152) can be bounded directly using Lemma 3 as follows:

$$\begin{aligned}
\mathbb{E}[\mathrm{Tr}(\frac{\boldsymbol{\Lambda}_k^{(j)}}{n})] &= \mathbb{E}[\mathrm{Tr}(\tau^2\frac{\boldsymbol{C}_t}{n})], \\
&= \frac{\tau^2}{n}\mathbb{E}[\sum_{i=1}^{p}C_{i,i}], \\
&\leq \frac{\tau^2}{n}pB^2(1+2R_q+R_{2q}). \qquad (164)
\end{aligned}$$

Combining the bounds on all three terms, we arrive at a recursion of the form $a_{k+1} \leq a_kJ_1 + J_2$, where $a_k \coloneqq \mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2]$. Unrolling the recursion and summing the resulting geometric series yields

$$\begin{aligned} a_{k+1} &\leq a_k J_1 + J_2, \\ &= (a_{k-1} J_1 + J_2) J_1 + J_2, \\ &= a_0 J_1^{k+1} + J_2 (1 + J_1 + J_1^2 + \cdots + J_1^k), \\ &= a_0 J_1^{k+1} + J_2 \frac{J_1^{k+1} - 1}{J_1 - 1}, \end{aligned} \tag{165}$$

$$a_k \leq a_0 J_1^k + J_2 \frac{J_1^k - 1}{J_1 - 1}, \tag{166}$$

$$\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] \leq \mathbb{E}[\|\boldsymbol{x}_0^{(j)}\|^2] J_1^k + J_2 \frac{J_1^k - 1}{J_1 - 1}, \tag{167}$$

where we can define J_1 and J_2 as

$$\begin{aligned} J_1 &= 2\frac{p}{n}\Bigg(\frac{\tau^2}{2}\Big(2B^2(1 + 2R_q + R_{2q})\Big) + \tau\Big(2p + \frac{1}{2}\Big)\Big(B\sqrt{2(1 + R_{2q})} + K(1 + R_q^{\frac{q-1}{q}})\Big) \\ &\qquad + 2\tau(L + |\phi(0)|) + \frac{\tau}{2p} + \frac{\tau}{2p} 2L^2\Bigg) + 1, \\ &= \frac{V(q, \tau, p)}{n} + 1, \end{aligned} \tag{168}$$

$$\begin{aligned} J_2 &= \frac{\tau^2}{n} p B^2 (1 + 2R_q + R_{2q}) \\ &\qquad + 2\Bigg(\frac{\tau p}{2n}(B\sqrt{2(1 + R_{2q})} + K\big(1 + R_q^{\frac{q-1}{q}}\big))\mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2] + \frac{\tau}{n} p \phi(0)^2\Bigg), \\ &= \frac{Z(q, \tau, p)}{n}. \end{aligned} \tag{169}$$

Recall that the initial moments of the estimates ($\boldsymbol{x}_{0,i}$) and all moments of the true components ($\boldsymbol{u}_i$) are assumed to be bounded by the assumptions of our theorem. Thus, for $k \leq nT$, we use the classical limit $(1 + a/n)^n \to e^a$ to conclude that $(1 + V/n)^{nT} \leq e^{VT}$ uniformly in n, arriving at

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] &\leq \mathbb{E}[\|\boldsymbol{x}_0^{(j)}\|^2]\Big(\frac{V(q, \tau, p)}{n} + 1\Big)^{nT} + \Big(\frac{Z(q, \tau, p)}{n}\Big)\frac{\Big(\frac{V(q,\tau,p)}{n} + 1\Big)^{nT} - 1}{\Big(\frac{V(q,\tau,p)}{n} + 1\Big) - 1}, \\ &\leq A_1(T), \end{aligned} \tag{170}$$

where $A_1(T)$ is a finite constant depending on T. We have established (150); the proof of (151) is analogous. This completes the proof. ■

For the squared norm of $\boldsymbol{\mathcal{G}}_k^{(j)}$, we can write the following expression:

$$\begin{aligned}\|\boldsymbol{\mathcal{G}}_k^{(j)}\|^2 = \frac{\tau^4}{4}\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{C}}^\top\tilde{\boldsymbol{C}}\boldsymbol{x}_k^{(j)} + \tau^2\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{M}}^\top\tilde{\boldsymbol{M}}\boldsymbol{x}_k^{(j)} + \tau^2\boldsymbol{u}^{(j)\top}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\boldsymbol{u}^{(j)} \\ + \tau^2\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{R}}^\top\tilde{\boldsymbol{R}}\boldsymbol{x}_k^{(j)} + \tau^2\phi(\boldsymbol{x}_k^{(j)})^\top\phi(\boldsymbol{x}_k^{(j)}) \\ + \tau^3\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{C}}^\top\tilde{\boldsymbol{M}}\boldsymbol{x}_k^{(j)} - \tau^3\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{C}}^\top\boldsymbol{\Psi}\boldsymbol{u}^{(j)} - 2\tau^2\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{M}}^\top\boldsymbol{\Psi}\boldsymbol{u}^{(j)} \\ - \tau^3\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{C}}^\top\tilde{\boldsymbol{R}}\boldsymbol{x}_k^{(j)} - 2\tau^2\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{M}}^\top\tilde{\boldsymbol{R}}\boldsymbol{x}_k^{(j)} + 2\tau^2\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{R}}^\top\boldsymbol{\Psi}\boldsymbol{u}^{(j)} \\ + \tau^3\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{C}}^\top\phi(\boldsymbol{x}_k^{(j)}) + 2\tau^2\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{M}}^\top\phi(\boldsymbol{x}_k^{(j)}) \\ - 2\tau^2\boldsymbol{x}_k^{(j)\top}\tilde{\boldsymbol{R}}^\top\phi(\boldsymbol{x}_k^{(j)}) - 2\tau^2\boldsymbol{u}^{(j)\top}\boldsymbol{\Psi}^\top\phi(\boldsymbol{x}_k^{(j)}). \end{aligned} \tag{171}$$

Taking expectations and applying the triangle inequality, we have to establish bounds on every term. We start by obtaining

$$\mathbb{E}[\boldsymbol{x}^\top\tilde{\boldsymbol{C}}^\top\tilde{\boldsymbol{C}}\boldsymbol{x}] \le \mathbb{E}[\|\tilde{\boldsymbol{C}}\|_2^2\,\|\boldsymbol{x}\|^2] \le \left(2pB^2(1+2R_q+R_{2q})\right)^2\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2], \tag{172}$$

$$\mathbb{E}[\boldsymbol{x}^\top\tilde{\boldsymbol{M}}^\top\tilde{\boldsymbol{M}}\boldsymbol{x}] \le \mathbb{E}[\|\tilde{\boldsymbol{M}}\|_2^2\,\|\boldsymbol{x}\|^2] \le 4p^4\left(B\sqrt{2(1+R_{2q})} + K(1+R_q^{\frac{q-1}{q}})\right)^2\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2], \tag{173}$$

$$\mathbb{E}[\boldsymbol{u}^\top\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\boldsymbol{u}] \le \mathbb{E}[\|\boldsymbol{\Psi}\|_2^2\,\|\boldsymbol{u}\|^2] \le p^2\left(B\sqrt{2(1+R_{2q})} + K\big(1+R_q^{\frac{q-1}{q}}\big)\right)^2\mathbb{E}[\|\boldsymbol{u}^{(j)}\|^2], \tag{174}$$

$$\mathbb{E}[\boldsymbol{x}^\top\tilde{\boldsymbol{R}}^\top\tilde{\boldsymbol{R}}\boldsymbol{x}] \le \mathbb{E}[\|\tilde{\boldsymbol{R}}\|_2^2\,\|\boldsymbol{x}\|^2] \le 4p^2(L+|\phi(0)|)^2\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2], \tag{175}$$

$$\mathbb{E}[\|\phi(\boldsymbol{x}_k^{(j)})\|^2] \le 2L^2\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] + 2p\phi(0)^2. \tag{176}$$

The same bounding logic can be applied to the cross terms by substituting the spectral norm bounds we obtained in the previous sections. All these terms, with the exception of $\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2]$, are independent of n. Having established that $\max_{k\le nT}\mathbb{E}[\|\boldsymbol{x}_k^{(j)}\|^2] \le A_1(T)$ in Lemma 6, we can conclude:

$$\max_{k\le nT}\mathbb{E}\|\boldsymbol{\mathcal{G}}_k^{(j)}\|_2^2 \le C(T). \tag{177}$$

The second condition of Assumption C.8 requires bounding the Frobenius norm of $\boldsymbol{\Lambda}_k$, we obtain

$$\|\boldsymbol{\Lambda}_k\|_F^2 = \tau^4\sum_i\sum_j((\boldsymbol{C}_t)_{i,j})^2. \tag{178}$$

Applying the element-wise bound we obtained from Lemma 3 yields

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\Lambda}_k\|_F^2] &= \mathbb{E}[\tau^4\sum_{i=1}^p\sum_{j=1}^p((\boldsymbol{C}_k)_{i,j})^2], \\ &\le \tau^4p^2B^4(1+2R_q+R_{2q})^2,\end{aligned} \tag{179}$$

which is again a finite constant independent of n and k. Consequently, we conclude that

$$\max_{k\le nT} \mathbb{E}\|\mathbf{\Lambda}_k\|_F^2 \le C(T). \tag{180}$$

C.9. Assumption 9

For each $T > 0$, there exists $C(T) < \infty$ such that $max_{k\le nT}\mathbb{E}[\|\mathbf{\Delta}_k^{(j)}\|^4] \le C(T)n^{-2}$, and for any $i \ne j$:

$$\max_{k\le nT}\mathbb{E}\left[\left|\mathbb{E}_k\left[\left(\mathbf{\Delta}_k^{(i)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(i)}]\right)^\top\left(\mathbf{\Delta}_k^{(j)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(j)}]\right)\right]\right|\right] \le \frac{C(T)}{n^2}. \tag{181}$$

