ON THE IMPORTANCE OF CONTRASTIVE LOSS IN MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, contrastive learning approaches (e.g., CLIP (Radford et al., 2021)) have received huge success in multimodal learning, where the model tries to minimize the distance between the representations of different views (e.g., image and its caption) of the same data point, while keep the representations of different data points away from each other. However, from a theoretical perspective, it is unclear how contrastive learning can learn to align the representations from different views efficiently, especially in cases where the data is not isotropic. In this work, we analyze the training dynamics of a simple multimodal contrastive learning model, and show that contrastive pairs are important for the model to efficiently balance the learned representations. In particular, we reveal a stage-wise behavior of the learning process: In the first stage, the model aligns the feature representations using positive pairs and the condition number grows in this stage. Then, in the second stage, the model reduces the condition number of the learned representations using negative pairs.

1 INTRODUCTION

One of the exceptional abilities of humans is to associate data from different modalities (such as texts and images) together. For example, when we hear the words "white dog", we can immediately align it with the image of a dog with the white color. When we hear the loud sound of the engine, we can imagine an expensive sport car passing nearby.

Recently, in machine learning, multimodal learning methods – Training the model to be able to align the data from different modules has become an increasingly popular research direction, especially in deep learning (He & Peng (2017); Stroud et al. (2020); Radford et al. (2021); Ramesh et al. (2021); Xu et al. (2021); Jia et al. (2021); Wang et al. (2022b)). Among them, the recent work CLIP (Radford et al. (2021)) shows remarkable quality results on aligning the features of text and images. The *contrastive learning* based method CLIP empirically outperforms many existing non-contrastive approaches (Grill et al. (2020); Chen & He (2021); He et al. (2020)). The major difference between contrastive approach and other approaches is that contrastive loss not only requires the learned representation from the same pair of data (i.e. positive pairs) to be positively aligned, but it also requires the data from different pairs (i.e. negative pairs) to be as negatively aligned as possible. In the paper, the authors also identify contrastive loss as the most critical part to the success of CLIP.

Empirical, CLIP has shown superior performance in multimodal learning on aligning features from the two modules. For example, the trained model can align the representation of words "black cat" with the representation of an image of a cat with black color, while aligning the words "yellow cat" with an image of a cat with yellow color instead of black one. However, from a theoretical perspective, the most fundamental questions are still largely open: In particular, how do contrastive pairs help in this new multimodal learning approach? How can the **non-convex** contrastive loss be efficiently minimized to learn features from both modules?

Unlike the prior theoretical works on contrastive learning which mostly focus on extracting features from one module (e.g., Arora et al. (2019); Jing et al. (2022); Pokle et al. (2022); Tian et al. (2021); Wen & Li (2021)), one main technical challenge of analyzing contrastive learning in a multimodal setting is how the model can be trained to align the feature representations $\boldsymbol{f_A}, \boldsymbol{f_B}$ from modules $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively. Due to the existence of negative pairs that emphasize on negative correlations

of $\boldsymbol{f_A}$ and $\boldsymbol{f_B}$, it is unclear that the model is still incentive to align the features from different modalities.

In this paper we make preliminary theoretical steps towards answering the fundamental theoretical questions of the importance of contrastive loss in multimodal learning. We assume the data from the two modules are of form $\boldsymbol{x_A} = \boldsymbol{A}\boldsymbol{z_A} + \boldsymbol{\xi_A}$ and $\boldsymbol{x_B} = \boldsymbol{B}\boldsymbol{z_B} + \boldsymbol{\xi_B}$, respectively, where $\boldsymbol{z_A}, \boldsymbol{z_B}$ are considered as the hidden signals, $\boldsymbol{A}, \boldsymbol{B}$ linear transformations from the signal to the observation and $\boldsymbol{\xi_A}, \boldsymbol{\xi_B}$ the noises. Similar linear models have also been used in previous works (Tian et al. (2021); Wen & Li (2021)) in the context of single-modal learning ($\boldsymbol{A} = \boldsymbol{B}$). The positive pair of the data shares the same signal $\boldsymbol{z_A} = \boldsymbol{z_B}$, and has different noises $\boldsymbol{\xi_A}, \boldsymbol{\xi_B}$ and transformations $\boldsymbol{A}, \boldsymbol{B}$. The goal is to learn features $\boldsymbol{f_A}, \boldsymbol{f_B}$ that align data from the two modules.

Under this setting, we make the following contributions:

1. We consider the challenging (but more practical) setting where the features in $\boldsymbol{A}$ and $\boldsymbol{B}$ are inhomogeneous, that is, the condition number of A and B can be $\omega(1)$. Prior works (Jing et al. (2022); Tian et al. (2021); Wen & Li (2021)) only consider cases where $\boldsymbol{A}$ and $\boldsymbol{B}$ are exactly column orthonormal matrices even in the simpler setting of single modality ($\boldsymbol{A} = \boldsymbol{B}$).
2. We consider feature learners $\boldsymbol{f_A}, \boldsymbol{f_B}$ with normalization, meaning that $\boldsymbol{f_A}, \boldsymbol{f_B}$ are always normalized to have (expected) norm one during training. Output normalization plays a critical rule in the practical success of contrastive learning and is also employed in CLIP, but it is rarely considered in theory due to the additional complexity of the division by norm.
3. We analyze the learning process of stochastic gradient descent from random initialization. We prove that contrastive learning converges efficiently to a nearly optimal solution, which indeed aligns the feature representation $\boldsymbol{f_A}$ and $\boldsymbol{f_B}$.
4. We also demonstrate the importance of negative pairs by comparing with training only over the positive pairs: We prove that although the latter can also learn to align $\boldsymbol{f_A}$ and $\boldsymbol{f_B}$, the features learned by contrastive learning with negative pairs is much more *uniform*, meaning that $\boldsymbol{f_A}, \boldsymbol{f_B}$ can recover all the singular vectors of $\boldsymbol{A}$ and $\boldsymbol{B}$. On the other hand, without negative pairs the learned representation is close to a rank one solution, meaning that $\boldsymbol{f_A}, \boldsymbol{f_B}$ will only focus on the top singular direction of $\boldsymbol{A}$ and $\boldsymbol{B}$.
5. We also perform simulations to further support our theory.

2 Related Works

Multimodal learning Despite the empirical success of multimodal learning, there are very few theoretical results on this topic. The one most related to ours is Huang et al. (2021), in which the authors show that, in certain cases, multimodal methods can provably perform better than single-modal models. However, the authors consider neither contrastive pairs nor the training dynamics.

Contrastive/Non-contrastive learning theory Another much richer line of research is about contrastive and non-contrastive methods in the context of single-modal self-supervised learning. Starting from Arora et al. (2019), many recent works have provided various explanations on why the representations learned with contrastive learning are useful in downstream tasks (Chuang et al. (2020); Tosh et al. (2021); Nozawa & Sato (2021); Wang et al. (2022a); HaoChen et al. (2021); Lee et al. (2021); Wang & Isola (2020)). These works mostly focus on the generalization aspect of the problem, and do not consider training. Among them, Wang & Isola (2020) also study the problem using the notions of alignment and uniformity, and demonstrate that balanced representations benefit downstream tasks. However, they do not provide guarantees on training. Another related line of research is about non-contrastive learning, where the necessity of negative examples is questioned. In this line of research, the optimization problem does get considered as non-contrastive losses have trivial collapsed solutions. Tian et al. (2021) show that, under certain conditions, non-contrastive learning methods can learn non-collapsed solutions. Jing et al. (2022) show that, even with negative examples, contrastive learning can still suffer from another type of collapse, where the learned representation only expand a low-dimensional subspace of the embedding space. In Pokle et al. (2022), the authors show that non-contrastive losses have many non-collapsed bad minima that the training

algorithm does not avoid. Another related work that takes optimization into consideration is Wen & Li (2021), in which the authors analyze the training dynamics of contrastive learning and show that, with random augmentation, neural networks can learn features that are suppressed by noises when no augmentations are used. Despite these works do consider the optimization problem, they only consider the case where the features are uniform, and only Wen & Li (2021) considers output normalization. We compare our result with the most relevant works in the next paragraph.

Comparison with Jing et al. (2022); Tian et al. (2021); Wen & Li (2021) The dimensional collapse problem reported in Jing et al. (2022) is not a real issue in our setting, since, in our case, the best the model can do is to recover the latent vector $\boldsymbol{z}$, up to some rotation. As a result, it is natural for the learned representations to span only a low-dimensional subspace of the embedding space $\mathbb{R}^m$. Here, the point of choosing $m \gg d$ is to make the optimization dynamics more regular, which is a common strategy in the study of over-parameterized neural networks. The main difference between our work and the analysis in Tian et al. (2021) and Wen & Li (2021) is we do not assume the inputs are isotropic. In our setting, the condition number can be as large as $\Theta(\log d)$. When the condition number is 1, one can imagine that, thanks to the symmetry, all directions will be learned simultaneously. Therefore, we do not need negative examples to prevent collapse (Tian et al. (2021)) or the negative examples do not play an important role in analysis (Wen & Li (2021)). On the other hand, when the condition number is larger than 1, we do need to use the negative examples to shrink the condition number, which corresponds to the second stage of our analysis.

3 Problem Setup

We consider a linear data-generating model. Similar linear models have also been used in previous works, in the context of single-modal learning (e.g., Lee et al. (2021); Wen & Li (2021)). Formally, we assume that the contrastive pairs $(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^-)$ are constructed as

$$\boldsymbol{x}_{\boldsymbol{A}}^+ = \boldsymbol{A}\boldsymbol{z}^+ + \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+, \quad \boldsymbol{x}_{\boldsymbol{B}}^- = \boldsymbol{B}\boldsymbol{z}^- + \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{B}}^-, \tag{1}$$

where $\boldsymbol{z}^\pm$, $\boldsymbol{\xi}_{\boldsymbol{A}}^-$, $\boldsymbol{\xi}_{\boldsymbol{B}}^-$ are independent random variables following the uniforms distributions over $\{\pm 1/\sqrt{d}\}^r$, $\{\pm 1/\sqrt{d}\}^{d-r}$ and $\{\pm 1/\sqrt{d}\}^{d-r}$, respectively, and $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times r}$, $\boldsymbol{A}_{\boldsymbol{\xi}}, \boldsymbol{B}_{\boldsymbol{\xi}} \in \mathbb{R}^{d\times(d-r)}$ are matrices with $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{B}^\top\boldsymbol{B} = \operatorname{diag}(\boldsymbol{\sigma}^2)$ and $\boldsymbol{A}_{\boldsymbol{\xi}}^\top\boldsymbol{A}_{\boldsymbol{\xi}} = \boldsymbol{B}_{\boldsymbol{\xi}}^\top\boldsymbol{B}_{\boldsymbol{\xi}} = \sigma_\xi^2\boldsymbol{I}_{d-r}$ for some $\boldsymbol{\sigma} \in \mathbb{R}_+^r$ and $\sigma_\xi \in \mathbb{R}_+$. In words, we first sample the latent vectors $\boldsymbol{z}^\pm \in \mathbb{R}^r$ independently, and encode them with $\boldsymbol{A}, \boldsymbol{B}$ to form the signal part of the input. Then, we sample the latent noises $\boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{\xi}_{\boldsymbol{B}}^- \in \mathbb{R}^{d-r}$, and encode them with $\boldsymbol{A}_{\boldsymbol{\xi}}, \boldsymbol{B}_{\boldsymbol{\xi}}$ to form the noise part of the input. Finally, we add the signal and noise parts together to obtain $(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^-)$. To generate a positive pair $(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)$, we use the same latent vector. That is, $\boldsymbol{x}_{\boldsymbol{A}}^+ = \boldsymbol{A}\boldsymbol{z}^+ + \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+$ and $\boldsymbol{x}_{\boldsymbol{B}}^+ = \boldsymbol{B}\boldsymbol{z}^+ + \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{B}}^+$. Note that the latent noises here are still independent. We use $\sigma_{\max}^2$ and $\sigma_{\min}^2$ to denote the maximum and minimum of $\sigma_1^2, \ldots, \sigma_r^2, \sigma_\xi^2$, respectively. We assume that $(\sigma_{\max}^2/\sigma_{\min}^2)\max\left\{1, (d-r)\sigma_\xi^2/(r\sigma_{\min}^2)\right\} \le c\log d$ for some small constant $c > 0$. Note that our results can easily be generalized to setting where the dimensions of $\boldsymbol{x}_{\boldsymbol{A}}$ and $\boldsymbol{x}_{\boldsymbol{B}}$ are not the same, since one can simply pad zeros at the end of each column of $\boldsymbol{A}$ and $\boldsymbol{B}$.

One way of interpreting this model is to view each coordinate of the latent vector $\boldsymbol{z}$ as an indicator for the presence/absence of a certain object, and the corresponding columns in $\boldsymbol{A}$ and $\boldsymbol{B}$ as the visual and word embeddings of this object, respectively.