First, by Lemma 6, the bound $\max_{k\le nT}\mathbb{E}[\|\mathbf{\Delta}_k^{(j)}\|^4] \le C(T)n^{-2}$ follows from the same argument used to verify Assumption C.8. For the latter condition, we expand the product:

$$\mathbb{E}_k\left[\left(\mathbf{\Delta}_k^{(i)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(i)}]\right)^\top\left(\mathbf{\Delta}_k^{(j)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(j)}]\right)\right] = \mathbb{E}_k[\mathbf{\Delta}_k^{(i)\top}\mathbf{\Delta}_k^{(j)}] - \mathbb{E}_k[\mathbf{\Delta}_k^{(i)\top}]\mathbb{E}_k[\mathbf{\Delta}_k^{(j)}], \tag{182}$$

From Equation 53, the cross terms are of order $O(1/n^2)$ when $i \ne j$. Combined with the definition of $\mathbf{\Delta}_k^{(j)}$, this yields $\mathbb{E}[\mathbf{\Delta}_k^{(j)\top}\mathbf{\Delta}_k^{(i)}] = O(1/n^2)$. Following the definition of $\boldsymbol{\mathcal{G}}_k^{(j)}$, we arrive at

$$\mathbb{E}_k\left[\left(\mathbf{\Delta}_k^{(i)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(i)}]\right)^\top\left(\mathbf{\Delta}_k^{(j)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(j)}]\right)\right] = -\frac{\boldsymbol{\mathcal{G}}_k^{(i)\top}\boldsymbol{\mathcal{G}}_k^{(j)}}{n^2} + O(1/n^2). \tag{183}$$

Finally, taking the outer expectation on both sides yields

$$\mathbb{E}\left[\left|\mathbb{E}_k\left[\left(\mathbf{\Delta}_k^{(i)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(i)}]\right)^\top\left(\mathbf{\Delta}_k^{(j)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(j)}]\right)\right]\right|\right] \le \mathbb{E}\left[\left|-\frac{\boldsymbol{\mathcal{G}}_k^{(i)\top}\boldsymbol{\mathcal{G}}_k^{(j)}}{n^2} + O(1/n^2)\right|\right]. \tag{184}$$

For the right hand side, we can establish the following bound as

$$\begin{aligned}
\mathbb{E}\left[\left|-\frac{\boldsymbol{\mathcal{G}}_k^{(i)\top}\boldsymbol{\mathcal{G}}_k^{(j)}}{n^2} + O(1/n^2)\right|\right] &\le \mathbb{E}\left[\left|\frac{[\boldsymbol{\mathcal{G}}_k^{(i)\top}\boldsymbol{\mathcal{G}}_k^{(j)}]}{n^2}\right|\right] + O(1/n^2), \\
&\le \frac{\sqrt{\mathbb{E}[\|\boldsymbol{\mathcal{G}}_k^{(i)}\|^2]\mathbb{E}[\|\boldsymbol{\mathcal{G}}_k^{(j)}]\|^2]}}{n^2} + O(1/n^2), \\
&\le \frac{\mathbb{E}[\|\boldsymbol{\mathcal{G}}_k^{(i)}\|^2] + \mathbb{E}[\|\boldsymbol{\mathcal{G}}_k^{(j)}\|^2]}{2n^2} + O(1/n^2).
\end{aligned} \tag{185}$$

In the verification of Assumption C.8, we showed that for some $C(T) < \infty$ we have

$$\max_{k\le nT} \mathbb{E}\|\boldsymbol{\mathcal{G}}_k^{(j)}\|^2 \le C(T). \tag{186}$$

Consequently, Assumption 9 is satisfied directly from Equation 185, which establishes our claim:

$$\max_{k\le nT}\mathbb{E}\left|[\mathbb{E}_k\left[\left(\mathbf{\Delta}_k^{(i)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(i)}]\right)^\top\left(\mathbf{\Delta}_k^{(j)} - \mathbb{E}_k[\mathbf{\Delta}_k^{(j)}]\right)\right]\right| \le \frac{C(T)}{n^2}. \tag{187}$$

C.10. Assumption 10

For each $b > 0$ and $T > 0$, the following PDE (in weak form) has a unique solution in $D([0,T], \mathcal{M}(\mathbb{R}^{2p}))$ for all bounded test functions $\varphi(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{C}^3(\mathbb{R}^{2p})$:

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle + \int_0^t \left\langle \nabla_{\boldsymbol{x}}^\top \varphi \; \boldsymbol{\mathcal{G}}(\boldsymbol{x}_{\hat{t}}, \boldsymbol{u}_{\hat{t}}, \boldsymbol{\Theta}_{\hat{t}} \sqcap b), \mu_{\hat{t}} \right\rangle \mathrm{d}\hat{t} + \frac{1}{2} \int_0^t \left\langle \mathrm{Tr}(\boldsymbol{\Lambda}(\boldsymbol{x}_{\hat{t}}, \boldsymbol{u}_{\hat{t}}, \boldsymbol{\Theta}_{\hat{t}} \sqcap b) \nabla_{\boldsymbol{x}}^2 \varphi), \mu_{\hat{t}} \right\rangle \mathrm{d}\hat{t}. \tag{188}$$

The sufficient conditions for this uniqueness assumption are:

10.1 $\langle \|\boldsymbol{u}\|^2, \mu_0 \rangle \leq L$ and $\langle \|\boldsymbol{x}\|^2, \mu_0 \rangle \leq V$, where L, V are two generic constants.

10.2 For any $\boldsymbol{x}, \tilde{\boldsymbol{x}}, \boldsymbol{u}$, and $\boldsymbol{\Theta}$, we have $\|\boldsymbol{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\Theta}) - \boldsymbol{\mathcal{G}}(\tilde{\boldsymbol{x}}, \boldsymbol{u}, \boldsymbol{\Theta})\| \leq L(1 + \|\boldsymbol{\Theta}\|_F)\|(\boldsymbol{x} - \tilde{\boldsymbol{x}})\|$.

10.3 $|\boldsymbol{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\Theta}) - \boldsymbol{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{u}, \tilde{\boldsymbol{\Theta}})| \leq L(1 + \|\boldsymbol{u}\| + \|\boldsymbol{x}\|)\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|_F$.

10.4 $|\boldsymbol{\mathcal{G}}(\boldsymbol{x}, \boldsymbol{u}, \tilde{\boldsymbol{\Theta}})| \leq L(1 + |\boldsymbol{u}| + |\boldsymbol{x}|)(\|\boldsymbol{\Theta}\|_F + 1)$.

10.5 For any $\boldsymbol{\Theta}$ and $\tilde{\boldsymbol{\Theta}}$, we have $|\boldsymbol{\Lambda}^{\frac{1}{2}}(\boldsymbol{\Theta}) - \boldsymbol{\Lambda}^{\frac{1}{2}}(\tilde{\boldsymbol{\Theta}})| \leq L\|\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}\|_F$.

10.6 $\Lambda^{\frac{1}{2}}(\boldsymbol{\Theta}) \leq L(1 + \|\boldsymbol{\Theta}\|_F)$.

Given our definitions of $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\Lambda}$, the verification of conditions (10.1)–(10.6) follows from standard algebraic arguments and is omitted for brevity. In particular, to satisfy Assumption 10.4, we required the function $\phi(x)$ to be Lipschitz. Following that we verified assumptions C.1-C.10 hold in our setting, the deterministic measure μ_t is the unique solution of the following partial differential equation:

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle + \int_0^t \left\langle \nabla_{\boldsymbol{x}} \varphi^\top \, \boldsymbol{\mathcal{G}}_s, \mu_s \right\rangle \mathrm{d}s + \int_0^t \frac{1}{2} \left\langle \mathrm{Tr}\big(\boldsymbol{\Lambda}_s \, \nabla_{\boldsymbol{x}}^2 \varphi\big), \mu_s \right\rangle \mathrm{d}s. \tag{189}$$

Furthermore, if μ_t admits a density $P_t(\boldsymbol{x}, \boldsymbol{u})$, the following strong form PDE holds:

$$\frac{\partial P_t}{\partial t} = -\nabla_{\boldsymbol{x}}\big(\boldsymbol{\mathcal{G}}_t \, P_t\big) + \frac{1}{2} \, \mathrm{Tr}\big(\boldsymbol{\Lambda}_t \, \nabla_{\boldsymbol{x}}^2 P_t\big). \tag{190}$$

In the theorem, the used time-embedding $k = \lfloor tn \rfloor$ with $t \geq 0$ and the considered assumptions are consistent with the prior mean-field literature [11, 18]. Furthermore, we note that the Lipschitz assumption on ϕ is a byproduct of our proof technique, whereas our experiments show that our results remain accurate beyond the Lipschitz class.