Now, we describe our learner model. We consider (normalized) linear feature learners. Define

$$\boldsymbol{f}_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}) := \frac{\boldsymbol{W}_{\boldsymbol{A}}^\top\boldsymbol{x}_{\boldsymbol{A}}}{\sqrt{\mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}}\left\|\boldsymbol{W}_{\boldsymbol{A}}^\top\boldsymbol{x}_{\boldsymbol{A}}\right\|^2}}, \quad \boldsymbol{f}_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{B}}) := \frac{\boldsymbol{W}_{\boldsymbol{B}}^\top\boldsymbol{x}_{\boldsymbol{B}}}{\sqrt{\mathbb{E}_{\boldsymbol{x}_{\boldsymbol{B}}}\left\|\boldsymbol{W}_{\boldsymbol{B}}^\top\boldsymbol{x}_{\boldsymbol{A}}\right\|^2}},$$

where $\boldsymbol{W}_{\boldsymbol{A}}, \boldsymbol{W}_{\boldsymbol{B}} \in \mathbb{R}^{d\times m}$ are the trainable parameters. In words, we first map the inputs $(\boldsymbol{x}_{\boldsymbol{A}}, \boldsymbol{x}_{\boldsymbol{B}})$ into the embedding space $\mathbb{R}^m$ using $\boldsymbol{W}_{\boldsymbol{A}}$ and $\boldsymbol{W}_{\boldsymbol{B}}$, and then apply batch normalization to the outputs. By saying the learned representations are aligned, we mean that $\boldsymbol{f}_{\boldsymbol{A}}$ and $\boldsymbol{f}_{\boldsymbol{B}}$ are close for positive pairs and far away from each other for negative pairs. Meanwhile, we say the learned representations are balanced if changing a small fraction of coordinates of $\boldsymbol{z}$ does not change the representation dramatically. See Section 4 for formal definitions.

One can easily verify that, in the population case, we have $\mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}} \left\|\boldsymbol{W}_{\boldsymbol{A}}^\top \boldsymbol{x}_{\boldsymbol{A}}\right\|^2 = \left\|\boldsymbol{W}_{\boldsymbol{A}}^\top \boldsymbol{A}\right\|_F^2 / d + \left\|\boldsymbol{W}_{\boldsymbol{A}}^\top \boldsymbol{A}_{\boldsymbol{\xi}}\right\|_F^2 / d$. For notational simplicity, we define $\boldsymbol{K}_{\boldsymbol{B}} = \boldsymbol{W}_{\boldsymbol{B}}^\top \boldsymbol{B}$, $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} = \boldsymbol{W}_{\boldsymbol{A}}^\top \boldsymbol{A}_{\boldsymbol{\xi}}$, and $\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} = \boldsymbol{W}_{\boldsymbol{B}}^\top \boldsymbol{B}_{\boldsymbol{\xi}}$. These are the matrices that directly map latent vectors to their final representations. We also define $N_{\boldsymbol{A}}^2 = (\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2)/d$ and $N_{\boldsymbol{B}}^2 = (\|\boldsymbol{K}_{\boldsymbol{B}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\|_F^2)/d$. Then our model can be rewritten as[1]

$$\boldsymbol{f}_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}) = \frac{\boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}_{\boldsymbol{A}} + \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}}, \quad \boldsymbol{f}_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{B}}) = \frac{\boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}_{\boldsymbol{B}} + \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{B}}}{N_{\boldsymbol{B}}}. \tag{2}$$

We initialize each entry of $\boldsymbol{W}_{\boldsymbol{A}}$ and $\boldsymbol{W}_{\boldsymbol{B}}$ using iid Gaussian $\mathcal{N}(0, 1/m)$. This scaling ensures the norm of outputs before initialization does not blow up as $m \to \infty$. We train our model using gradient descent over the following contrastive loss $\mathcal{L}$. First, we define[2]

$$S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) = \frac{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{B}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)},$$

$$S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) = \frac{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)},$$

where K is positive constant controlling the strength of negative samples and $\tau_t \in (0, 1]$ is the inverse temperature. Then, we define the contrastive loss as

$$\mathcal{L} := \mathcal{L}_{\boldsymbol{A}} + \mathcal{L}_{\boldsymbol{B}} := -\mathbb{E} \log S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - \mathbb{E} \log S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+). \tag{3}$$

By the non-contrastive loss, we mean

$$\hat{\mathcal{L}} := -\mathbb{E} \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle. \tag{4}$$

Formally, the training algorithm is defined as follows.

Algorithm 3.1 (Training algorithm)**.** *Let $\tilde{\mathcal{L}}$ be $\mathcal{L}$ in the contrastive case and $\hat{\mathcal{L}}$ in the non-contrastive case. At each step, we first sample a batch of positive/negative $\{(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+, \boldsymbol{x}_{\boldsymbol{A}}^-, \boldsymbol{x}_{\boldsymbol{B}}^-)\}_{i=1}^N$, use them to compute the empirical version of $\tilde{\mathcal{L}}$, and update the weight matrices using $\boldsymbol{W}_{\boldsymbol{A}} \leftarrow \boldsymbol{W}_{\boldsymbol{A}} - \tau_t^{-2} \eta \nabla_{\boldsymbol{W}_{\boldsymbol{A}}} \tilde{\mathcal{L}}$ and $\boldsymbol{W}_{\boldsymbol{B}} \leftarrow \boldsymbol{W}_{\boldsymbol{B}} - \tau_t^{-2} \eta \nabla_{\boldsymbol{W}_{\boldsymbol{B}}} \tilde{\mathcal{L}}$.*

In the non-contrastive case, we always use $\tau_t = 1$[3], and we repeat the above process until gradient descent converges to an approximate stationary point. In the contrastive case, we first use a small $\tau_t = 1/\operatorname{poly}(d)$, run the process for $T_1 = \operatorname{poly}(d)$ iterations, switch to $\tau_t = 1$, and run the process for another $T_2 - T_1 = \operatorname{poly}(d)$ iterations.

4 Main Results

In words, our results say that though both contrastive and non-contrastive methods can align the representations, the representations learned by contrastive methods are more balanced. First, we need to define what do "alignment" and "balance" mean here. We still use $\boldsymbol{f}_{\boldsymbol{A}}$ and $\boldsymbol{f}_{\boldsymbol{B}}$ to denote the embeddings, but our definitions here will be architecture-agnostic. After some general discussion, we also discuss these definitions in our specific setting.

Definition 4.1 (Alignment)**.** *We define the alignment score as*

$$\Gamma_{\texttt{Align}} := \frac{1}{2} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^\pm, \boldsymbol{x}_{\boldsymbol{B}}^\pm} \left\{ \mathbb{1}\left\{ \left\|\boldsymbol{f}_{\boldsymbol{A}}^+ - \boldsymbol{f}_{\boldsymbol{B}}^+\right\| < \left\|\boldsymbol{f}_{\boldsymbol{A}}^+ - \boldsymbol{f}_{\boldsymbol{B}}^-\right\| \right\} + \mathbb{1}\left\{ \left\|\boldsymbol{f}_{\boldsymbol{A}}^+ - \boldsymbol{f}_{\boldsymbol{B}}^+\right\| < \left\|\boldsymbol{f}_{\boldsymbol{A}}^- - \boldsymbol{f}_{\boldsymbol{B}}^+\right\| \right\} \right\}.$$

Namely, $\Gamma_{\texttt{Align}}$ is the accuracy of classifying if the latent vectors of the input pair $(\boldsymbol{x}_{\boldsymbol{A}}, \boldsymbol{x}_{\boldsymbol{B}})$ are the same. We say that the learned representations are aligned if $\Gamma_{\texttt{Align}} \approx 1$.

Note that the notion of alignment introduced here is stronger than matching the positive pairs, which can achieved by simply mapping all inputs to one single embedding. In that case, $\Gamma_{\texttt{Align}}$ will be 0 (or 0.5 if we choose to break tie randomly instead of using strict inequality).

[1] See Section D for discussions on the sample complexity.

[2] We use $\boldsymbol{f}_{\boldsymbol{A}}^+$ as a shorthand for $\boldsymbol{f}_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+)$, similarly for other combinations of $\boldsymbol{A}, \boldsymbol{B}$ and $\pm$.

[3] Note that, in the non-contrastive case, changing τ_t only changes the learning rate.

Definition 4.2 (Balance)**.** *We define the balance score as*

$$\Gamma_{\texttt{Balance}} := \frac{\|\boldsymbol{\Sigma}_f\|_F^2}{\|\boldsymbol{\Sigma}\|_2^2} \quad \textit{where} \quad \boldsymbol{\Sigma}_f := \mathop{\mathbb{E}}_{\boldsymbol{x_A}} \left\{ \boldsymbol{f_A} \boldsymbol{f_A}^\top \right\}.$$

Namely, $\Gamma_{\texttt{Balance}}$ *is the stable rank of the covariance matrix of the output embeddings. We say that the learned representations are balanced if* $\Gamma_{\texttt{Balance}} \geq \alpha r$ *for some* $\alpha \approx 1$.

We make some short remarks on Definition 4.2. First, we only use $\boldsymbol{f_A}$ in this definition because if the embeddings are well-aligned, $\boldsymbol{f_A}\boldsymbol{f_A}^\top$ and $\boldsymbol{f_B}\boldsymbol{f_B}^\top$ should be approximately the same. Second, the stable rank is usually used as a surrogate for the actual rank that is less sensitive to small singular values (Rudelson & Vershynin (2007)). To see why this makes sense, note that $\|\boldsymbol{\Sigma}_f\|_F^2 / \|\boldsymbol{\Sigma}\|_2^2 = \sum_{k=1}^m \kappa_k^2 / \max_{k\in[m]} \kappa_k^2$, where $\kappa_1, \ldots, \kappa_m$ are the singular values of $\boldsymbol{\Sigma}_f$. If all nonzero κ_k are the same, then this recover the actual rank. The intuition behind the use of rank is that, the more independent latent variables the model learns, the higher the rank of the representations needs to be.

In Jing et al. (2022), the authors also use rank to measure the degree of "dimensional collapse". Unlike their argument, here we only require $\Gamma_{\texttt{Balance}}$ to be at least αr, instead of αm, for the representations to be called balanced because even if we can recover the underlying latent vectors, the rank is still at most r. Hence, it does not make much sense to expect the learned representations can span the entire embedding space, especially in our setting. Finally, note that a sufficient condition for $\Gamma_{\texttt{Balance}} \geq \alpha r$ is that the ratio of the largest and r-largest singular values is at most $\sqrt{\alpha}$. In other words, after excluding those singular values that should be 0, the condition number is approximately 1.

Why balance representations are important? In our setting, one simple example of aligned but unbalanced representations is $\boldsymbol{W_A}^\top = \mathrm{diag}(\boldsymbol{\nu})\boldsymbol{A}^{-1}$ and $\boldsymbol{W_B}^\top = \mathrm{diag}(\boldsymbol{\nu})\boldsymbol{B}^{-1}$ with $\nu_1 = 1$ and $\nu_2 = \cdots = \nu_r = \exp(-d)$. This model maps inputs whose latent vector is $\boldsymbol{z}$ to $\mathrm{diag}(\boldsymbol{\mu})\boldsymbol{z}$, up to some normalization, for both modules, whence it has $\Gamma_{\texttt{Align}} = 1$. Meanwhile, one can verify that this model has $\Gamma_{\texttt{Balance}}/r \leq 2/r \approx 0$. The problem of this model is that the model overly emphasize on z_1. And for a data with $z_1 = 0$, the output of the model will be very non-robust to noise due to the small $\nu_2, \nu_3, \cdots, \nu_r$, which is not ideal. This model can be made balanced by replacing $\mathrm{diag}(\boldsymbol{\mu})$ with $\boldsymbol{I}_r$, in which case the model directly recover the latent vector $\boldsymbol{z}$. As a result, all changes in $\boldsymbol{z}$ will be reflected in the final representation in a faithful way.

Similar notion has also been studied in Wang & Isola (2020) under the name "uniformity" in the context of self-supervised learning, and they also report that balanced representations lead to better performance in downstream tasks, though, unlike our result, they do not provide guarantees on training. Still, this further suggests that learning balanced representations is a reasonable and important goal.

With these two definitions, we can now state our main results.

Theorem 4.3. *Suppose that the network width* $m = \mathrm{poly}(d)$ *is sufficiently large, the learning rate* $\eta = 1/\mathrm{poly}(d)$ *is sufficiently small, and we generate sufficiently, but still polynomially, many samples at each step to compute the loss*[4]*. Suppose that*

$$\frac{\sigma_{\max}^2}{\sigma_{\min}^2} \max\left\{1, \frac{(d-r)\sigma_\xi^2}{r\sigma_{\min}^2}\right\} \leq c \log d,$$

for some small constant $c > 0$.

(a) ***Non-contrastive loss.*** *There exists a* $\boldsymbol{\sigma} \in \mathbb{R}^r$ *satisfying the above assumptions such that, after* $\mathrm{poly}(d)$ *iterations, Algorithm 3.1 will converge to an approximate stationary point, at which the learned representations are aligned but not balanced, that is,* $\Gamma_{\texttt{Align}} \approx 1$ *but* $\Gamma_{\texttt{Balance}}/r \approx 0$.