C.11. Derivation of Corollary 2

By selecting an appropriate test function φ in (104), we can derive a closed-form system of ODEs. Specifically, we consider the case $\phi(x) = 0$, which consequently enforces $R_{i,j} = 0$. We provide additional total and marginal trajectories for the limiting PDE under $\phi(x) = 0$ in Figure 4, where our theory demonstrates close agreement with numerical simulations. To explicitly obtain the macroscopic evolution for the indices (i, j), we instantiate the test function as

$$\varphi_{l,m}(\boldsymbol{x}, \boldsymbol{u}) := x_l u_m. \tag{191}$$

Then the weak pairing recovers the overlap matrix:

$$\langle \varphi_{l,m}, \mu_t \rangle = \int x_l u_m \, \mu_t(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{u}) = Q_{t,l,m}. \tag{192}$$

Moreover, since $\varphi_{l,m}$ is linear in $\boldsymbol{x}$, we have $\nabla^2_{\boldsymbol{x}} \varphi_{l,m} \equiv 0$, so the diffusion term in (104) vanishes. Therefore, the evolution reduces to

$$\frac{d}{dt} Q_{t,l,m} = \int u_m \, \omega_l \, \mu_t(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{u}). \tag{193}$$

This follows directly from (99) by setting $\phi(x) = 0$ and $R_{i,j} = 0$, yielding the ODE in summation form:

$$\begin{aligned} \frac{d}{dt} Q_{lj} = -\tau^2 \left(\frac{1}{2} Q_{lj} \langle \gamma_l^2 \rangle + \sum_{i<l} Q_{i,j} \langle \gamma_i \gamma_l \rangle \right) + \tau \, \psi_l^{(j)} - \tau Q_{lj} \boldsymbol{Q}_l^\top \boldsymbol{\psi}_l \\ - \tau \sum_{i<l} Q_{i,j} \Big(\boldsymbol{Q}_i^\top \boldsymbol{\psi}_l + \boldsymbol{Q}_l^\top \boldsymbol{\psi}_i \Big). \end{aligned} \tag{194}$$

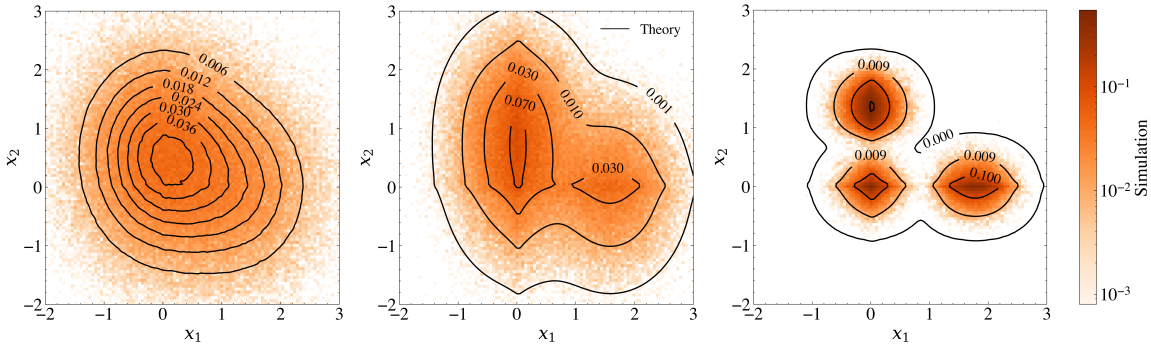
Expressing the above expression in matrix form using the previously defined matrices $\boldsymbol{C}_t$, $\boldsymbol{M}_t$ and $\boldsymbol{\Psi}_t$ yields the final closed-form ODE for the matrix $\boldsymbol{Q}_t$:

$$\frac{d}{dt} \boldsymbol{Q} = -\frac{\tau^2}{2} \mathcal{T}(\boldsymbol{C}_t) \boldsymbol{Q} - \tau \mathcal{T}(\boldsymbol{M}_t) \boldsymbol{Q} + \tau \, \boldsymbol{\Psi}_t^\top, \tag{195}$$

where the matrix operator $\mathcal{T}(\cdot)$ is defined in the formal statement of Theorem 1.

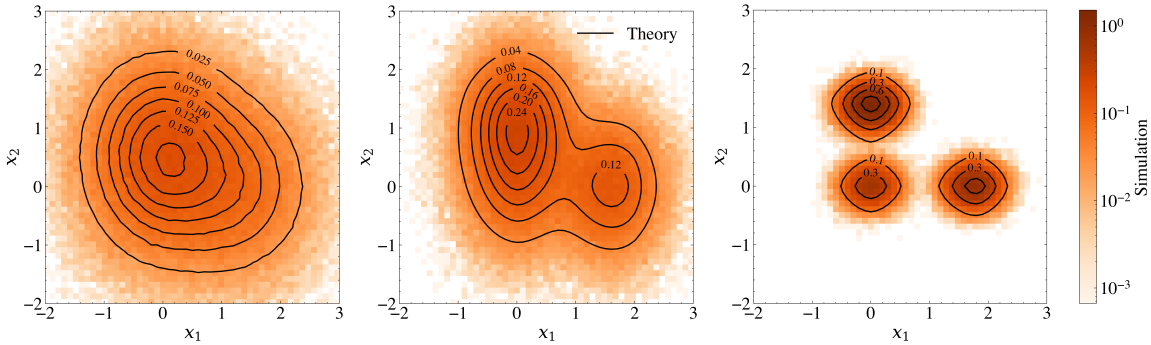
Appendix D. Evolution of the limiting density

For completeness, we provide the joint and marginal probability densities. Figure 3 and Figure 4 demonstrates that the strong-form PDE in Corollary 1 accurately tracks the empirical evolution of the joint density in simulations. Corresponding marginal density plots are given in Figure 5 and 6.



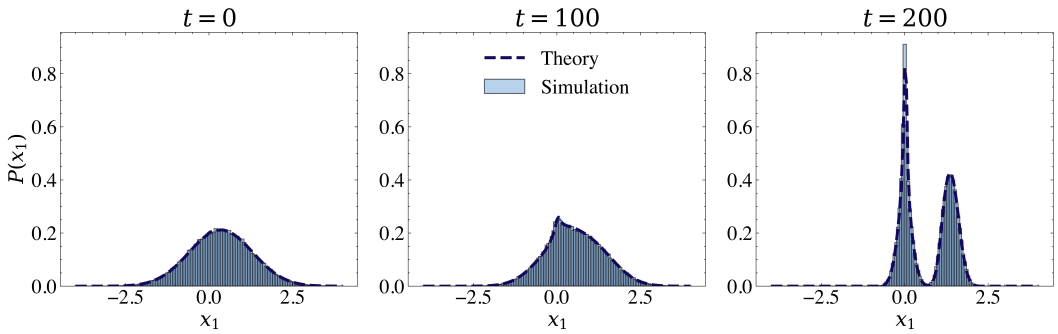
(a) $t = 0$ (b) $t = 100$ (c) $t = 200$

Figure 3: Evolution of the joint probability limiting density for $p = 2$. Comparison between theoretical predictions (contours) and Monte Carlo simulations (heatmaps) for $P_t(x_1, x_2, u_1, u_2)$ at times $t = 0, 100, 200$. Here, we use the setting of Example 1. We set $\beta_1 = 1$ for c_1 while β_2 is set to 0 for c_2. In simulations, component vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are drawn from sparse distributions: $\mathbb{P}(u = 1/\sqrt{\rho_i}) = \rho_i$ and $\mathbb{P}(u = 0) = 1 - \rho_i$ with sparsity levels $\rho_1 = 0.5$ and $\rho_2 = 0.3$. This choice enables a clear visualization of the convergence behavior, with the components evolving toward three distinct regions in the $(\mathbf{x}_1, \mathbf{x}_2)$ space. We employ $\phi(x) = 0.1\ \mathrm{sgn}(x)$, corresponding to L_1 regularization commonly used in practice to induce sparsity. Here, $n = 5000$, and $\tau = 0.01$.

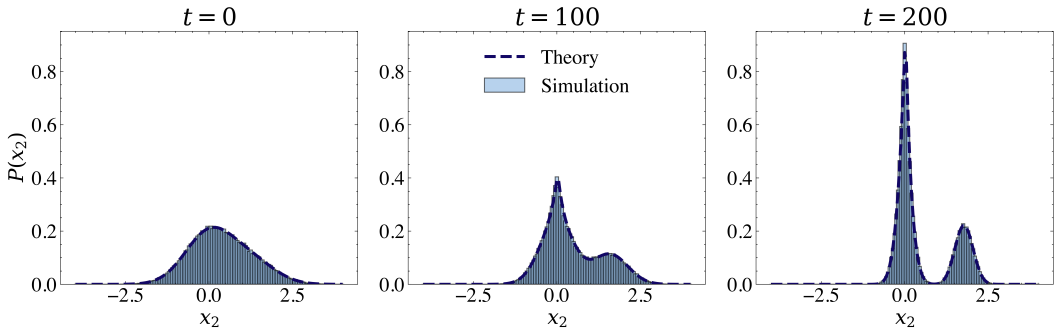


(a) $t = 0$ (b) $t = 100$ (c) $t = 200$

Figure 4: Evolution of the joint limiting probability density for $p = 2$. Comparison between theoretical predictions (contours) and Monte Carlo simulations (heatmaps) for $P_t(x_1, x_2, u_1, u_2)$ at times $t = 0, 100, 200$. The setting corresponds to sparse component vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ for $\phi(x) = 0$. Here we use the setting of Example 1. For c_1, we set $\beta_1 = 1$, for c_2, we set $\beta_2 = 0$. In simulations, component vectors are drawn from distributions ($\mathbb{P}(u = 1/\sqrt{\rho_i}) = \rho_i$ and $\mathbb{P}(u = 0) = 1 - \rho_i$) with sparsity levels $\rho_1 = 0.5$ and $\rho_2 = 0.3$. This choice enables a clear visualization of the convergence behavior , with the components evolving toward three distinct regions in the $(\mathbf{x}_1, \mathbf{x}_2)$ space. Here, $n = 1000$, $\tau = 0.01$

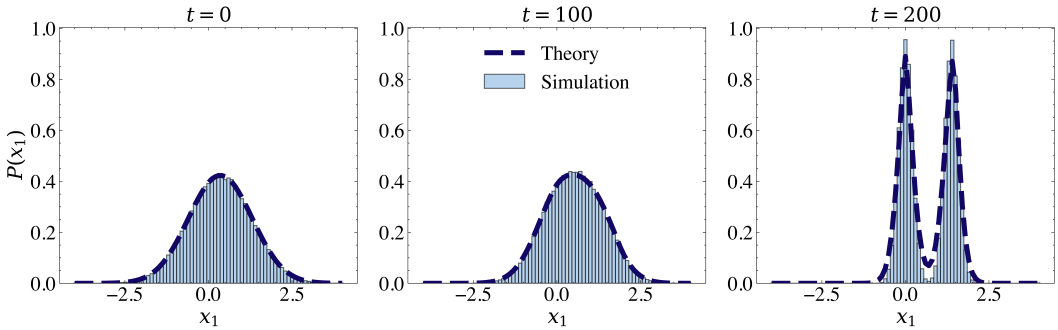


(a) Marginal density of the first component x_1 at times $t = 0, 100$, and 200.