(b) ***Contrastive loss.*** *There exists* $\tau_0^2 = 1/\mathrm{poly}(d)$ *and* $T_1 = \mathrm{poly}(d)$ *such that, for any valid* $\boldsymbol{\sigma}$*, Algorithm 3.1 will reach a point after* $\mathrm{poly}(d)$ *iterations at which the learned representations are aligned and balanced, that is,* $\Gamma_{\texttt{Align}} \approx 1$ *and* $\Gamma_{\texttt{Balance}}/r \approx 1$.

[4]To make the proof cleaner, we write it in terms of gradient flow over population loss. See Section D for discussions on the discretization of gradient flow.

We close this section with sufficient conditions for the learned representations to be aligned and balanced in our setting. First, in our specific setting, the most natural way for a model to achieve $\Gamma_{\texttt{Align}} \approx 1$ is to learn $\|\boldsymbol{f_A} - \boldsymbol{f_B}\| \approx \|\boldsymbol{K}_0(\boldsymbol{z_A} - \boldsymbol{z_B})\|$ for some $\boldsymbol{K}_0 \in \mathbb{R}^{m\times r}$ with $\boldsymbol{K}^\top \boldsymbol{K} \succ 0$. In this case, the distance between positive pairs is always close to 0 while the distance between negative pairs being positive. Recall that the output of our model is $\boldsymbol{f_A} = (\boldsymbol{K_A z_A} + \boldsymbol{K_{A,\xi}\xi_A})/N_{\boldsymbol{A}}$ and $\boldsymbol{f_B} = (\boldsymbol{K_B z_B} + \boldsymbol{K_{B,\xi}\xi_B})/N_{\boldsymbol{B}}$. Hence, we have the following sufficient condition.

Lemma 4.4 (Sufficent condition for aligned representations)**.** *If $\|\boldsymbol{K_A}\|_F \gg \|\boldsymbol{K_{A,\xi}}\|_F$, $\|\boldsymbol{K_B}\|_F \gg \|\boldsymbol{K_{B,\xi}}\|_F$, $\boldsymbol{K_A} \approx \boldsymbol{K_B}$, and the condition number of $\boldsymbol{K_A}$ is upper bounded by $\mathrm{poly}(d)$, then the learned representations are aligned.*

The first two conditions imply the signal parts dominate the outputted representations, the third condition gives the existence of $\boldsymbol{K}_0$, and the final condition makes sure the smallest singular value $\boldsymbol{K}_0^\top \boldsymbol{K}_0$ of is still at least $1/\mathrm{poly}(d)$ after normalization.

Now we consider the balance of the learned representations. Suppose that our representations are already aligned, in the sense of Lemma 4.4. Then we have

$$\boldsymbol{\Sigma}_f = \mathop{\mathbb{E}}_{\boldsymbol{x_A}} \left\{ \boldsymbol{f_A} \boldsymbol{f_A}^\top \right\} = \mathop{\mathbb{E}}_{\boldsymbol{z}} \left\{ \frac{\boldsymbol{K_A z z}^\top \boldsymbol{K_A}^\top}{\|\boldsymbol{K_A}\|_F^2 / d} \right\} = \frac{\boldsymbol{K_A K_A}^\top}{\|\boldsymbol{K_A}\|_F^2}.$$

Recall the relationship between $\Gamma_{\texttt{Balance}}$ and the effective condition number of $\boldsymbol{\Sigma}_f$. We have the following sufficient condition.

Lemma 4.5 (Sufficient condition for balanced representations)**.** *Suppose that our model is aligned, in the sense of Lemma 4.4. If the condition number of $\boldsymbol{K_A}^\top \boldsymbol{K_A}$ is close to 1, then the learned representations are balanced.*

Note that, since the columns of $\boldsymbol{A}$ are not orthonormal, even at initialization, the condition number of $\boldsymbol{K_A}^\top \boldsymbol{K_A}$ is not close to 1. In other words, a condition-number-reducing stage is necessary for the model to learn balanced representations.

5 Training Dynamics and Proof Outline

Our proof is based on characterizing the dynamics of gradient descent over the contrastive loss. We first choose a small inverse temperature $\tau_t^2 = 1/\mathrm{poly}(d)$ and run gradient descent for $\mathrm{poly}(d)$ iterations. This stage is called Stage 1. Then, we set $\tau_t^2 = 1$ and run gradient descent for another $\mathrm{poly}(d)$ iterations. This stage is called Stage 2. The use of different τ_t is mainly for technical purpose since this gives cleaner separation between stages. We observe similar stage-wise behavior when using a uniform τ_t^2. See Figure 1 for simulation results. We also report results for the non-contrastive loss there.

5.1 The training dynamics

Instead of tracking the parameters $\boldsymbol{W_A}$ and $\boldsymbol{W_B}$ directly, we will track $\boldsymbol{K_A}, \boldsymbol{K_{A,\xi}}, \boldsymbol{K_B}, \boldsymbol{K_{B,\xi}}$, the matrices that directly map latent signals and noises to the final representations. For the case with contrastive pairs, one can show that the dynamics of $\boldsymbol{K_A}$ is governed by the following equation

$$\begin{aligned}
\dot{\boldsymbol{K}}_{\boldsymbol{A}} = & \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \left(\frac{\boldsymbol{f}_{\boldsymbol{B}}^+(\boldsymbol{z}^+)^\top}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\boldsymbol{K_A}}{N_{\boldsymbol{A}}^2 d} \right) \right\} \mathrm{diag}(\boldsymbol{\sigma}^2) \\
& - K \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^\pm} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \left(\frac{\boldsymbol{f}_{\boldsymbol{B}}^-(\boldsymbol{z}^+)^\top}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^- \right\rangle \frac{\boldsymbol{K_A}}{N_{\boldsymbol{A}}^2 d} \right) \right\} \mathrm{diag}(\boldsymbol{\sigma}^2) \\
& - K \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^\pm, \boldsymbol{x}_{\boldsymbol{B}}^+} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \left(\frac{\boldsymbol{f}_{\boldsymbol{B}}^+(\boldsymbol{z}^-)^\top}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^-, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\boldsymbol{K_A}}{N_{\boldsymbol{A}}^2 d} \right) \right\} \mathrm{diag}(\boldsymbol{\sigma}^2).
\end{aligned} \tag{5}$$

See Lemma A.3 for the calculation. The equations for the other $\boldsymbol{K}$-matrices can be derived similarly. Note that we rescale the gradients by $1/\tau_t^2$ so that $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K_A}$ does not shrink with τ_t^2. For the non-contrastive case, the equation is

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K_A} = \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+} \left\{ \frac{\boldsymbol{f}_{\boldsymbol{B}}^+(\boldsymbol{z}^+)^\top}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\boldsymbol{K_A}}{N_{\boldsymbol{A}}^2 d} \right\} \mathrm{diag}(\boldsymbol{\sigma}^2). \tag{6}$$

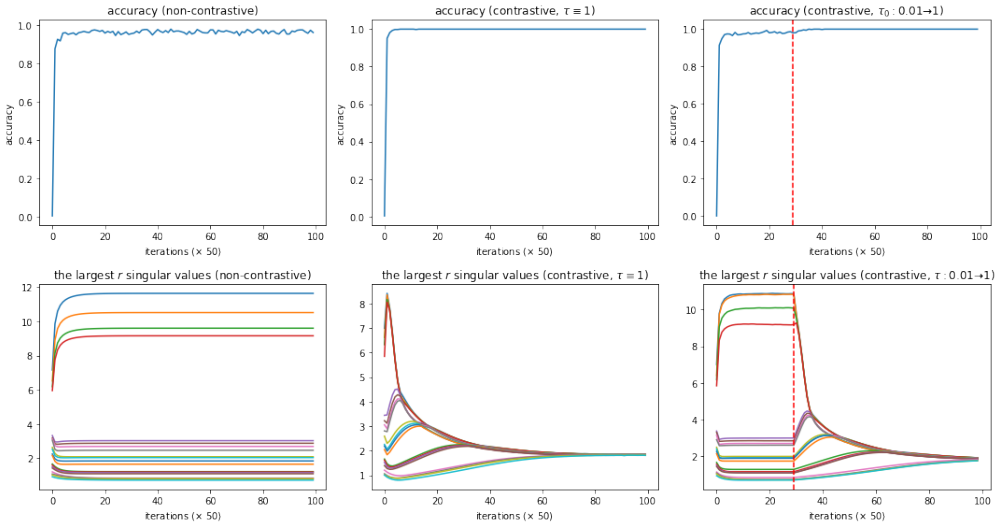

Figure 1: Simulation results. The first row reports the accuracies of different approaches on the classifying positive/negative pairs problem, and the second row reports the values of the largest r singular values. From left to right, the columns correspond to the non-contrastive method, contrastive method with $\tau_t \equiv 1$ throughout the entire process, and contrastive method with τ_t switch from 0.01 to 1 at the vertical dashed line. One can make several observations here. (a) All methods can quickly attain near 100% accuracy. (b) Only contrastive methods will reduce the condition number to approximately 1. (c) Even when $\tau_t \equiv 1$, we still have the stage-wise behavior, where the models first align the representations in Stage 1, and then balance the representations in Stage 2.

Note that the RHS resembles the first term of (5). This is not a coincidence. We will establish the approximate equivalence between the non-contrastive approach and the contrastive approach with a small inverse temperature τ_t^2 (cf. Section 5.3 and Lemma B.9).

5.2 The infinite-width dynamics

The overall proof strategy is to first characterize dynamics of the infinite-width limit, which is much simpler compared to (5), and then control the error introduced by discretizing the infinite-width network using polynomially many neurons. This discretization is one of the main technical challenges of the proof. In general, in order to track the infinite-width dynamics, an exponentially large network is needed (Mei et al. (2018)).

The basic idea is to use Taylor expansion around the infinite-width trajectory to factor out the first order error terms and show that, either they drive the process towards the infinite-width trajectory, or the error growth introduced by them is slower than the convergence rate.

Here, for the ease of presentation, we will focus on the noiseless infinite-width dynamics and, in particular, the evolution of the condition number. Recall that we use iid Gaussian to initialize the entries of $\boldsymbol{W_A}$ and $\boldsymbol{W_B}$. Hence, in the $m \to \infty$ limit, different columns of $\boldsymbol{K_A}$ and $\boldsymbol{K_B}$ are orthogonal to each other, at least at initialization. Moreover, one can verify that, thanks to the symmetry, this holds throughout the entire training procedure. Meanwhile, by symmetry, the corresponding quantities in modules $\boldsymbol{A}$ and $\boldsymbol{B}$ are always the same. Namely, when $m \to \infty$, we have $\boldsymbol{K_A}^\top \boldsymbol{K_A} = \boldsymbol{K_B}^\top \boldsymbol{K_B} = \mathrm{diag}(\boldsymbol{\kappa}^2)$ and $\boldsymbol{K_A}^\top \boldsymbol{K_B} = \mathrm{diag}(\hat{\boldsymbol{\kappa}}^2)$ for some $\boldsymbol{\kappa}, \hat{\boldsymbol{\kappa}} \in \mathbb{R}^r$. As a result, in order to characterize the dynamics of $\boldsymbol{K_A}, \boldsymbol{K_B}$, it suffices to look at $\boldsymbol{\kappa}^2$ and $\hat{\boldsymbol{\kappa}}^2$. One can show that, in this noiseless infinite-width limit, we have (cf. Lemma A.7)

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\kappa_p^2 &= 4\left(1-\tilde{S}\right)\left(\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2} - \frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2}\right)\sigma_p^2 - 4\left(1-\tilde{S}\right)\left(\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2}T_p - \frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\sigma_p^2, \\
\frac{\mathrm{d}}{\mathrm{d}t}\hat{\kappa}_p^2 &= 4\left(1-\tilde{S}\right)\left(\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2} - \frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2}\right)\sigma_p^2 - 4\left(1-\tilde{S}\right)\left(\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2}T_p - \frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\sigma_p^2,
\end{aligned} \tag{7}$$

where $\tilde{S}$ is a $\Theta(1)$ quantity depending on $\hat{\boldsymbol{\kappa}}$ and $\boldsymbol{\kappa}$, $T_p = \tanh\left(\tau_t^2 \hat{\kappa}_p^2 / \|\boldsymbol{\kappa}\|^2\right)$, and $\tilde{T} = \sum_{k=1}^r \left(\hat{\kappa}_p^2 / \|\boldsymbol{\kappa}\|^2\right) T_p$. The first term comes from the first term of (5) and the second term from the second and third terms of (5). In other words, the first term of (7) comes from the positive pairs, while the second term from the negative pairs.

5.3 Stage 1

In this subsection, we describe the dynamics of gradient descent in Stage 1, and explain how we control the growth of the errors and condition number.

Equivalence of Stage 1 and non-contrastive methods Recall (5). Since we use a small τ_t^2 in Stage 1, we have $\exp(\tau^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-) \approx 1$ and, as a result, $S_{\boldsymbol{A}} \approx S_{\boldsymbol{B}} \approx (1+K)^{-1}$. Then, note that $\boldsymbol{z}^\pm, \boldsymbol{\xi}_{\boldsymbol{A}}^\pm, \boldsymbol{\xi}_{\boldsymbol{B}}^\pm$ are independent mean zero random variables. Hence, the second and third terms of (5) are approximately 0. (This can also be inferred from (7) by noting that $T_p \approx 0$ when $\tau_t^2 \approx 0$.) Meanwhile, for the first term, since $2 - S_{\boldsymbol{A}} - S_{\boldsymbol{B}}$ is approximately constant, it approximately equals the RHS of (6), up to a constant factor. This gives the equivalence of the dynamics Stage 1 and the non-contrastive method. See Lemma B.9 for a more formal proof.