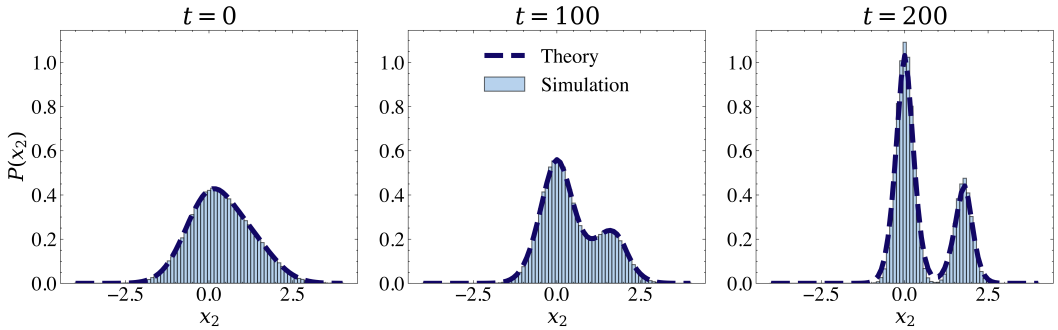


(b) Marginal density of the second component x_2 at times $t = 0, 100$, and 200.

Figure 5: Evolution of the limiting marginal densities, corresponding to the total density in Figure 3. The setting corresponds to sparse component vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, for $\phi(x) = 0.1sgn(x)$. Here we use the setting of Example 1: we set $\beta_1 = 1$ for c_1, and set $\beta_2 = 0$ for c_2. In Monte Carlo simulations, component vectors are drawn from distributions ($\mathbb{P}(u = 1/\sqrt{\rho_i}) = \rho_i$ and $\mathbb{P}(u = 0) = 1 - \rho_i$) with sparsity levels $\rho_1 = 0.5$ and $\rho_2 = 0.3$. The dashed dark blue curves correspond to the PDE prediction, while the light blue histograms represent Monte Carlo simulations. Here $n = 5000$ and $\tau = 0.01$.



(a) Marginal density of the first component x_1 at times $t = 0, 100$, and 200.



(b) Marginal density of the second component x_2 at times $t = 0, 100$, and 200.

Figure 6: Evolution of the limiting marginal densities corresponding to the total density in Figure 4. The setting corresponds to sparse component vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, for $\phi(x) = 0$. Here we use the setting of Example 1. For c_1, we set $\beta_1 = 1$, for c_2, we set $\beta_2 = 0$. In Monte Carlo simulations, component vectors are drawn from sparse distributions ($\mathbb{P}(u = 1/\sqrt{\rho_i}) = \rho_i$ and $\mathbb{P}(u = 0) = 1 - \rho_i$) with sparsity levels $\rho_1 = 0.5$ and $\rho_2 = 0.3$. The dashed dark blue curves correspond to the PDE prediction, while the light blue histograms represent Monte Carlo simulations. Here $n = 1000$ and $\tau = 0.01$.

Appendix E. Derivation of the ODEs for cubic nonlinearity: $f(x) = \pm x^3$

For the cubic nonlinearity $f(x) = \pm x^3$ and $p = 2$, we explicitly derive the system of coupled ODEs. We begin by computing the expectations with respect to the random variables $c_{i,t}$ and e_i for $i \in \{1, 2\}$. For notational brevity, we omit the index t and focus on the positive case; the derivation for $f(x) = -x^3$ follows immediately by symmetry. We start with the expression for γ_i:

$$\begin{aligned}
\gamma_i &= \left(c_1 Q_{i,1} + c_2 Q_{i,2} + e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2}\right)^3, \\
&= (c_1 Q_{i,1})^3 + (c_2 Q_{i,2})^3 + \left(e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2}\right)^3 \\
&\quad + 3(c_1 Q_{i,1})^2 (c_2 Q_{i,2}) + 3(c_1 Q_{i,1})^2 e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2} \\
&\quad + 3(c_2 Q_{i,2})^2 (c_1 Q_{i,1}) + 3(c_2 Q_{i,2})^2 e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2} \\
&\quad + 3\left(e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2}\right)^2 (c_1 Q_{i,1}) + 3\left(e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2}\right)^2 (c_2 Q_{i,2}) \\
&\quad + 6(c_1 Q_{i,1})(c_2 Q_{i,2}) e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2}. \qquad (196)
\end{aligned}$$

For the derivative $\gamma_i^{'}$ we obtain the following:

$$\gamma_i^{'} = 3\left(c_1 Q_{i,1} + c_2 Q_{i,2} + e_1\sqrt{1 - Q_{i,1}^2 - Q_{i,2}^2}\right)^2. \qquad (197)$$

Let $m_{i,j} := \langle c_i^j \rangle$ denote the j-th moment of the variable c_i, also recall that $\langle c_i \rangle = 0$ and $\langle c_i^2 \rangle = 1$, this yields the following expectations:

$$\langle \gamma_i^{'} \rangle = 3(Q_{i,1}^2 + Q_{i,2}^2 + 1 - Q_{i,1}^2 - Q_{i,2}^2) = 3, \qquad (198)$$

$$\langle c_1 \gamma_i \rangle = Q_{i,1}^3 m_{1,4} + 3Q_{i,2}^2 Q_{i,1} + 3(1 - Q_{i,1}^2 - Q_{i,2}^2) Q_{i,1} = Q_{i,1}^3 (m_{1,4} - 3) + 3Q_{i,1}, \qquad (199)$$

$$\langle c_2 \gamma_i \rangle = Q_{i,2}^3 (m_{2,4} - 3) + 3Q_{i,2}. \qquad (200)$$

Furthermore, we require the expectations of the squared terms γ_i^2 and cross term $\gamma_1\gamma_2$, which we calculate as follows:

$$\begin{aligned}
\langle \gamma_1^2 \rangle &= Q_{1,1}^6 m_{1,6} + Q_{1,2}^6 m_{2,6} + 15(1 - Q_{1,1}^2 - Q_{1,2}^2)^3 + 15Q_{1,2}^4 (1 - Q_{1,2}^2) m_{2,4} \\
&\quad + 15Q_{1,1}^4 (1 - Q_{1,1}^2) m_{1,4} + 45(1 - Q_{1,1}^2 - Q_{1,2}^2)^2 (Q_{1,1}^2 + Q_{1,2}^2) \\
&\quad + 90Q_{1,1}^2 Q_{1,2}^2 (1 - Q_{1,1}^2 - Q_{1,2}^2) + 20Q_{1,2}^3 Q_{1,1}^3 m_{1,3} m_{2,3}\ , \qquad (201)
\end{aligned}$$

$$\begin{aligned}
\langle \gamma_2^2 \rangle &= Q_{2,1}^6 m_{1,6} + Q_{2,2}^6 m_{2,6} + 15(1 - Q_{2,1}^2 - Q_{2,2}^2)^3 + 15Q_{2,2}^4 (1 - Q_{2,2}^2) m_{2,4} \\
&\quad + 15Q_{2,1}^4 (1 - Q_{2,1}^2) m_{1,4} + 45(1 - Q_{2,1}^2 - Q_{2,2}^2)^2 (Q_{2,1}^2 + Q_{2,2}^2) \\
&\quad + 90Q_{2,1}^2 Q_{2,2}^2 (1 - Q_{2,1}^2 - Q_{2,2}^2) + 20Q_{2,2}^3 Q_{2,1}^3 m_{1,3} m_{2,3}\ , \qquad (202)
\end{aligned}$$

$$
\begin{aligned}
\langle \gamma_1\gamma_2 \rangle = \Bigg\langle & \left(c_1 Q_{1,1} + c_2 Q_{1,2} + e_1\sqrt{1 - Q_{1,1}^2 - Q_{1,2}^2}\right)^3 \\
& \times \left(c_1 Q_{2,1} + c_2 Q_{2,2} + e_2\sqrt{1 - Q_{2,1}^2 - Q_{2,2}^2}\right)^3 \Bigg\rangle, \\
= Q_{1,1}^3 & \left(Q_{2,1}^3 m_{1,6} + Q_{2,2}^3 m_{1,3} m_{2,3} + 3(1 - Q_{2,1}^2) Q_{2,1} m_{1,4}\right) \\
& + Q_{1,2}^3 \left(Q_{2,2}^3 m_{2,6} + Q_{2,1}^3 m_{1,3} m_{2,3} + 3(1 - Q_{2,2}^2) Q_{2,2} m_{2,4}\right) \\
& + 3Q_{1,1}^2 Q_{1,2} \Bigg(Q_{2,2}^3 m_{2,4} + 3Q_{2,1}^2 Q_{2,2} m_{1,4} \\
& \qquad + 3Q_{2,2}^2 Q_{2,1} m_{1,3} m_{2,3} + 3(1 - Q_{2,1}^2 - Q_{2,2}^2) Q_{2,2}\Bigg) \\
& + 3Q_{1,2}^2 Q_{1,1} \Bigg(Q_{2,1}^3 m_{1,4} + 3Q_{2,2}^2 Q_{2,1} m_{2,4} \\
& \qquad + 3Q_{2,1}^2 Q_{2,2} m_{1,3} m_{2,3} + 3(1 - Q_{2,1}^2 - Q_{2,2}^2) Q_{2,1}\Bigg) \\
& + 3Q_{1,1}\left(1 - Q_{1,1}^2 - Q_{1,2}^2\right)\left(Q_{2,1}^3 m_{1,4} + 3(1 - Q_{2,1}^2) Q_{2,1}\right) \\
& + 3Q_{1,2}\left(1 - Q_{1,1}^2 - Q_{1,2}^2\right)\left(Q_{2,2}^3 m_{2,4} + 3(1 - Q_{2,2}^2) Q_{2,2}\right) . \qquad (203)
\end{aligned}
$$

To streamline the presentation of the ODEs, we define the following functions representing the expectations of $\gamma_i\gamma_j$:

$$F(Q_{1,1}, Q_{2,1}, Q_{1,2}, Q_{2,2}) \coloneqq \langle \gamma_1\gamma_2 \rangle, \qquad (204)$$

$$W(Q_{i,1}, Q_{i,2}) \coloneqq \langle \gamma_i^2 \rangle. \qquad (205)$$

We also define the function D for notational convenience, which appears in the drift coefficient of the second estimate:

$$
\begin{aligned}
D(Q_{1,1}, Q_{2,1}, Q_{1,2}, Q_{2,2}) = {} & Q_{1,1} Q_{2,1}^3 (m_{1,4} - 3) + Q_{1,2} Q_{2,2}^3 (m_{2,4} - 3) \\
& + Q_{2,1} Q_{1,1}^3 (m_{1,4} - 3) + Q_{2,2} Q_{1,2}^3 (m_{2,4} - 3). \qquad (206)
\end{aligned}
$$

Finally, for the drift and diffusion terms we arrive at the following final expressions:

$$\begin{aligned}\omega_2 &= \tau\Big(3x_2^{(j)} + u_1^{(j)}Q_{2,1}^3(m_{1,4}-3) + u_2^{(j)}Q_{2,2}^3(m_{2,4}-3)\Big) - \frac{x_2^{(j)}\tau^2}{2}W(Q_{2,1},Q_{2,2}) \\ &\quad - x_1^{(j)}\tau^2 F(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) - x_1^{(j)}\tau D(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) \\ &\quad - x_2^{(j)}\tau\Big(3 + Q_{2,1}^4[m_{1,4}-3] + Q_{2,2}^4[m_{2,4}-3]\Big), \\ &= \tau\Big(u_1^{(j)}Q_{2,1}^3(m_{1,4}-3) + u_2^{(j)}Q_{2,2}^3(m_{2,4}-3)\Big) - \frac{x_2^{(j)}\tau^2}{2}W(Q_{2,1},Q_{2,2}) \\ &\quad - x_1^{(j)}\tau^2 F(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) - x_1^{(j)}\tau D(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) \\ &\quad - x_2^{(j)}\tau\Big(Q_{2,1}^4[m_{1,4}-3] + Q_{2,2}^4[m_{2,4}-3]\Big),\end{aligned} \tag{207}$$

Consequently, for the first estimate we get

$$\begin{aligned}\omega_1 &= \tau\Big(u_1^{(j)}Q_{1,1}^3(m_{1,4}-3) + u_2^{(j)}Q_{1,2}^3(m_{2,4}-3)\Big) \\ &\quad - \frac{x_1^{(j)}\tau^2}{2}W(Q_{1,1},Q_{1,2}) - x_1^{(j)}\tau\Big(Q_{1,1}^4(m_{1,4}-3) + Q_{1,2}^4(m_{2,4}-3)\Big).\end{aligned} \tag{208}$$

Combining these results, we arrive at the following system of coupled ODEs for $f(x) = +x^3$:

$$\begin{aligned}\frac{d}{dt}Q_{1,1} &= \tau Q_{1,1}^3(m_{1,4}-3) - \frac{Q_{1,1}\tau^2}{2}W(Q_{1,1},Q_{1,2}) \\ &\quad - \tau Q_{1,1}\Big(Q_{1,1}^4(m_{1,4}-3) + Q_{1,2}^4(m_{2,4}-3)\Big),\end{aligned} \tag{209}$$

$$\begin{aligned}\frac{d}{dt}Q_{1,2} &= \tau Q_{1,2}^3(m_{2,4}-3) - \frac{Q_{1,2}\tau^2}{2}W(Q_{1,1},Q_{1,2}) \\ &\quad - \tau Q_{1,2}\Big(Q_{1,1}^4(m_{1,4}-3) + Q_{1,2}^4(m_{2,4}-3)\Big),\end{aligned} \tag{210}$$

$$\begin{aligned}\frac{d}{dt}Q_{2,1} &= \tau Q_{2,1}^3(m_{1,4}-3) - \frac{\tau^2}{2}Q_{2,1}\,W(Q_{2,1},Q_{2,2}) \\ &\quad - \tau^2 Q_{1,1}F(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) \\ &\quad - \tau Q_{1,1}\,D(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) \\ &\quad - \tau Q_{2,1}\Big(Q_{2,1}^4(m_{1,4}-3) + Q_{2,2}^4(m_{2,4}-3)\Big),\end{aligned} \tag{211}$$

$$\begin{aligned}\frac{d}{dt}Q_{2,2} &= \tau\big(Q_{2,2}^3(m_{2,4}-3)\big) - \frac{\tau^2}{2}Q_{2,2}\,W(Q_{2,1},Q_{2,2}) \\ &\quad - \tau^2 Q_{1,2}\,F(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) \\ &\quad - \tau Q_{1,2}\,D(Q_{1,1},Q_{2,1},Q_{1,2},Q_{2,2}) \\ &\quad - \tau Q_{2,2}\Big(Q_{2,1}^4(m_{1,4}-3) + Q_{2,2}^4(m_{2,4}-3)\Big).\end{aligned} \tag{212}$$

Consequently, it follows directly for $f(x) = \pm x^3$:

$$\frac{dQ_{1,1}}{dt} = \pm\tau Q_{1,1}^3(m_{1,4} - 3) - \frac{Q_{1,1}\tau^2}{2} W(Q_{1,1}, Q_{1,2}) \mp \tau Q_{1,1}\Big(Q_{1,1}^4(m_{1,4} - 3) + Q_{1,2}^4(m_{2,4} - 3)\Big), \tag{213}$$

$$\frac{dQ_{1,2}}{dt} = \pm\tau Q_{1,2}^3(m_{2,4} - 3) - \frac{Q_{1,2}\tau^2}{2} W(Q_{1,1}, Q_{1,2}) \mp \tau Q_{1,2}\Big(Q_{1,1}^4(m_{1,4} - 3) + Q_{1,2}^4(m_{2,4} - 3)\Big), \tag{214}$$

$$\begin{aligned}\frac{dQ_{2,1}}{dt} = &\pm\tau Q_{2,1}^3(m_{1,4} - 3) - \frac{\tau^2}{2} Q_{2,1}\, W(Q_{2,1}, Q_{2,2}) \\ &- \tau^2 Q_{1,1}\, F(Q_{1,1}, Q_{2,1}, Q_{1,2}, Q_{2,2}) \\ &\mp \tau Q_{1,1}\, D(Q_{1,1}, Q_{2,1}, Q_{1,2}, Q_{2,2}) \\ &\mp \tau Q_{2,1}\Big(Q_{2,1}^4(m_{1,4} - 3) + Q_{2,2}^4(m_{2,4} - 3)\Big),\end{aligned} \tag{215}$$

$$\begin{aligned}\frac{dQ_{2,2}}{dt} = &\pm\tau Q_{2,2}^3(m_{2,4} - 3) - \frac{\tau^2}{2} Q_{2,2}\, W(Q_{2,1}, Q_{2,2}) \\ &- \tau^2 Q_{1,2}\, F(Q_{1,1}, Q_{2,1}, Q_{1,2}, Q_{2,2}) \\ &\mp \tau Q_{1,2}\, D(Q_{1,1}, Q_{2,1}, Q_{1,2}, Q_{2,2}) \\ &\mp \tau Q_{2,2}\Big(Q_{2,1}^4(m_{1,4} - 3) + Q_{2,2}^4(m_{2,4} - 3)\Big).\end{aligned} \tag{216}$$

Appendix F. Steady state analysis

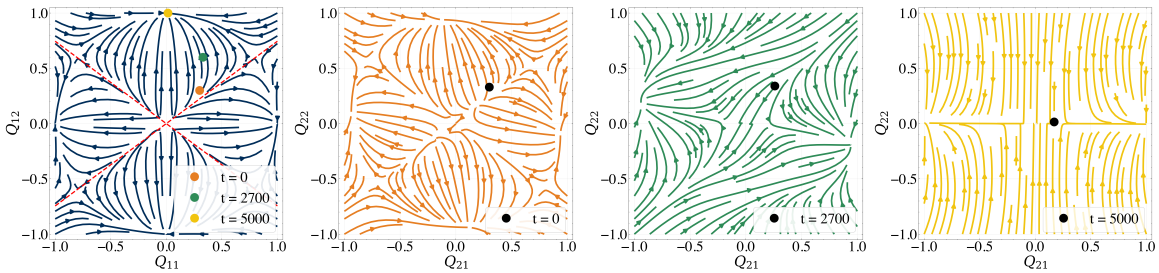
We derive insights from the ODEs describing the learning dynamics. Specifically, steady-state analysis of the limiting ODEs reveals a nontrivial interplay between the learning rate τ, initialization, and the higher-order moment structure of the components. By analyzing the equilibrium $\frac{d\boldsymbol{Q}}{dt} = 0$ in (8), we give general conditions for learnability and competition. In the rest of this section, we first study the learnability boundary, then show a staircase behavior in the total number of learned components, derive a condition (a boundary) for competition to occur, and finally, we demonstrate the instability induced by competition.