Now, we consider the contrastive The main result of Stage 1 is as follows.

Lemma 5.1 (Informal versin of Lemma B.1)**.** *Under the assumptions of Theorem 4.3, the finite-width dynamics closely track the infinite-width one throughout Stage 1, which takes at most* $\mathrm{poly}(d)$ *iterations, and, at the end of Stage 1, we have, for any* $p, q \in [r]$ *and* $s \in [d-r]$*,*

$$\frac{\hat{\kappa}_p^2}{\kappa_p^2} \approx 1, \quad \frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_s\|^2}{\kappa_p^2} \approx 0, \quad \frac{\kappa_p^2}{\kappa_q^2} \leq O\left(\sqrt{d}\right). \tag{8}$$

Moreover, there exists a $\boldsymbol{\sigma} \in \mathbb{R}^r$ *such that* $\max_{p,q} \kappa_p^2/\kappa_q^2 = \Omega(\sqrt{d})$ *at the end of Stage 1.*

In words, (8) says, at the end of Stage 1, $\boldsymbol{K}_{\boldsymbol{A}} \approx \boldsymbol{K}_{\boldsymbol{B}}$ in the relative sense, the noise-signal ratio is small, and the condition number is bounded by $O(\sqrt{d})$. By Lemma 4.4, the first two conditions of (8) imply that the learned representations are aligned. The proof of this lemma can be found in Section B. Basically, we couple the convergence of $\hat{\kappa}_p^2/\kappa_p^2$ and the noise-signal ratio with the growth of discretization error and condition number. The main tool we use is the following nonlinear version of Gronwall's lemma.

Lemma 5.2. *Let* A_t *be a positive process. Let* X_t *and* Y_t *be defined as* $\dot{X}_t \leq -A_t X_t, \dot{Y}_t \leq \beta A_t X_t Y_t$*, with* X_0, Y_0, β *being positive. Then, for any* $T \geq 0$*, we have* $Y_T \leq Y_0 \exp(\beta X_0)$*.*

Here, X_t represent the progress we have made and Y_t the error we wish to control. In our case, X_t is the maximum between $1 - \hat{\kappa}_p^2/\kappa_p^2$ and the noise-signal ratio, and Y_t the discretization error and condition number. This lemma says that, if the error growth rate depends on the progress, then by coupling these two processes, we can make sure the error does not blow up. The point of this lemma is that, with coupling, we do not need a very tight estimation on the convergence time nor the error growth rate.

5.4 Stage 2

In Stage 2, τ_t^2 is no longer $o(1)$, and now the second term of (7) comes into play. We show that, in this stage, the model will reduce the condition number of $\boldsymbol{K}_{\boldsymbol{A}}$ to approximately 1. By Lemma 4.5, this implies that the learned representations are balanced. Formally, we have the following lemma. The proof can be found in Section C.

Lemma 5.3. *Informal version of Lemma C.1 Under the assumptions of Theorem 4.3, the finite-width dynamics still closely track the infinite-one throughout Stage 2, which again takes at most* $\mathrm{poly}(d)$ *iterations, and, at the end of Stage 2, we have, for any* $p, q \in [r]$ *and* $s \in [d-r]$*,*

$$\frac{\hat{\kappa}_p^2}{\kappa_p^2} \approx 1, \quad \frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_s\|^2}{\kappa_p^2} \approx 0, \quad \frac{\kappa_p^2}{\kappa_q^2} \approx 1. \tag{9}$$

The way to controlling $\hat{\kappa}_p^2/\kappa_p^2$ and the noise-signal ratio is similar to Stage 1. For the condition number, note that when $\tau_t^2 = 1$ and $\hat{\boldsymbol{\kappa}} \approx \boldsymbol{\kappa}$, the first term of (7) becomes close to 0 while the second term becomes $-4(1-\tilde{S})\kappa_p^2/\left\|\boldsymbol{\kappa}\right\|^2\left(T_p - \tilde{T}\right)\sigma_p^2$. Recall $T_p = \tanh\left(\tau_t^2\hat{\kappa}_p^2/\left\|\boldsymbol{\kappa}\right\|^2\right)$ is positively correlated with $\hat{\kappa}_p^2/\left\|\boldsymbol{\kappa}\right\|^2$ and $\tilde{T}$ is a weight average of these T_p's, this term will push $\hat{\kappa}_p^2/\left\|\boldsymbol{\kappa}\right\|^2$ towards their average and, consequently, reduces the condition number. To obtain a satisfactory convergence rate, we consider the ratio directly and compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\kappa_p^2}{\kappa_q^2} \approx -\frac{4(1-\tilde{S})}{\left\|\boldsymbol{\kappa}\right\|^2}\left(\left(T_p - \tilde{T}\right)\sigma_p^2 - \left(T_q - \tilde{T}\right)\sigma_q^2\right)\frac{\kappa_p^2}{\kappa_q^2}.$$

Suppose that κ_p^2 is the largest and κ_q^2 is the smallest singular values. Then, we have $T_p - \tilde{T} \geq 0$ and $T_q - \tilde{T} \leq 0$. As a result,

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\kappa_p^2}{\kappa_q^2} \lesssim -\frac{4(1-\tilde{S})\sigma_{\min}^2}{\left\|\boldsymbol{\kappa}\right\|^2}\left(T_p - T_q\right)\frac{\kappa_p^2}{\kappa_q^2}.$$

In Stage 2, controlling the discretization error is more complicated than in Stage 1 because of the existence of contrastive terms and the fact that we can no longer approximate $\exp(\tau_t^2 \boldsymbol{f_A}\cdot\boldsymbol{f_B})$ with a constant. For the ease of presentation, we assume here the noise-signal ratio is 0 and $\boldsymbol{K_A} = \boldsymbol{K_B}$, and explain how to control $\left\langle\overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle$ for $p \neq q \in [r]$ under these assumptions. In the actual proof, we do handle these rigorously. For notational simplicity, we write $\boldsymbol{K_{AB}} = \boldsymbol{K_A}^\top\boldsymbol{K_B}$. First, we write

$$\begin{aligned}\left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^-}\right\rangle = \frac{\left\langle \boldsymbol{K_A}^\top\boldsymbol{K_B}\boldsymbol{z}^+, \boldsymbol{z}^-\right\rangle}{\left\|\boldsymbol{K_A}\right\|_F^2/d} &= \sum_{k=1}^r \frac{[\boldsymbol{K_{AB}}]_{k,k}}{\left\|[\boldsymbol{K_A}]_F^2\right\|/d} z_k^+ z_k^- + \sum_{i\neq j}\frac{[\boldsymbol{K_{AB}}]_{i,j}}{\left\|[\boldsymbol{K_A}]_F^2\right\|/d} z_i^+ z_j^- \\ &=: I_{+,-} + \delta_{+,-}.\end{aligned}$$

Then we have $\exp(\boldsymbol{f_A^+}\cdot\boldsymbol{f_B^-}) \approx \exp(I_{+,-})(1+\delta_{+,-})$. Repeat this process for the $S_{\boldsymbol{A}}$, $S_{\boldsymbol{B}}$ and exp in (5), ignore all higher order terms, and eventually one can obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle \approx [\boldsymbol{Q}_1]_{q,p}\frac{\kappa_q/\kappa_p\sigma_p^2 + \kappa_p/\kappa_q\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d},$$

for some matrix $\boldsymbol{Q} \in \mathbb{R}^{r\times r}$. Moreover, one can show that

$$[\boldsymbol{Q}_1]_{p,q} \approx -(1-\tilde{S})\left(2 - T_p^2 - T_q^2 + \tilde{S}(2 - T_p - T_q)^2\right)\frac{[\boldsymbol{K_{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}.$$

Note that the coefficient is negative. As a result, $\left\langle\overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle$ will move towards 0.

6 Conclusion and Discussion

In this work, we study the role of contrastive pairs in multimodal learning, and show that contrastive pairs are important for the model to learn representations that are both **aligned and balanced**. Our work extends previous results in several directions: First, we consider the more complicated multimodal learning problem. Meanwhile, our data generating model is inhomogeneous, and we show that in this case non-contrastive method can collapse to an approximately rank-1 solution while contrastive method can learn all features. We also includes output normalization in our analysis, a technique that is widely used in practice but is still under-studied in theory.

However, despite the complexity of the analysis, our model is still linear, which is very different from the models used in practice. Also, for the results on non-contrastive methods, we do not consider more advanced training techniques such as Grill et al. (2020) and Chen & He (2021). We leave the analysis of these more practical techniques for future work.

REFERENCES

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A Theoretical Analysis of Contrastive Unsupervised Representation Learning, February 2019. URL http://arxiv.org/abs/1902.09229. arXiv:1902.09229 [cs, stat].

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. *arXiv:2106.04156 [cs, stat]*, August 2021. URL http://arxiv.org/abs/2106.04156. arXiv: 2106.04156.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

X. He and Y. Peng. Fine-grained image classification via combining vision and language. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7332–7340, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.775. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.775.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What Makes Multi-Modal Learning Better than Single (Provably). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10944–10956. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/5aa3405a3f865c10f420a4a7b55cbff3-Paper.pdf.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Jason D Lee, Qi Lei, Nikunj Saunshi, and JIACHENG ZHUO. Predicting What You Already Know Helps: Provable Self-Supervised Learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 309–323. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/02e656adee09f8394b402d9958389b7d-Paper.pdf.

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*, 115(33):E7665–E7671, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1806579115. URL https://www.pnas.org/content/115/33/E7665.

Kento Nozawa and Issei Sato. Understanding Negative Samples in Instance Discriminative Self-supervised Representation Learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=pZ5X_svdPQ.

Ashwini Pokle, Jinjin Tian, Yuchen Li, and Andrej Risteski. Contrasting the landscape of contrastive and non-contrastive learning, March 2022. URL http://arxiv.org/abs/2203.15702. arXiv:2203.15702 [cs, stat].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv: 2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ramesh21a.html.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):21–es, jul 2007. ISSN 0004-5411. doi: 10.1145/1255443.1255449. URL https://doi.org/10.1145/1255443.1255449.

Jonathan C. Stroud, Zhichao Lu, Chen Sun, Jia Deng, Rahul Sukthankar, Cordelia Schmid, and David A. Ross. Learning video representations from textual web supervision, 2020. URL https://arxiv.org/abs/2007.14937.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10268–10278. PMLR, July 2021. URL https://proceedings.mlr.press/v139/tian21a.html.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models, April 2021. URL http://arxiv.org/abs/2008.10150. arXiv:2008.10150 [cs, stat].

Tongzhou Wang and Phillip Isola. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939. PMLR, November 2020. URL https://proceedings.mlr.press/v119/wang20k.html. ISSN: 2640-3498.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a Ladder: A New Theoretical Understanding of Contrastive Learning via Augmentation Overlap. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=ECvgmYVyeUz.

Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks, 2022b.

Zixin Wen and Yuanzhi Li. Toward Understanding the Feature Learning Process of Self-supervised Contrastive Learning. *arXiv:2105.15134 [cs, stat]*, May 2021. URL http://arxiv.org/abs/2105.15134. arXiv: 2105.15134.

Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model, 2021.

A GRADIENT CALCULATION

In this section, we compute the gradients and the equations governing some related quantities. We first consider the general finite-width case, and then the infinite-width case, in which we have simple formulas for many quantities of interests.

A.1 THE FINITE-WIDTH CASE

First, we prove the following auxiliary lemma.

Lemma A.1. *For any $\boldsymbol{x}_{\boldsymbol{A}}^{+}$ and $\boldsymbol{x}_{\boldsymbol{B}}^{-}$, we have*

$$\nabla_{\boldsymbol{W}_{\boldsymbol{A}}} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-}) = \frac{\tau_t^2 \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-})}{N_{\boldsymbol{A}}} \left(\boldsymbol{x}_{\boldsymbol{A}}^{+} (\boldsymbol{f}_{\boldsymbol{B}}^{-})^{\top} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^{\top} + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^{\top} \right)/d}{N_{\boldsymbol{A}}} \right).$$

Then we compute the gradients.