F.1. Learnability boundary

We first consider the decoupled regime, where $\boldsymbol{Q}_t$ remains nearly diagonal (up to some permutations and negligible off-diagonal entries). Consequently, for the steady state of the overlap matrix denoted by $\boldsymbol{Q}_s$, the equilibrium condition for (8) reduces to

$$\tau\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{Q}_s^2) - \frac{\tau^2}{2}\mathrm{diag}(\mathbb{E}_{\boldsymbol{c_t},\boldsymbol{e}}\left[f^2(\boldsymbol{v})\right])\boldsymbol{Q}_s = 0. \tag{217}$$

Equation (217) holds for any nonlinearity f, under our assumptions given in Theorem 1. However, since the boundary depends explicitly on f, a closed-form condition in terms of the learning rate and the higher-order moments is intractable. To obtain an explicit



(a) (b) (c) (d)

Figure 7: Phase portraits of the coupled 2×2 system in Figure 1(b). First row dynamics $(Q_{1,1}, Q_{1,2})$ evolve independently in (a), with markers representing specific timestamps: $t = 0$ (orange), $t = 2700$ (green), and $t = 5000$ (yellow), and the dashed red line represents the theoretically predicted competition boundary in (236). In (b), (c), and (d), we plot the second row dynamics $(Q_{2,1}, Q_{2,2})$ at the corresponding times, with black markers indicating the ODE-predicted state. As the first row evolves, it progressively tilts the stability landscape of the second row. Vector field colors in (b–d) match the first-row temporal states for visual alignment.

criterion, we revisit the cubic $f(x) = \pm x^3$; which reduces the dependence to finitely many moments, yielding an exact learnability boundary.

For the remainder of this analysis, we adopt $f(x) = -x^3$ as an example case. For the cubic branch, the steady-state condition $\dot{\boldsymbol{Q}}_t = 0$ in (213) yields a seventh-order polynomial in $Q_{j,j}$. By substituting $q = Q_{j,j}^2$ for the diagonal entries—consistent with the first regime—the dynamics reduce to the form $\dot{Q}_{j,j} = Q_{j,j} P(q)$, where $P(q) = \xi q^3 + \eta q^2 + \zeta q + \varpi$ is a cubic polynomial with $\varpi = 15$. A non-trivial steady state, representing the learning phase where $\boldsymbol{Q} \neq 0$, exists if and only if $P(q) = 0$ admits at least one positive real root.

For a cubic polynomial $P(q) = \xi q^3 + \eta q^2 + \zeta q + \varpi = 0$ where $\varpi > 0$, to have at least one positive root, we have to consider two cases : $\xi > 0$ vs. $\xi < 0$.

Case1 ; $\xi > 0$ Since $P(0) = \varpi > 0$ and $\lim_{q\to\infty} P(q) = \infty$, the curve begins positive and ends positive. For a positive real root to exist, the local minimum must dip below the q-axis. The stationary points are found where $P'(q) = 0$:

$$3\xi q^2 + 2\eta q + \zeta = 0 \implies q_{\min} = \frac{-\eta + \sqrt{\eta^2 - 3\xi\zeta}}{3\xi} \ . \tag{218}$$

(Requires $\eta^2 - 3\xi\zeta > 0$ for real turning points).

For roots to exist, the value at the minimum must be non-positive:

$$P(q_{\min}) \leq 0 \ . \tag{219}$$

Substituting $q_{\min}$ into $P(q)$ yields the condition that the cubic discriminant Δ must be non-negative.

$$\Delta = \eta^2\zeta^2 - 4\xi\zeta^3 - 4\eta^3\varpi - 27\xi^2\varpi^2 + 18\xi\eta\zeta\varpi \geq 0 \ . \tag{220}$$

We also require that this real positive solution lies between our physical meaningful space. We analyze this by transforming the system into the reciprocal function $y = 1/q$.

Let $R(y) = q^3 P(1/q)$. The coefficients reverse order:

$$R(y) = \varpi y^3 + \zeta y^2 + \eta y + \xi \ . \tag{221}$$

The slope of the reciprocal function is

$$R'(y) = 3\varpi y^2 + 2\zeta y + \eta, \tag{222}$$

$$R'(1) = 3\varpi + 2\zeta + \eta \ . \tag{223}$$

The condition $3\varpi + 2\zeta + \eta < 0$ implies $R'(1) < 0$. Since $\varpi > 0$, $R(y) \to +\infty$ as $y \to \infty$. A negative slope at $y = 1$ indicates that the local dip (and thus the root y^*) must occur at a value greater than 1:

$$y^* > 1 \implies \frac{1}{q^*} > 1 \implies q^* < 1 \ . \tag{224}$$

Thus, the condition defines the region where the positive root is confined to the interval $(0, 1)$.

Case 2: $\xi < 0$ In this regime, since $P(0) = \varpi > 0$ and $\lim_{q\to\infty} P(q) = -\infty$, the Intermediate Value Theorem guarantees the existence of at least one positive real root $q^* > 0$.

To ensure this root lies within the meaningful interval $q^* \in (0, 1)$, we examine the reciprocal polynomial. Since $R(y) \to \infty$ as $y \to \infty$, the condition $R(1) < 0$ is sufficient to ensure that the root y^* (where $R(y^*) = 0$) satisfies $y^* > 1$, which implies $q^* < 1$.

The condition $R(1) < 0$ translates to:

$$\varpi + \zeta + \eta + \xi < 0 \tag{225}$$

However, **Case 2 is physically inadmissible in our context**. Given that $\xi+\eta+\zeta+\varpi = (m_6 - 15) + 15 < 0$, it follows that $m_6 < 0$. Since the sixth moment m_6 represents an even power expectation, it is non-negative by definition ($m_6 \geq 0$). This contradiction eliminates Case 2, leaving Case 1 as the sole determinant of our learnability boundary.

Finally, we arrived at the following learnability boundary. Define

$$\xi := (m_6 - 15) - 15(m_4 - 3), \quad \eta := (m_4 - 3)\left(15 - \frac{2}{\tau}\right), \quad \zeta := \frac{2(m_4 - 3)}{\tau}, \quad \varpi := 15.$$

Then, *learning occurs if*

$$\eta^2\zeta^2 - 4\xi\zeta^3 - 4\eta^3\varpi - 27\xi^2\varpi^2 + 18\xi\eta\zeta\varpi > 0 \qquad \text{and} \qquad 3\varpi + 2\zeta + \eta < 0. \tag{226}$$

The simultaneous satisfaction of these inequalities defines *the learnable region*, a subset of the (τ, m_4, m_6) space. To isolate the role of τ, Figure 2(a) illustrates cross-sections in the (m_4, m_6) plane for various values of τ. As τ increases, the learnable region becomes increasingly constrained, persisting only for a narrow set of moment combinations. Conversely, as $\tau \to 0$, the constraint relaxes and the stability boundary approaches a vertical asymptote $m_4 = 3$, so that the boundary is controlled solely by the sign of the fourth cumulant, $\kappa_4 = m_4 - 3$, (alternative characterization of non-Gaussianity to raw moments [52, 53]).

F.2. Staircase behavior

Figure 2(b) shows that, as the component moment pairs (m_4, m_6) cross the learnability boundaries, the number of recoverable components changes in discrete steps with the learning rate τ. The thresholds in Figure 2(b) match the boundary crossings in Figure 2(a), confirming that the mean-field theory captures the phase transitions. Competition shifts the staircase to smaller values of τ: although the discrete structure persists, stable recovery requires more conservative learning rates. Thus, learning more components requires smaller τ, with the precise thresholds determined jointly by the data moments and the initialization regime.