Lemma A.2. *We have*

$$\begin{aligned}
\nabla_{\boldsymbol{W}_{\boldsymbol{A}}} \mathcal{L} = & -\tau_t^2 \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \right) \left(\frac{\boldsymbol{x}_{\boldsymbol{A}}^{+} (\boldsymbol{f}_{\boldsymbol{B}}^{+})^{\top}}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^{\top} + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^{\top} \right)}{N_{\boldsymbol{A}}^2 d} \right) \right\} \\
& + \frac{K\tau_t^2}{N_{\boldsymbol{A}}} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{\pm}} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \left(\boldsymbol{x}_{\boldsymbol{A}}^{+} (\boldsymbol{f}_{\boldsymbol{B}}^{-})^{\top} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^{\top} + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^{\top} \right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\} \\
& + \frac{K\tau_t^2}{N_{\boldsymbol{A}}} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{\pm}, \boldsymbol{x}_{\boldsymbol{B}}^{+}} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{-} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \left(\boldsymbol{x}_{\boldsymbol{A}}^{-} (\boldsymbol{f}_{\boldsymbol{B}}^{+})^{\top} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{-}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^{\top} + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^{\top} \right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\}.
\end{aligned}$$

The formula for $\nabla_{\boldsymbol{W}_{\boldsymbol{B}}} \mathcal{L}$ can be obtained by interchanging the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Instead of tracking $\boldsymbol{W}_{\boldsymbol{A}}$ and $\boldsymbol{W}_{\boldsymbol{B}}$ directly, we are going to track $\boldsymbol{K}_{\boldsymbol{A}}$, $\boldsymbol{K}_{\boldsymbol{B}}$, $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}$ and $\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}$. The dynamics of them is governed by the following equation. As a direct corollary of Lemma A.2, we have the following.

Lemma A.3. *We have*

$$\begin{aligned}
& \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{K}_{\boldsymbol{A}} \\
= & \tau_t^2 \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \right) \left(\frac{(\boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^{+} + \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{B}})(\boldsymbol{z}^{+})^{\top}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d} \right) \right\} \operatorname{diag}(\boldsymbol{\sigma}^2) \\
& - K\tau_t^2 \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{\pm}} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \left(\frac{\boldsymbol{f}_{\boldsymbol{B}}^{-}(\boldsymbol{z}^{+})^{\top}}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d} \right) \right\} \operatorname{diag}(\boldsymbol{\sigma}^2) \\
& - K\tau_t^2 \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{\pm}, \boldsymbol{x}_{\boldsymbol{B}}^{+}} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{-} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \left(\frac{\boldsymbol{f}_{\boldsymbol{B}}^{+}(\boldsymbol{z}^{-})^{\top}}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{-}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d} \right) \right\} \operatorname{diag}(\boldsymbol{\sigma}^2),
\end{aligned}$$

and

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} = & \frac{\tau_t^2}{N_{\boldsymbol{A}}} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \right) \left(\boldsymbol{f}_{\boldsymbol{B}}^{+}(\boldsymbol{\xi}_{\boldsymbol{A}}^{+})^{\top} - \frac{\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\} \sigma_\xi^2 \\
& - \frac{K\tau_t^2}{N_{\boldsymbol{A}}} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{\pm}} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \left(\boldsymbol{f}_{\boldsymbol{B}}^{-}(\boldsymbol{\xi}_{\boldsymbol{A}}^{+})^{\top} - \frac{\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\} \sigma_\xi^2 \\
& - \frac{K\tau_t^2}{N_{\boldsymbol{A}}} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{\pm}, \boldsymbol{x}_{\boldsymbol{B}}^{+}} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{-} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \left(\boldsymbol{f}_{\boldsymbol{B}}^{+}(\boldsymbol{\xi}_{\boldsymbol{A}}^{-})^{\top} - \frac{\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{-}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\} \sigma_\xi^2.
\end{aligned}$$

We can rewrite the above result as follows.

Corollary A.4. *Define*

$$
\begin{aligned}
Q_0 := &- \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \right\} \\
&+ K \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^\pm} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^- \right\rangle \right\} \\
&+ K \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^\pm, \boldsymbol{x}_{\boldsymbol{B}}^+} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \left\langle \boldsymbol{f}_{\boldsymbol{A}}^-, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \right\},
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{Q}_1 := &\, \mathbb{E} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \boldsymbol{z}^+ (\boldsymbol{z}^+)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{z}^- (\boldsymbol{z}^+)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{z}^+ (\boldsymbol{z}^-)^\top d \right\}, \\
\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}} := &\, \mathbb{E} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \boldsymbol{\xi}_{\boldsymbol{B}}^+ (\boldsymbol{z}^+)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{B}}^- (\boldsymbol{z}^+)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{B}}^+ (\boldsymbol{z}^-)^\top d \right\}, \\
\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}} := &\, \mathbb{E} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \boldsymbol{\xi}_{\boldsymbol{A}}^+ (\boldsymbol{z}^+)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{A}}^+ (\boldsymbol{z}^-)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{A}}^- (\boldsymbol{z}^+)^\top d \right\}, \\
\boldsymbol{Q}_2 := &\, \mathbb{E} \left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \boldsymbol{\xi}_{\boldsymbol{B}}^+ (\boldsymbol{\xi}_{\boldsymbol{A}}^+)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{B}}^- (\boldsymbol{\xi}_{\boldsymbol{A}}^+)^\top d \right\} \\
&- K\, \mathbb{E} \left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{B}}^+ (\boldsymbol{\xi}_{\boldsymbol{A}}^-)^\top d \right\}.
\end{aligned}
$$

We have

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{K}_{\boldsymbol{A}} &= \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_1 \operatorname{diag}(\boldsymbol{\sigma}^2) + \frac{\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}} \operatorname{diag}(\boldsymbol{\sigma}^2) + \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d} Q_0 \operatorname{diag}(\boldsymbol{\sigma}^2), \\
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{K}_{\boldsymbol{B}} &= \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_1^\top \operatorname{diag}(\boldsymbol{\sigma}^2) + \frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}} \operatorname{diag}(\boldsymbol{\sigma}^2) + \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{B}}^2 d} Q_0 \operatorname{diag}(\boldsymbol{\sigma}^2), \\
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} &= \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top \sigma_\xi^2 + \frac{\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_2 \sigma_\xi^2 + \frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d} Q_0 \sigma_\xi^2, \\
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} &= \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}^\top \sigma_\xi^2 + \frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \boldsymbol{Q}_2^\top \sigma_\xi^2 + \frac{\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}}{N_{\boldsymbol{B}}^2 d} Q_0 \sigma_\xi^2.
\end{aligned}
$$

We are interested in each column of $\boldsymbol{K}_{\boldsymbol{A}}$ and $\boldsymbol{K}_{\boldsymbol{B}}$, whose dynamics is given by the next lemma. The next lemma also decompose the dynamics along the radial and tangent directions.

Lemma A.5. *For any $p \in [r]$ and $q \in [d-r]$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 = 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_p^2,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\|^2 = 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_1^\top]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\|^2}{N_{\boldsymbol{B}}^2 d}Q_0\sigma_p^2,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2 = 2\frac{\left\langle [\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q, [\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top]_q\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_2]_q\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\right\|^2 = 2\frac{\left\langle [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q, [\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}^\top]_q\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q, [\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_2^\top]_q\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2,$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\left(\frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} + \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\right)\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d},$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right)^\top\right)\left(\frac{[\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_1^\top]_p}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|} + \frac{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|}\right)\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d},$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right)\left(\frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|} + \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_2]_q}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d},$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right)^\top\right)\left(\frac{[\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|} + \frac{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_2^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}.$$

Finally, as a simple corollary of Lemma A.1, we have the following result on the gradients of the non-contrastive loss.

Lemma A.6. *For the non-contrastive loss (4), we have*

$$\nabla_{\boldsymbol{W}_{\boldsymbol{A}}}\hat{\mathcal{L}} = \mathbb{E}\left\{\frac{\boldsymbol{x}_{\boldsymbol{A}}^+(\boldsymbol{f}_{\boldsymbol{B}}^+)^\top}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+\right\rangle\frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^\top + \boldsymbol{A}_{\boldsymbol{\xi}}\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^\top\right)}{N_{\boldsymbol{A}}^2 d}\right\}$$

As a corollary, we have, in this case,

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A}} = \mathbb{E}\left\{\frac{\boldsymbol{f}_{\boldsymbol{B}}^+(\boldsymbol{z}^+)^\top}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+\right\rangle\frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d}\right\}\operatorname{diag}(\boldsymbol{\sigma}^2),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} = \mathbb{E}\left\{\frac{\boldsymbol{f}_{\boldsymbol{B}}^+(\boldsymbol{\xi}_{\boldsymbol{A}}^+)^\top}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+\right\rangle\frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d}\right\}\sigma_\xi^2.$$

OMITTED PROOF OF THIS SUBSECTION

Proof of Lemma A.1. We compute

$$\begin{aligned}
\nabla_{\boldsymbol{W}_{\boldsymbol{A}}}\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-) &= \tau_t^2\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-)\nabla_{\boldsymbol{W}_{\boldsymbol{A}}}\left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^-\right\rangle \\
&= \tau_t^2\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-)\nabla_{\boldsymbol{W}_{\boldsymbol{A}}}\frac{\left\langle \boldsymbol{W}_{\boldsymbol{A}}^\top\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^-\right\rangle}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}/\sqrt{d}} \\
&= \tau_t^2\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-)\frac{\nabla_{\boldsymbol{W}_{\boldsymbol{A}}}\left\langle \boldsymbol{W}_{\boldsymbol{A}}^\top\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^-\right\rangle}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}/\sqrt{d}} \\
&\quad - \tau_t^2\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-)\left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^-\right\rangle\frac{\nabla_{\boldsymbol{W}_{\boldsymbol{A}}}\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}}.
\end{aligned}$$

For the first term, we have $\nabla_{\boldsymbol{W_A}} \left\langle \boldsymbol{W_A}^\top \boldsymbol{x_A^+}, \boldsymbol{f_B^-} \right\rangle = \boldsymbol{x_A^+}(\boldsymbol{f_B^-})^\top$. For the second term, we have

$$\frac{\nabla_{\boldsymbol{W_A}} \sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}} = \frac{\nabla_{\boldsymbol{W_A}} \|\boldsymbol{K_A}\|_F^2 + \nabla_{\boldsymbol{W_A}} \|\boldsymbol{K_{A,\xi}}\|_F^2}{2(\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2)} = \frac{\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top}{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}.$$

Hence,

$$\begin{aligned}
\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-}) &= \tau_t^2 \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-}) \frac{\boldsymbol{x_A^+}(\boldsymbol{f_B^-})^\top}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}/\sqrt{d}} \\
&\quad - \tau_t^2 \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-}) \left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^-} \right\rangle \frac{\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top}{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2} \\
&= \frac{\tau_t^2 \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}/\sqrt{d}} \left(\boldsymbol{x_A^+}(\boldsymbol{f_B^-})^\top - \left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^-} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}} \right).
\end{aligned}$$

□

Proof of Lemma A.2. We compute

$$\begin{aligned}
\nabla_{\boldsymbol{W_A}} \mathcal{L}_{\boldsymbol{A}} &= -\mathbb{E} \frac{\nabla_{\boldsymbol{W_A}} S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+})}{S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+})} \\
&= -\mathbb{E} \left\{ \frac{1}{S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+})} \nabla_{\boldsymbol{W_A}} \frac{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+})}{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+}) + K\, \mathbb{E}_{\boldsymbol{z}^-} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})} \right\} \\
&= -\mathbb{E} \left\{ \frac{1}{S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+})} \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+})}{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+}) + K\, \mathbb{E}_{\boldsymbol{z}^-} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})} \right\} \\
&\quad + \mathbb{E} \left\{ \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+}) + K\, \mathbb{E}_{\boldsymbol{z}^-} \nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})}{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+}) + K\, \mathbb{E}_{\boldsymbol{z}^-} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})} \right\}.
\end{aligned}$$

By Lemma A.1, the first term is

$$\begin{aligned}
&-\mathbb{E} \left\{ \frac{1}{S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+})} \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+})}{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+}) + K\, \mathbb{E}_{\boldsymbol{z}^-} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})} \right\} \\
&= -\frac{\tau_t^2}{N_{\boldsymbol{A}}} \mathop{\mathbb{E}}_{\boldsymbol{x_A^+}, \boldsymbol{x_B^+}} \left\{ \boldsymbol{x_A^+}(\boldsymbol{f_B^+})^\top - \left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^+} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}} \right\},
\end{aligned}$$

and the second term is

$$\begin{aligned}
&\mathbb{E} \left\{ \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+}) + K\, \mathbb{E}_{\boldsymbol{x_B^-}} \nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})}{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+}) + K\, \mathbb{E}_{\boldsymbol{z}^-} \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})} \right\} \\
&= \frac{\tau_t^2}{N_{\boldsymbol{A}}} \mathbb{E} \left\{ S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+}) \left(\boldsymbol{x_A^+}(\boldsymbol{f_B^+})^\top - \left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^+} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}} \right) \right\} \\
&\quad + \frac{K\tau_t^2}{N_{\boldsymbol{A}}} \mathbb{E} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+}) \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})}{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+})} \left(\boldsymbol{x_A^+}(\boldsymbol{f_B^-})^\top - \left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^-} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}} \right) \right\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\nabla_{\boldsymbol{W_A}} \mathcal{L}_{\boldsymbol{A}} = &-\frac{\tau_t^2}{N_{\boldsymbol{A}}} \mathop{\mathbb{E}}_{\boldsymbol{x_A^+}, \boldsymbol{x_B^+}} \left\{ \left(1 - S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+})\right) \left(\boldsymbol{x_A^+}(\boldsymbol{f_B^+})^\top - \left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^+} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}} \right) \right\} \\
&+ \frac{K\tau_t^2}{N_{\boldsymbol{A}}} \mathbb{E} \left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x_A^+}, \boldsymbol{x_B^+}) \exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^-})}{\exp(\tau_t^2 \boldsymbol{f_A^+} \cdot \boldsymbol{f_B^+})} \left(\boldsymbol{x_A^+}(\boldsymbol{f_B^-})^\top - \left\langle \boldsymbol{f_A^+}, \boldsymbol{f_B^-} \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K_A}^\top + \boldsymbol{A_\xi}\boldsymbol{K_{A,\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K_A}\|_F^2 + \|\boldsymbol{K_{A,\xi}}\|_F^2}} \right) \right\}.
\end{aligned}$$

Then, for $\nabla_{\boldsymbol{W_A}} \mathcal{L}_{\boldsymbol{B}}$, we compute

$$\begin{aligned}
\nabla_{\boldsymbol{W_A}} \mathcal{L}_{\boldsymbol{B}} &= -\mathbb{E}\, \frac{\nabla_{\boldsymbol{W_A}} S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)}{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)} \\
&= -\mathbb{E}\left\{ \frac{1}{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)} \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \right\} \\
&\quad + \mathbb{E}\left\{ \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K \nabla_{\boldsymbol{W_A}}\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \right\}.
\end{aligned}$$

Again, by Lemma A.1, the first term is

$$\begin{aligned}
&-\mathbb{E}\left\{ \frac{1}{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)} \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \right\} \\
&= -\frac{\tau_t^2}{N_{\boldsymbol{A}}}\, \mathbb{E}\left\{ \boldsymbol{x}_{\boldsymbol{A}}^+ (\boldsymbol{f}_{\boldsymbol{B}}^+)^\top - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^\top + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right\},
\end{aligned}$$

and the second term is

$$\begin{aligned}
&\mathbb{E}\left\{ \frac{\nabla_{\boldsymbol{W_A}} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K \nabla_{\boldsymbol{W_A}}\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+) + K\, \mathbb{E}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \right\} \\
&= \frac{\tau_t^2}{N_{\boldsymbol{A}}}\, \mathbb{E}\left\{ S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \left(\boldsymbol{x}_{\boldsymbol{A}}^+ (\boldsymbol{f}_{\boldsymbol{B}}^+)^\top - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^\top + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\} \\
&\quad + \frac{K\tau_t^2}{N_{\boldsymbol{A}}}\, \mathbb{E}\left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \left(\boldsymbol{x}_{\boldsymbol{A}}^- (\boldsymbol{f}_{\boldsymbol{B}}^+)^\top - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^-, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^\top + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\nabla_{\boldsymbol{W_A}} \mathcal{L}_{\boldsymbol{B}} &= -\frac{\tau_t^2}{N_{\boldsymbol{A}}}\, \mathbb{E}\left\{ (1 - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)) \left(\boldsymbol{x}_{\boldsymbol{A}}^+ (\boldsymbol{f}_{\boldsymbol{B}}^+)^\top - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^\top + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\} \\
&\quad + \frac{K\tau_t^2}{N_{\boldsymbol{A}}}\, \mathbb{E}\left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \left(\boldsymbol{x}_{\boldsymbol{A}}^- (\boldsymbol{f}_{\boldsymbol{B}}^+)^\top - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^-, \boldsymbol{f}_{\boldsymbol{B}}^+ \right\rangle \frac{\left(\boldsymbol{A}\boldsymbol{K}_{\boldsymbol{A}}^\top + \boldsymbol{A}_{\boldsymbol{\xi}} \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}^\top\right)/\sqrt{d}}{\sqrt{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2 + \|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}} \right) \right\}.
\end{aligned}$$

Combine these formulas together and we complete the proof. □

A.2 THE INFINITE-WIDTH CASE

Now we consider the noiseless infinite-width dynamics. The results of this subsection will not be used in the proof. It mainly serves as a way to give intuitions on how the dynamics looks. As we have discussed in the main text, in this noiseless infinite-width case, it suffices to track $\kappa_p^2 = \|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2$ and $\hat{\kappa}_p^2 = \langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p \rangle$.

Lemma A.7. *In the noiseless infinite-width case, we have*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \kappa_p^2 &= 4\left(1 - \tilde{S}\right) \left(\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2} - \frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2} \frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2} \right) \sigma_p^2 - 4\left(1 - \tilde{S}\right) \left(\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2} T_p - \frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2} \tilde{T} \right) \sigma_p^2, \\
\frac{\mathrm{d}}{\mathrm{d}t} \hat{\kappa}_p^2 &= 4\left(1 - \tilde{S}\right) \left(\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2} - \frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2} \frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2} \right) \sigma_p^2 - 4\left(1 - \tilde{S}\right) \left(\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2} T_p - \frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2} \tilde{T} \right) \sigma_p^2,
\end{aligned}$$

Proof. First, note hat

$$\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle = \frac{\left\langle \boldsymbol{K}_{\boldsymbol{BA}} \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle}{\left\| \boldsymbol{\kappa} \right\|^{2} / d} = \frac{\left\langle \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle_{\hat{\boldsymbol{\kappa}}^{2}}}{\left\| \boldsymbol{\kappa} \right\|^{2} / d}.$$

This implies that (a) $\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle$ does not depend on the actual value of $\boldsymbol{z}^{+}$, and (b) if we flip the signs of $z_p^{\pm}$ simultaneously, then the value of $\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle$ remain unchanged. To compute $S_{\boldsymbol{A}}$, we then need to take expectation over $\boldsymbol{z}^{-}$. We compute

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^{-}} \exp\left(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-}\right) = \prod_{k=1}^{r} \mathop{\mathbb{E}}_{z_k^{-}} \exp\left(\tau_t^2 \frac{\hat{\kappa}_k^2 z_k^{+} z_k^{-}}{\left\| \boldsymbol{\kappa} \right\|^{2} / d}\right) = \prod_{k=1}^{r} \cosh\left(\frac{\tau_t^2 \hat{\kappa}_k^2}{\left\| \boldsymbol{\kappa} \right\|^{2}}\right) =: Z_c.$$

Again, it does not depend on the actual value of $\boldsymbol{z}^{+}$. One can conduct similar calculation for $S_{\boldsymbol{B}}$ and, consequently, we have $S_{\boldsymbol{A}} \equiv S_{\boldsymbol{B}} \equiv \tilde{S}$ for some $\tilde{S}$ that depends on $\boldsymbol{\kappa}$ and $\hat{\boldsymbol{\kappa}}$ but not on $\boldsymbol{z}^{+}$. Then, we can rewrite (5) as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{K}_{\boldsymbol{A}} &= 2\left(1 - \tilde{S}\right) \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}} \left\{ \frac{\boldsymbol{f}_{\boldsymbol{B}}^{+}(\boldsymbol{z}^{+})^{\top}}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{+} \right\rangle \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^{2} d} \right\} \operatorname{diag}(\boldsymbol{\sigma}^{2}) \\
&\quad - 2K\tilde{S} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{\pm}} \left\{ \frac{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \left(\frac{\boldsymbol{f}_{\boldsymbol{B}}^{-}(\boldsymbol{z}^{+})^{\top}}{N_{\boldsymbol{A}}} - \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^{2} d} \right) \right\} \operatorname{diag}(\boldsymbol{\sigma}^{2}) \\
&= 2\left(1 - \tilde{S}\right) \left(\frac{\boldsymbol{K}_{\boldsymbol{B}}}{\left\| \boldsymbol{\kappa} \right\|^{2}} - \frac{\left\| \hat{\boldsymbol{\kappa}} \right\|^{2}}{\left\| \boldsymbol{\kappa} \right\|^{2}} \frac{\boldsymbol{K}_{\boldsymbol{A}}}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) \operatorname{diag}(\boldsymbol{\sigma}^{2}) \\
&\quad - \frac{2K\tilde{S}}{\exp\left(\frac{\tau_t^2 \left\| \hat{\boldsymbol{\kappa}} \right\|^{2}}{\left\| \boldsymbol{\kappa} \right\|^{2}}\right)} \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{\pm}} \left\{ \exp\left(\frac{\tau_t^2 \left\langle \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle_{\hat{\boldsymbol{\kappa}}^{2}} d}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) \left(\frac{\boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^{-}(\boldsymbol{z}^{+})^{\top}}{\left\| \boldsymbol{\kappa} \right\|^{2} / d} - \frac{\left\langle \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle_{\hat{\boldsymbol{\kappa}}^{2}}}{\left\| \boldsymbol{\kappa} \right\|^{2} / d} \frac{\boldsymbol{K}_{\boldsymbol{A}}}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) \right\} \operatorname{diag}(\boldsymbol{\sigma}^{2}).
\end{aligned}$$

Note that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{z}^{\pm}} \left\{ \exp\left(\frac{\tau_t^2 \left\langle \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle_{\hat{\boldsymbol{\kappa}}^{2}} d}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) z_p^{-} z_q^{+} d \right\} \\
&= \mathbb{1}\{p = q\} \mathop{\mathbb{E}}_{z_p^{\pm}} \left\{ \exp\left(\frac{\tau_t^2 \hat{\kappa}_k^2 z_k^{+} z_k^{-} d}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) z_p^{-} z_p^{+} d \right\} \prod_{k \neq p} \mathop{\mathbb{E}}_{z_k^{\pm}} \left\{ \exp\left(\frac{\tau_t^2 \hat{\kappa}_k^2 z_k^{+} z_k^{-} d}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) \right\} \\
&= \mathbb{1}\{p = q\} \sinh\left(\frac{\tau_t^2 \hat{\kappa}_k^2}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) \prod_{k \neq p} \cosh\left(\frac{\tau_t^2 \hat{\kappa}_k^2}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) \\
&= \mathbb{1}\{p = q\} Z_c T_p.
\end{aligned}$$

Meanwhile, note that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{\pm}} \left\{ \exp\left(\frac{\tau_t^2 \left\langle \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle_{\hat{\boldsymbol{\kappa}}^{2}} d}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) \frac{\left\langle \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle_{\hat{\boldsymbol{\kappa}}^{2}}}{\left\| \boldsymbol{\kappa} \right\|^{2} / d} \right\} &= \sum_{k=1}^{r} \frac{\hat{\kappa}_k^2}{\left\| \boldsymbol{\kappa} \right\|^{2} / d} \mathop{\mathbb{E}}_{\boldsymbol{z}^{\pm}} \left\{ \exp\left(\frac{\tau_t^2 \left\langle \boldsymbol{z}^{+}, \boldsymbol{z}^{-} \right\rangle_{\hat{\boldsymbol{\kappa}}^{2}} d}{\left\| \boldsymbol{\kappa} \right\|^{2}} \right) z_k^{+} z_k^{-} \right\} \\
&= Z_c \sum_{k=1}^{r} \frac{\hat{\kappa}_k^2}{\left\| \boldsymbol{\kappa} \right\|^{2}} T_k \\
&=: Z_c \tilde{T}.
\end{aligned}$$

Thus,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K_A} &= 2\left(1-\tilde{S}\right)\left(\frac{\boldsymbol{K_B}}{\|\boldsymbol{\kappa}\|^2}-\frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{\boldsymbol{K_A}}{\|\boldsymbol{\kappa}\|^2}\right)\mathrm{diag}(\boldsymbol{\sigma}^2)\\ &\quad -\frac{2K\tilde{S}Z_c}{\exp\left(\frac{\tau_t^2\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\right)}\left(\frac{\boldsymbol{K_B}}{\|\boldsymbol{\kappa}\|^2}\mathrm{diag}\left([T_k]_{k\in[r]}\right)-\frac{\boldsymbol{K_A}}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\mathrm{diag}(\boldsymbol{\sigma}^2)\\ &= 2\left(1-\tilde{S}\right)\left(\frac{\boldsymbol{K_B}}{\|\boldsymbol{\kappa}\|^2}-\frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{\boldsymbol{K_A}}{\|\boldsymbol{\kappa}\|^2}\right)\mathrm{diag}(\boldsymbol{\sigma}^2)\\ &\quad -2\left(1-\tilde{S}\right)\left(\frac{\boldsymbol{K_B}}{\|\boldsymbol{\kappa}\|^2}\mathrm{diag}\left([T_k]_{k\in[r]}\right)-\frac{\boldsymbol{K_A}}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\mathrm{diag}(\boldsymbol{\sigma}^2).\end{aligned}$$

As a corollary, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}[\boldsymbol{K_A}]_p = 2\left(1-\tilde{S}\right)\left(\frac{[\boldsymbol{K_B}]_p}{\|\boldsymbol{\kappa}\|^2}-\frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{[\boldsymbol{K_A}]_p}{\|\boldsymbol{\kappa}\|^2}\right)\sigma_p^2 - 2\left(1-\tilde{S}\right)\left(\frac{[\boldsymbol{K_B}]_p}{\|\boldsymbol{\kappa}\|^2}T_p-\frac{[\boldsymbol{K_A}]_p}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\sigma_p^2.$$

Hence,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\kappa_p^2 &= 2\left\langle [\boldsymbol{K_A}]_p, \frac{\mathrm{d}}{\mathrm{d}t}[\boldsymbol{K_A}]_p\right\rangle\\ &= 4\left(1-\tilde{S}\right)\left(\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2}-\frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2}\right)\sigma_p^2 - 2\left(1-\tilde{S}\right)\left(\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2}T_p-\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\sigma_p^2.\end{aligned}$$

By symmetry, for $\boldsymbol{K_B}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}[\boldsymbol{K_B}]_p = 2\left(1-\tilde{S}\right)\left(\frac{[\boldsymbol{K_A}]_p}{\|\boldsymbol{\kappa}\|^2}-\frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{[\boldsymbol{K_B}]_p}{\|\boldsymbol{\kappa}\|^2}\right)\sigma_p^2 - 2\left(1-\tilde{S}\right)\left(\frac{[\boldsymbol{K_A}]_p}{\|\boldsymbol{\kappa}\|^2}T_p-\frac{[\boldsymbol{K_B}]_p}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\sigma_p^2.$$

Then, we can compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\kappa}_p^2 = 4\left(1-\tilde{S}\right)\left(\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2}-\frac{\|\hat{\boldsymbol{\kappa}}\|^2}{\|\boldsymbol{\kappa}\|^2}\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2}\right)\sigma_p^2 - 4\left(1-\tilde{S}\right)\left(\frac{\kappa_p^2}{\|\boldsymbol{\kappa}\|^2}T_p-\frac{\hat{\kappa}_p^2}{\|\boldsymbol{\kappa}\|^2}\tilde{T}\right)\sigma_p^2.$$

□

B STAGE 1

In this section, we show that the following hold:

(a) $\boldsymbol{K_A} \approx \boldsymbol{K_B}$ after Stage 1.
(b) The noise-signal ratio is small after Stage 1.
(c) The condition number is $O(\sqrt{d})$ in Stage 1.
(d) The trajectory is still close to the infinite-width one in Stage 1.