F.3. Competition boundary

To analytically characterize the precise conditions for the competition regime, we return to the example case of $p = 2$. We begin by analyzing the dynamics of the first row of $\boldsymbol{Q} \in \mathbb{R}^{2\times 2}$, namely $Q_{1,1}$ and $Q_{1,2}$, as these govern the evolution of the second row ($Q_{2,1}$ and $Q_{2,2}$). The first row satisfies the ODE system:

$$\begin{aligned}\frac{dQ_{1,1}}{dt} &= \tau\Big(\langle c_1\gamma_1\rangle - Q_{1,1}\langle\gamma_1'\rangle\Big) - \frac{\tau^2}{2}Q_{1,1}\langle\gamma_1^2\rangle \\ &\quad - \tau Q_{1,1}\Big(Q_{1,1}(\langle c_1\gamma_1\rangle - Q_{1,1}\langle\gamma_1^{'}\rangle) + Q_{1,2}(\langle c_2\gamma_1\rangle - Q_{1,2}\langle\gamma_1^{'}\rangle)\Big),\end{aligned} \tag{227}$$

$$\begin{aligned}\frac{dQ_{1,2}}{dt} &= \tau\Big(\langle c_2\gamma_1\rangle - Q_{1,2}\langle\gamma_1'\rangle\Big) - \frac{\tau^2}{2}Q_{1,2}\langle\gamma_1^2\rangle \\ &\quad - \tau Q_{1,2}\Big(Q_{1,1}(\langle c_1\gamma_1\rangle - Q_{1,1}\langle\gamma_1^{'}\rangle) + Q_{1,2}(\langle c_2\gamma_1\rangle - Q_{1,2}\langle\gamma_1^{'}\rangle)\Big).\end{aligned} \tag{228}$$

To understand the boundary between the basins of attraction for the two components, we analyze the relative growth rates of the order parameters $Q_{1,1}$ and $Q_{1,2}$. Dividing the update equations for $\dot{Q}_{1,1}$ and $\dot{Q}_{1,2}$ by $Q_{1,1}$ and $Q_{1,2}$ respectively yields the logarithmic derivatives. Physically, the term $\frac{\dot{Q}}{Q}$ represents the relative growth rate of the component, rather than its absolute speed:

$$\frac{d}{dt}\ln(Q_{1,i}) = \frac{\dot{Q}_{1,i}}{Q_{1,i}}. \tag{229}$$

This normalization allows us to compare the competitive advantage of one component over the other, independent of their current magnitudes. To find the separatrix, we examine the evolution of the ratio between the two components. We compute the difference between their relative growth rates:

$$\frac{d}{dt}\ln\left(\frac{Q_{1,2}}{Q_{1,1}}\right) = \frac{\dot{Q}_{1,2}}{Q_{1,2}} - \frac{\dot{Q}_{1,1}}{Q_{1,1}}. \tag{230}$$

Substituting the explicit expressions for $\frac{\dot{Q}_{1,1}}{Q_{1,1}}$ and $\frac{\dot{Q}_{1,2}}{Q_{1,2}}$:

$$\frac{\dot{Q}_{1,1}}{Q_{1,1}} = \tau\Big(\frac{\langle c_1\gamma_1\rangle}{Q_{1,1}} - \langle\gamma_1'\rangle\Big) - \tau\Big(Q_{1,1}(\langle c_1\gamma_1\rangle - Q_{1,1}\langle\gamma_1'\rangle) + Q_{1,2}(\langle c_2\gamma_1\rangle - Q_{1,2}\langle\gamma_1'\rangle)\Big) - \frac{\tau^2}{2}\langle\gamma_1^2\rangle, \tag{231}$$

$$\frac{\dot{Q}_{1,2}}{Q_{1,2}} = \tau\Big(\frac{\langle c_2\gamma_1\rangle}{Q_{1,2}} - \langle\gamma_1'\rangle\Big) - \tau\Big(Q_{1,1}(\langle c_1\gamma_1\rangle - Q_{1,1}\langle\gamma_1'\rangle) + Q_{1,2}(\langle c_2\gamma_1\rangle - Q_{1,2}\langle\gamma_1'\rangle)\Big) - \frac{\tau^2}{2}\langle\gamma_1^2\rangle. \tag{232}$$

Subtracting these two equations eliminates the second and the third term, yielding

$$\frac{\dot{Q}_{1,2}}{Q_{1,2}} - \frac{\dot{Q}_{1,1}}{Q_{1,1}} = \tau\left(\frac{\langle c_2\gamma_1\rangle}{Q_{1,2}} - \frac{\langle c_1\gamma_1\rangle}{Q_{1,1}}\right). \tag{233}$$

The separatrix is defined as the boundary where the ratio between the components is stationary ($\frac{d}{dt}(Q_{1,2}/Q_{1,1}) = 0$). Setting the difference to zero yields

$$\tau\left(\frac{\langle c_2\gamma_1\rangle}{Q_{1,2}} - \frac{\langle c_1\gamma_1\rangle}{Q_{1,1}}\right) = 0 \implies \frac{\langle c_2\gamma_1\rangle}{Q_{1,2}} = \frac{\langle c_1\gamma_1\rangle}{Q_{1,1}}. \tag{234}$$

We can express this boundary in terms of the nonlinearity $f(\cdot)$, by substituting the definition of γ_i for clarity, as

$$\frac{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[c_1 f(\upsilon_1)]}{Q_{1,1}} = \frac{\mathbb{E}_{\boldsymbol{c},\boldsymbol{e}}[c_2 f(\upsilon_1)]}{Q_{1,2}}. \tag{235}$$

This equation remains valid for an arbitrary nonlinearity. For $f(x) = \pm x^3$, once the learnability conditions hold, the learned component is selected by initialization through a boundary in the $(Q_{1,1}, Q_{1,2})$ phase plane. Specifically, Equation (235) yields the theoretical competition boundary:

$$Q_{1,2} = \pm\sqrt{\frac{(m_{1,4}-3)}{(m_{2,4}-3)}}\,Q_{1,1}, \tag{236}$$

where $m_{j,4}$ denotes the fourth moment of component c_j. This relation defines an implicit cross-shaped separatrix, which can be observed in Figure 7(a) as the red dashed lines, partitioning the phase space into distinct basins of attraction. While the general competition condition holds for complex nonlinearities like $\tanh(x)$, the contribution of all higher-order moments of c_j yields more intricate boundaries, whose analysis lies beyond the scope of this paper.

F.4. Instability induced by competition

The degree of coupling in the initialization governs both the efficiency and the permissible learning rate of the recovery process. In Figure 2(b), we already observed that coupled

initialization shifts the staircase curve toward smaller values of τ relative to the decoupled case, indicating that the competition regime necessitates more conservative learning rates to recover all components. Figure 2(c) demonstrates this effect directly. As the ratio off-diagonal to diagonal entries in the initial overlap matrix increases (transitioning from decoupled to fully coupled regime) the maximum permissible learning rate $\tau_{\max}$ required for full recovery decreases significantly. This required reduction in learning rate, along with component competition, results in a sharp nonlinear increase in convergence time. The close agreement between our theoretical predictions and these empirical trajectories confirms that the high-dimensional dynamics are fundamentally constrained by an initialization-induced trade-off. Stronger initial coupling necessitates smaller learning rates, which in turn prolongs the time required for all components to converge.

F.5. Further analysis of the decoupled regime

In the decoupled initialization regime with $p = 2$ and cubic nonlinearity, we further analyze the stability of the fixed points using Jacobian analysis, numerically. In this regime, Jacobian matrix is diagonal with all off-diagonal entries equal to zero. Thus, the eigenvalues are simply the partial derivatives of the functions governing the ODEs, evaluated at the fixed points. Choosing a representative entry $Q_{1,1}$ we evaluate the partial derivative of the governing function ($h_{1,1} = \dot{Q}_{1,1}$) for the first component:

$$\frac{1}{\tau}\frac{\partial h_{1,1}}{\partial Q_{1,1}} = -3Q_{1,1}^2(m_{1,4}-3) - \frac{\tau}{2}\left(W(Q_{1,1},0) + Q_{1,1}\left.\frac{\partial W(Q_{1,1},Q_{1,2})}{\partial Q_{1,1}}\right|_{Q_{1,2}=0}\right) + 5Q_{1,1}^4(m_{1,4}-3)\ , \tag{237}$$

where

$$\left.\frac{\partial W(Q_{1,1},Q_{1,2})}{\partial Q_{1,1}}\right|_{Q_{1,2}=0} = -90Q_{1,1}^5 m_{1,4} + 6Q_{1,1}^5 m_{1,6} + 180Q_{1,1}^5 + 60Q_{1,1}^3 m_{1,4} - 180Q_{1,1}^3, \tag{238}$$

and

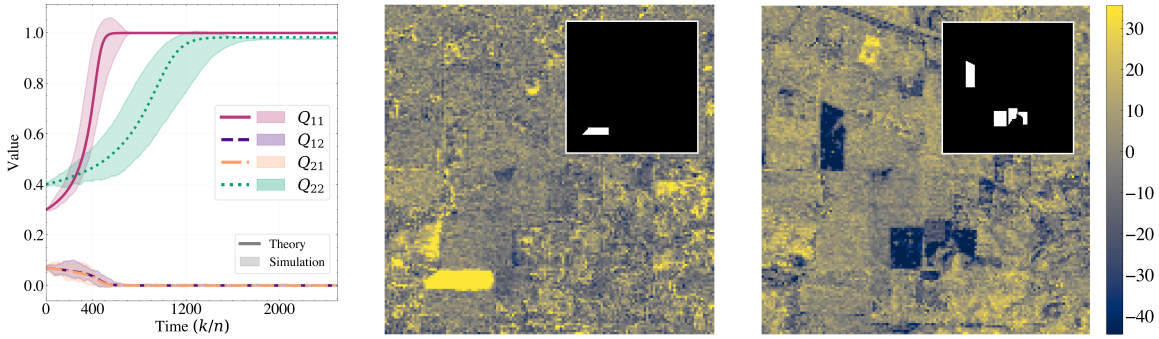
$$W(Q_{1,1},0) = Q_{1,1}^6 m_{1,6} + 15(1-Q_{1,1}^2)^3 + 15Q_{1,1}^4(1-Q_{1,1}^2)m_{1,4} + 45(1-Q_{1,1}^2)^2(Q_{1,1}^2). \tag{239}$$

Finally, for the partial derivative expression we arrive at

$$\frac{1}{\tau}\frac{\partial h_{1,1}}{\partial Q_{1,1}} = Q_{1,1}^6\frac{7\tau}{2}(-m_{1,6}+15m_{1,4}-30) + Q_{1,1}^4(m_{1,4}-3)(-\frac{75\tau}{2}+5) - 3Q_{1,1}^2(m_{1,4}-3) - \frac{15\tau}{2}\ . \tag{240}$$

We numerically determine the fixed points by identifying the roots where the evolution equations vanish. To assess stability, we evaluate the sign of the partial derivative $\frac{\partial h_{1,1}}{\partial Q_{1,1}}$ at these points. Consequently, a fixed point is stable if

$$\left.\frac{\partial h_{1,1}}{\partial Q_{1,1}}\right|_{Q_{1,1}^*} < 0\ . \tag{241}$$



(a) Cosine similarity dynamics.