We formalize the main results of Stage 1 bellow.

Lemma B.1 (Stage 1)**.** *Under the assumption of Theorem 4.3. We can choose a sufficiently (polynomially) large m and a sufficiently (inverse polynomially) small τ_t^2 that may depend on the δ's that appear in this lemma so that the following statement holds.*

Let T_1 be the earliest time all the following hold:

$$\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle \geq 1-\delta_-, \qquad \forall p\in[r],$$

$$\frac{\|[\boldsymbol{K_{A,\xi}}]_q\|}{\|[\boldsymbol{K_A}]_p\|} \leq \delta_{N/S}, \qquad \forall p\in[r], q\in[d-r],$$

where $\delta_-, \delta_{N/S} \in 1/\operatorname{poly}(d)$ *are two given parameters. We have* $T_1 \le \operatorname{poly}(d)$*. Moreover, at any time* $t \in [0, T_1]$*, we have* $\kappa_0 := \max_{p,q\in[r]} \|[\boldsymbol{K_A}]_p\| / \|[\boldsymbol{K_A}]_q\| \le O(\sqrt{d})$ *and*

$$\begin{aligned}
&\|[\boldsymbol{K_A}]_p\|^2 = (1 \pm \delta_{\boldsymbol{A/B}}) \|[\boldsymbol{K_B}]_p\|^2, \ \|[\boldsymbol{K_{A,\xi}}]_q\|^2 = (1 \pm \delta_{\boldsymbol{A/B}}) \|[\boldsymbol{K_{B,\xi}}]_q\|^2, && \forall p \in [r], q \in [d-r], \\
&\left|1 - \frac{\|[\boldsymbol{K_{A,\xi}}]_p\|}{\|[\boldsymbol{K_{A,\xi}}]_q\|}\right| \le \delta_{\xi,\kappa_0}, && \forall p, q \in [d-r], \\
&\max\left\{\left|\left\langle \overline{[\boldsymbol{K_C}]_p}, \overline{[\boldsymbol{K_D}]_q}\right\rangle\right| : \boldsymbol{C}, \boldsymbol{D} \in \{\boldsymbol{A}, \boldsymbol{B}\}\right\} \le \delta_{\boldsymbol{AB},\perp}, && \forall p \ne q \in [r], \\
&\max\left\{\left|\left\langle \overline{[\boldsymbol{K_C}]_p}, \overline{[\boldsymbol{K_{D,\xi}}]_q}\right\rangle\right|, \left|\left\langle \overline{[\boldsymbol{K_{A,\xi}}]_s}, \overline{[\boldsymbol{K_{B,\xi}}]_q}\right\rangle\right| : \boldsymbol{C}, \boldsymbol{D} \in \{\boldsymbol{A}, \boldsymbol{B}\}\right\} \le \delta_{\xi,\perp}, && \forall p \in [r], q, s \in [d-r],
\end{aligned} \tag{10}$$

where $\delta_{\boldsymbol{A/B}}, \delta_{\xi,\kappa_0}, \delta_{\boldsymbol{AB},\perp}, \delta_{\xi,\perp} \in 1/\operatorname{poly}(d)$ *are given parameters.*

Basically, the conditions in (10) mean that the norm of the corresponding columns are roughly the same, and the columns of all these matrices are approximately orthogonal to each other. Both of them are true in the infinite-width limit, and by some standard concentration argument, one can make all these errors to be arbitrarily inverse-polynomially small, at least at initialization. Note that, as a simple corollary of (10), we have

$$N_{\boldsymbol{A}} = N_{\boldsymbol{B}} \left(1 \pm \sqrt{\delta_{\boldsymbol{A/B}}}\right).$$

For notational simplicity, we also define

$$\rho_- = \max_{p\in[r]} \left\{1 - \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle\right\}, \quad \rho_{N/S} = \max_{\substack{p\in[r] \\ q\in[d-r]}} \frac{\|[\boldsymbol{K_{A,\xi}}]_q\|}{\|[\boldsymbol{K_A}]_p\|}.$$

The main tool we use to control the condition number and the discretization error is the following nonlinear version of Gronwall's lemma.

Lemma B.2. *Let* A_t *be a positive process. Let* X_t *and* Y_t *be defined as*

$$\dot{X}_t \le -A_t X_t, \quad \dot{Y}_t \le \alpha A_t X_t Y_t,$$

with X_0, Y_0, α *being positive. Then, for any* $T \ge 0$*, we have* $Y_T \le Y_0 \exp(\alpha X_0)$*.*

Remark. Here, X_t represent the progress we have made and Y_t the error. In our case, X_t is the maximum between $1 - \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle$ and the noise-signal ratio, and Y_t the discretization error. This lemma says that, if the error growth rate depends on the progress, then by coupling these two processes, we can make sure the error does not blow up. The point of this lemma is that, with coupling, we do not need a very tight estimation on the convergence time nor the error growth rate. ♣

Proof. The solution of this ODE system is given by

$$X_T = X_0 \exp\left(-\int_0^T A_t \,\mathrm{d}t\right), \quad Y_T = Y_0 \exp\left(\alpha X_0 \int_0^T A_t \exp\left(-\int_0^t A_s \,\mathrm{d}s\right) \mathrm{d}t\right).$$

Note that

$$\int_0^T A_t \exp\left(-\int_0^t A_s \,\mathrm{d}s\right) \mathrm{d}t = -\int_0^T \mathrm{d} \exp\left(-\int_0^t A_s \,\mathrm{d}s\right) = 1 - \exp\left(-\int_0^T A_t \,\mathrm{d}t\right).$$

Hence,

$$Y_T = Y_0 \exp\left(\alpha X_0 \left(1 - \exp\left(-\int_0^T A_t \,\mathrm{d}t\right)\right)\right) \le Y_0 \exp\left(\alpha X_0\right).$$

□

The organization of this section is as follows. In Section B.1, we derive estimations for the $\boldsymbol{Q}$-matrices defined in Corollary A.4 and use them to simplify the equations governing the training dynamics. In Section B.2, we estimate the rate at which $1 - \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p} \right\rangle$ and the noise-signal ratio converge to 0 and the growth rate of the condition number. In Section B.3, we estimate the growth rate of the discretization error. Then, in Section B.4, we prove Lemma B.1, the main lemma of Stage 1. Finally, we prove the negative result for non-contrastive learning in Section B.5.

B.1 Estimations for Q and the dynamics

Thanks to Lemma A.5, in order to analyze the dynamics, it suffices to estimate the $\boldsymbol{Q}$-matrices. In this subsection, we derive estimations for them and use these estimations to simplify the equations in Lemma A.5. Recall that, in Stage 1, τ_t^2 is small. Hence, $\exp(\tau_t^2 \langle \boldsymbol{f_A^+}, \boldsymbol{f_B^-} \rangle) = 1 \pm O(\tau_t^2)$. With this approximation, one can derive the following estimation for $S_{\boldsymbol{A}}$ and $S_{\boldsymbol{B}}$.

Lemma B.3 (Estimations for S)**.** *In Stage 1, we have*

$$S_{\boldsymbol{A}} = \frac{1}{1+K} \pm O_z\left(\tau_t^2\right) \pm O\left(\tau_t^2 d \delta_{\xi,\perp} \rho_{N/S}\right),$$

and the same is also true for $S_{\boldsymbol{B}}$. Here, $O_z(\tau_t^2)$ means $O(\tau_t^2)$ and this does not depend on $\boldsymbol{\xi_A}$ nor $\boldsymbol{\xi_B}$.

The proof of this lemma and all following lemma is deferred to the end of this subsection. Note that we derive a slight finer estimations for the noise-related part. This additional $\rho_{N/S}$ will be used cancel with terms like $\|[\boldsymbol{K_A}]_p\| \,/\, \|[\boldsymbol{K_{A,\xi}}]_q\|$, at the cost of a κ_0 factor, in later analysis. We emphasize here that O_z does not depend on the noises so that later we can argue $\mathbb{E}_{\boldsymbol{\xi}}\left\{O_z(\tau_t^2)\boldsymbol{\xi}\right\} = 0$. With this lemma, we now derive estimations for $\boldsymbol{Q}_1$, $\boldsymbol{Q}_{1,\boldsymbol{\xi}}$, $\boldsymbol{Q}_2$ and Q_0, respectively.

Lemma B.4 (Estimations for $\boldsymbol{Q}_1$)**.** *In Stage 1, we have*

$$\boldsymbol{Q}_1 = \frac{2K}{1+K}\boldsymbol{I}_d \pm O\left(d\tau_t^2\right).$$

Lemma B.5 (Estimations for $\boldsymbol{Q}_{1,\boldsymbol{\xi}}$ and $\boldsymbol{Q}_2$)**.** *In Stage 1, we have*

$$\max\left\{\|\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}\|_F, \|\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}\|_F, \|\boldsymbol{Q}_2\|_F\right\} = \pm O\left(\tau_t^2 d^2 \rho_{N/S} \delta_{\xi,\perp}\right).$$

Lemma B.6 (Estimations for Q_0)**.** *In Stage 1, we have*

$$Q_0 = -\frac{2K}{1+K}\frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(d\tau_t^2\right).$$

The proof of Lemma B.6 is essentially the same as the proof of Lemma B.4 so we omit it. With these three lemmas, we can now simplify Lemma A.5 as follows.

Corollary B.7. *In Stage 1, for any $p \in [r]$ and $q \in [d-r]$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K_A}]_p\|^2 = \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(\frac{\langle [\boldsymbol{K_A}]_p, [\boldsymbol{K_B}]_p\rangle}{\|[\boldsymbol{K_A}]_p\|^2} - \frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\right)\|[\boldsymbol{K_A}]_p\|^2$$
$$\pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}} + \tau_t^2 d\right)\kappa_p^2\right),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K_{A,\xi}}]_q\|^2 = -\frac{4K}{1+K}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\|[\boldsymbol{K_{A,\xi}}]_q\|^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\|[\boldsymbol{K_{A,\xi}}]_q\|^2\right),$$

The formulas for $\boldsymbol{K_B}$ and $\boldsymbol{K_{B,\xi}}$ can be obtained by interchanging the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Corollary B.8. *In Stage 1, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K_A}]_p} = \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(\boldsymbol{I} - \overline{[\boldsymbol{K_A}]_p}\left(\overline{[\boldsymbol{K_A}]_p}\right)^\top\right)\overline{[\boldsymbol{K_B}]_p} \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(\tau_t^2 d^2 \kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K_{A,\xi}}]_q} = \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\tau_t^2 d^3 \kappa_0 \delta_{\xi,\perp}\right).$$

Interchange the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$ and one can obtain the formulas for $\boldsymbol{K_B}$ and $\boldsymbol{K_{B,\xi}}$.

Note that the above results also imply the following lemma.