(b) Learned mask — Class 1.

(c) Learned mask — Class 2.

Figure 8: Multi-component ICA on the Indian Pines dataset. In (a) ODE predictions accurately track the empirical dynamics for $\tau = 0.01$, $\beta_1 = 1$, and $\beta_2 = 0.6$. The shaded regions indicate two standard deviations across 20 runs. In (b), learned spatial masks align with the underlying true classes 1 and 2; ground-truth annotations are shown as black&white insets (top right).

Appendix G. Multi-component online ICA for hyperspectral remote sensing

To validate the predictive power of our theory in a practical setting, we conduct an experiment on hyperspectral remote sensing [54]. Specifically, we use the Indian Pines hyperspectral dataset [55], where each pixel is represented by a 200-dimensional spectral vector. Given this dimensionality, the system operates near the high-dimensional regime, allowing for a robust test of our theory. Following the pre-processing steps detailed in section G.1, we extract canonical signal profiles by averaging the ground-truth masked vectors within each class and these spectral signatures serve as the latent components that we then mix for the ICA problem. As illustrated in Figure 8(a), the ODE predictions accurately characterize the learning dynamics. Figure 8(b) and 8(c) provide a comparison between the learned masks and the corresponding ground-truth masks, obtained via the post-processing steps described in Section G.2. The empirical trajectories remain well within the theoretically predicted regions, affirming our theoretical results.

G.1. Pre-processing

We utilize the Indian Pines hyperspectral dataset, denoted as a tensor $\mathcal{Y} \in \mathbb{R}^{H \times W \times B}$, where $B = 224$. We use the *corrected* version of the dataset, in which water absorption and high-noise spectral bands have been removed to ensure signal quality, resulting in an spectral dimension of $n = 200$. The data are flattened into a matrix $\boldsymbol{Y}_{\text{raw}} \in \mathbb{R}^{N \times n}$, where N is the total number of pixels, and centered by subtracting the mean, yielding $\boldsymbol{Y}_c$. To whiten the data, we compute the empirical covariance matrix $\boldsymbol{\Sigma} = \frac{1}{N}\boldsymbol{Y}_{\text{c}}^\top \boldsymbol{Y}_{\text{c}}$ and perform Singular Value Decomposition (SVD):

$$\boldsymbol{\Sigma} = \boldsymbol{V}_s \boldsymbol{S} \boldsymbol{V}_s^\top. \tag{242}$$

where $\boldsymbol{V}_s$ is an orthogonal matrix containing the singular vectors (representing the principal components of the data), and $\boldsymbol{S}$ is a diagonal matrix of the corresponding singular values (representing the variance along each component). The whitening matrix is constructed as $\boldsymbol{W} = \mathbf{S}^{-1/2}\boldsymbol{V}_s^\top$. The whitened dataset is obtained via the linear transformation $\tilde{\boldsymbol{Y}} = \mathbf{Y}_c\boldsymbol{W}^\top$, ensuring that the covariance of $\tilde{\boldsymbol{Y}}$ is the identity matrix $\boldsymbol{I}$.

We derive a set of ground truth vectors $\boldsymbol{u}_i$ from the labeled hyperspectral classes. Let $\mathcal{C}_i$ denote the set of pixel indices belonging to class i, and $\tilde{\boldsymbol{y}}_l$ denote $l-th$ row of the matrix $\tilde{\boldsymbol{Y}}$. We compute the class centroids in the whitened space:

$$\boldsymbol{u}_i = \frac{1}{|\mathcal{C}_i|}\sum_{l\in\mathcal{C}_i}\tilde{\mathbf{y}}_l. \tag{243}$$

Since raw class centroids may be correlated, we enforce orthogonality to satisfy the ICA independence assumption. For this, we construct the matrix $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p]$ and apply QR decomposition. The resulting orthogonal components are then re-scaled to maintain a norm of $\sqrt{n}$. Having established the independent ground truth components, we construct the data as described in (1). For $p = 2$, we select Classes 6 and 13, corresponding to Grass Trees and Wheat in the Indian Pines scene, as our two ground-truth components. In the setting of Example 1, we set $\beta_1 = 1$ for Class 13 and $\beta_2 = 0.6$ for Class 6, with learning rate $\tau = 0.01$.

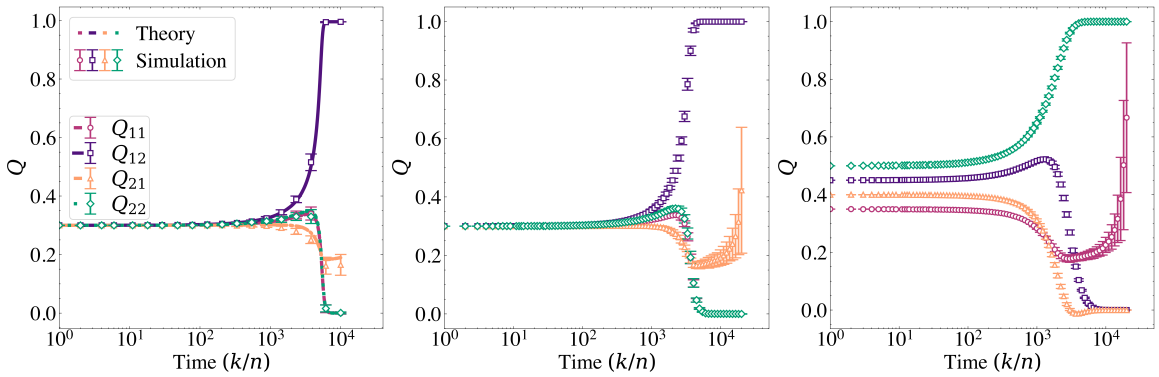
G.2. Post-processing

After obtaining the learned components which we reshape back to the spatial grid to yield activation maps $\mathbf{A} \in \mathbb{R}^{H\times W\times 2}$. Since online ICA recovers the independent components only up to permutation and sign, we align each learned component with a ground-truth class by computing a matching score against the centroid-derived targets. To enable qualitative assessment of recovery, the activation maps are plotted using a globally shared color scale. The corresponding ground-truth binary mask is then displayed as a localized inset overlaid directly on the learned activation map to facilitate direct visual comparison.

Appendix H. Competition in other orthogonalization schemes and nonlinearities

The identified regimes of decoupling and competition are not unique to specific configurations; rather, they are observable across other orthogonalization schemes and nonlinearities. In this section, we provide supplementary plots of the learning dynamics to demonstrate that our results are not restricted to Gram-Schmidt orthogonalization or to cubic nonlinearities.

Orthogonality may be enforced through several methodologies: sequential schemes such as QR decomposition (utilizing Householder or Gram-Schmidt), symmetric approaches via Löwdin (polar) orthogonalization, Cayley-transform updates, or Riemannian retractions on the Stiefel manifold.[56–59]. We simulate two orthogonalization schemes, specifically QR with Householder, which is a sequential method, also Löwdin orthogonalization, where estimate vectors are orthogonalized concurrently rather than sequentially as in the QR approach. In Figure 9(b), and Figure 9(c) we can clearly observe the competition between



(a) $f(x) = \tanh(x)$ with GS (b) QR with Householder (sequential) (c) Löwdin via polar (symmetric)

Figure 9: Competition is also visible across other nonlinearities and orthogonalization schemes. In a) our limiting ODE and the simulations are plotted with $f(x) = \tanh(x)$ with Gram-Schmidt process, with $\tau = 0.01$. In b), QR via Householder transformations, with $\tau = 0.001, Q_{0,i,j} = 0.3$ In c) Löwdin orthogonalization was used with $\tau = 0.001$ and $Q_{0,1,1}, Q_{0,1,2}, Q_{0,2,1}, Q_{0,2,2} = 0.35, 0.45, 0.4, 0.5$. In all of the above simulations $n = 1000, \beta_1 = 0.2, \beta_2 = 1$ were used, with 10 Monte Carlo averages, 2 standard deviation error bars for simulations.

components, as one of the estimates is first forced to unlearn, then eventually wins the competition.

Next, given that our Theorem 1 holds for any nonlinearity satisfying the underlying assumptions, we can evaluate other score functions and directly compare theoretical predictions with empirical simulations. Using Equation (8), we analyze the case where $f(x) = \tanh(x)$. Under the competition regime initializations, we obtain the results shown in Figure 9(a). We observe strong agreement between our theoretical ODE trajectories and empirical Monte Carlo simulations, further validating that our theory and regime identification hold for arbitrary nonlinearities.