Lemma B.9. *The dynamics of the non-contrastive method and the Stage 1 dynamics are equivalent, up to a multiplicative constant and some higher order terms.*

OMITTED PROOFS OF THIS SUBSECTION

Proof of Lemma B.3. First, we write

$$\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle = \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^{+}, \boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} + \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^{+}, \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{B}}^{-} \right\rangle + \left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{A}}^{+}, \boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} + \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{A}}^{+}, \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{B}}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}.$$

For the second term, we write

$$\frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^{+}, \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{B}}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} = \sum_{i \in [r], j \in [d-r]} \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_i\|^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \frac{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_j\|}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_i\|} \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_i}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_j} \right\rangle z_i^{+} \xi_{\boldsymbol{B},j}^{-} d.$$

For each summand, we have $\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_j\| / \|[\boldsymbol{K}_{\boldsymbol{A}}]_i\| \leq O(\rho_{N/S})$, $\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_i}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_j} \right\rangle = \pm O(\delta_{\xi,\perp})$, and $|z_i^{+} \xi_{\boldsymbol{B},j}^{-} d| \leq 1$. Hence, $\frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^{+}, \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{B}}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} = \pm O(d \delta_{\xi,\perp} \rho_{N/S})$. The same is also true for $\frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{A}}^{+}, \boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}$ and the third term. Therefore,

$$\left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle = \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^{+}, \boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} \pm O\left(d \delta_{\xi,\perp} \rho_{N/S}\right).$$

Then, we compute

$$\begin{aligned} \exp\left(\tau_t^2 \left\langle \boldsymbol{f}_{\boldsymbol{A}}^{+}, \boldsymbol{f}_{\boldsymbol{B}}^{-} \right\rangle\right) &= \exp\left(\tau_t^2 \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^{+}, \boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^{-} \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right) \left(1 \pm O\left(\tau_t^2 d \delta_{\xi,\perp} \rho_{N/S}\right)\right) \\ &= 1 \pm O_z\left(\tau_t^2\right) \pm O\left(\tau_t^2 d \delta_{\xi,\perp} \rho_{N/S}\right). \end{aligned}$$

Thus,

$$S_{\boldsymbol{A}} = \frac{1}{1+K} \pm O_z\left(\tau_t^2\right) \pm O\left(\tau_t^2 d \delta_{\xi,\perp} \rho_{N/S}\right).$$

The proof for $S_{\boldsymbol{B}}$ is essentially the same. □

Proof of Lemma B.4. Recall that

$$\begin{aligned} \boldsymbol{Q}_1 := \mathbb{E}\left\{\left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+})\right) \boldsymbol{z}^{+} (\boldsymbol{z}^{+})^{\top} d\right\} \\ - K\, \mathbb{E}\left\{\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{-})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \boldsymbol{z}^{-} (\boldsymbol{z}^{+})^{\top} d\right\} \\ - K\, \mathbb{E}\left\{\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) \exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{-} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^{+} \cdot \boldsymbol{f}_{\boldsymbol{B}}^{+})} \boldsymbol{z}^{+} (\boldsymbol{z}^{-})^{\top} d\right\}. \end{aligned}$$

Since $S_{\boldsymbol{A}} = (1+K)^{-1} \pm O(\tau_t^2)$ and $S_{\boldsymbol{B}} = (1+K)^{-1} \pm O(\tau_t^2)$, we have

$$2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^{+}, \boldsymbol{x}_{\boldsymbol{B}}^{+}) = 2 - \frac{2}{1+K} \pm O(\tau_t^2) = \frac{2K}{1+K} \pm O(\tau_t^2).$$

As a result,

$$\boldsymbol{Q}_1 = \frac{2K}{1+K} \mathbb{E}\left\{\boldsymbol{z}^{+} (\boldsymbol{z}^{+})^{\top} d\right\} - \frac{K}{1+K} \mathbb{E}\left\{\boldsymbol{z}^{-} (\boldsymbol{z}^{+})^{\top} d\right\} - \frac{K}{1+K} \mathbb{E}\left\{\boldsymbol{z}^{+} (\boldsymbol{z}^{-})^{\top} d\right\} \pm O\left(d \tau_t^2\right).$$

Note that $\mathbb{E}\left\{\boldsymbol{z}^{-} (\boldsymbol{z}^{+})^{\top}\right\} = 0$ and $\mathbb{E}\left\{\boldsymbol{z}^{+} (\boldsymbol{z}^{+})^{\top}\right\} = \boldsymbol{I}_d / d$. Therefore,

$$\boldsymbol{Q}_1 = \frac{2K}{1+K} \boldsymbol{I}_d \pm O\left(d \tau_t^2\right).$$

□

Proof of Lemma B.5. Recall that

$$\begin{aligned}\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}} := \mathbb{E}\left\{\left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \boldsymbol{\xi}_{\boldsymbol{A}}^+(\boldsymbol{z}^+)^\top d\right\} \\ - K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{A}}^+(\boldsymbol{z}^-)^\top d\right\} \\ - K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2 \boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{\xi}_{\boldsymbol{A}}^-(\boldsymbol{z}^+)^\top d\right\}.\end{aligned}$$

Note that if some quantity X does not depend on $\boldsymbol{\xi}$, then $\mathbb{E}\{X\boldsymbol{\xi}\} = 0$. Hence, by Lemma B.3, we have

$$\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}} = \pm O\left(\tau_t^2 d^2 \rho_{N/S} \delta_{\xi,\perp}\right).$$

The proof for $\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}$ and $\boldsymbol{Q}_2$ is essentially the same. □

Proof of Corollary B.7. Recall that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 &= 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_p^2 \\ &= \sum_{i=1}^{3}\mathrm{T}_i\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\right).\end{aligned}$$

We now estimate these terms one-by-one. For T_1, we have

$$\begin{aligned}\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p\rangle &= \langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle [\boldsymbol{Q}_1]_{p,p} + \sum_{k\neq p}\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_k\rangle [\boldsymbol{Q}_1]_{k,p} \\ &= \frac{2K}{1+K}\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle \pm O\left(\tau_t^2 d\kappa_p^2\right) \pm O\left(\tau_t^2 d\kappa_p^2\kappa_0\delta_{\boldsymbol{AB},\perp}\right) \\ &= \frac{2K}{1+K}\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle \pm O\left(\tau_t^2 d\kappa_p^2\right).\end{aligned}$$

Hence,

$$\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\right) = \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d\kappa_p^2\right).$$

For T_2, we compute

$$\begin{aligned}\mathrm{T}_2\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\right) &= \frac{2\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{d-r}\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k\rangle [\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{k,p} \\ &= \pm O(1)\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{d-r}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k\right\|}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k}\right\rangle [\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{k,p} \\ &= \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\rho_{N/S}^2\delta_{\xi,\perp}^2\kappa_p^2\right).\end{aligned}$$

Finally, for T_3, we compute

$$\begin{aligned}\mathrm{T}_3\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\right) &= 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{N_{\boldsymbol{A}}^2 d}\left(-\frac{2K}{1+K}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}}, \boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \pm O\left(d\tau_t^2\right)\right)\sigma_p^2 \\ &= -\frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1 \pm \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\frac{\langle \boldsymbol{K}_{\boldsymbol{A}}, \boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 \pm O\left(\frac{\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d\sigma_p^2\right) \\ &= -\frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}}, \boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}} + \tau_t^2 d\right)\kappa_p^2\right).\end{aligned}$$

Combine these together and we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 &= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\rangle - \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 \\ &\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\rho_{N/S}^2\delta_{\xi,\perp}^2\kappa_p^2\right) \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}+\tau_t^2 d\right)\kappa_p^2\right) \\ &= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\frac{\left\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\rangle}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2} - \frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 \\ &\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}+\tau_t^2 d\right)\kappa_p^2\right).\end{aligned}$$

Interchange the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$ and we obtain the formula for $\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\|^2$. Now we consider $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}$. We write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2 &= 2\frac{\left\langle[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q,[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top]_q\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q,[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_2]_q\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2 \\ &= \sum_{i=1}^{3}\mathrm{T}_i\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\right).\end{aligned}$$

For T_1, we compute

$$\begin{aligned}\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\right) &= 2\sum_{k=1}^{r}\frac{\langle[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q,[\boldsymbol{K}_{\boldsymbol{B}}]_k\rangle\,[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,k}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 \\ &= \frac{2\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{r}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\right\|}{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\right\rangle[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,k} \\ &= \pm O(1)\frac{2\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\sum_{k=1}^{r}\frac{\kappa_0}{\rho_{N/S}}\delta_{\xi,\perp}\tau_t^2 d^2\rho_{N/S}\delta_{\xi,\perp} \\ &= \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\kappa_0\delta_{\xi,\perp}^2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\right).\end{aligned}$$

Similarly, one can show that this bound also holds for T_2. Finally, for T_3, by Lemma B.6, we have

$$\begin{aligned}\mathrm{T}_3\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\right) &= 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1\pm\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\left(-\frac{2K}{1+K}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\pm O\left(d\tau_t^2\right)\right)\sigma_\xi^2 \\ &= -\frac{4K}{1+K}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2 \\ &\quad \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2+\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2\right).\end{aligned}$$

Combine these together and we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2 = -\frac{4K}{1+K}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2+\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2\right).$$

Interchange the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$ and we obtain the formula for $\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}$. □

Proof of Corollary B.8. We write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} &= \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &= \mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right) + \mathrm{T}_2\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right).\end{aligned}$$

For T_1, we have

$$\begin{aligned}\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right) &= \sum_{k=1}^{r}\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\|\,[\boldsymbol{Q}_1]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &= \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|\,[\boldsymbol{Q}_1]_{p,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \sum_{k\neq p}\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\|\,[\boldsymbol{Q}_1]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}.\end{aligned}$$

For the first term, by Lemma B.4, we have

$$\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|\,[\boldsymbol{Q}_1]_{p,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} = \left(1 \pm \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\left(\frac{2K}{1+K} \pm O\left(d\tau_t^2\right)\right) = \frac{2K}{1+K} \pm O\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right).$$

Also by Lemma B.4, for each summand in the second term, we have

$$\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\|\,[\boldsymbol{Q}_1]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} = \pm O\left(\tau_t^2 d\kappa_0\right).$$

Therefore,

$$\begin{aligned}\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right) &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p} \\ &\quad \pm \frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{r} O\left(\tau_t^2 d\kappa_0 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}.\end{aligned}$$

For T_2, by Lemma B.5, we have

$$\begin{aligned}\mathrm{T}_2\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right) &= \frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{d-r}\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k}\frac{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k\|\,[\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \\ &= \frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{d-r} O\left(\tau_t^2 d^2\rho_{N/S}^2\delta_{\xi,\perp}\right)\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k}.\end{aligned}$$

Combine these together, and we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p} \\ &\quad \pm \frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{r} O\left(\tau_t^2 d\kappa_0 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k} \pm \frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sum_{k=1}^{d-r} O\left(\tau_t^2 d^2\rho_{N/S}^2\delta_{\xi,\perp}\right)\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k} \\ &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p} \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).\end{aligned}$$

Now, we consider $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}$. Again, we write

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q} &= \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_2]_q}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &=: \mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right) + \mathrm{T}_2\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right).\end{aligned}$$

For the first term, by Lemma B.5, we have

$$\begin{aligned}
\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right) &= \sum_{k=1}^{r}\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\|\,[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\
&= \pm\sum_{k=1}^{r} O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\kappa_0}{\rho_{N/S}}\tau_t^2 d^2\rho_{N/S}\delta_{\xi,\perp}\right) \\
&= \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\kappa_0\delta_{\xi,\perp}\right).
\end{aligned}$$

The same bound also hold for T_2. In fact, we can have a slightly sharper bound for it because we no longer have $\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\| \,/\, \|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|$. Combine these and we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q} = \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\kappa_0\delta_{\xi,\perp}\right).$$

□

Proof of Lemma B.9. The proof of Corollary B.8 and Corollary B.7, *mutatis mutandis*, yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A}} &= \frac{2K}{1+K}\frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\operatorname{diag}(\boldsymbol{\sigma}^2) - \frac{2K}{1+K}\frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\operatorname{diag}(\boldsymbol{\sigma}^2) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\|\boldsymbol{K}_{\boldsymbol{A}}\|_F\right), \\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} &= -\frac{2K}{1+K}\frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}^2 d}d\tau_t^2\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F\right).
\end{aligned}$$

Recall from Lemma A.6 that, in the non-contrastive case, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A}} &= \mathbb{E}\left\{\frac{\boldsymbol{f}_{\boldsymbol{B}}^+(\boldsymbol{z}^+)^\top}{N_{\boldsymbol{A}}} - \langle\boldsymbol{f}_{\boldsymbol{A}}^+,\boldsymbol{f}_{\boldsymbol{B}}^+\rangle\frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d}\right\}\operatorname{diag}(\boldsymbol{\sigma}^2) \\
&= \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\operatorname{diag}(\boldsymbol{\sigma}^2) - \frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d}\operatorname{diag}(\boldsymbol{\sigma}^2), \\
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} &= -\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d}\sigma_\xi^2.
\end{aligned}$$

Note that they are exactly the same, except for a $2K/(1+K)$ factor and some higher order error terms. □

B.2 Convergence rate and the condition number

In this subsection, we estimate the rate at which $1-\rho_-$ and $\rho_{N/S}$ converge to 0 and the growth rate of the condition number.

Lemma B.10 (Convergence rate of ρ_-)**.** *In Stage 1, we have, for any $p \in [r]$,*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle &= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1+\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right)\left(1-\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Now we consider the noise-signal ratio. For some technical reason, instead of characterizing the dynamics of $\rho_{N/S}$, we consider

$$\hat{\rho}_{N/S} := \frac{\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F}{\|\boldsymbol{K}_{\boldsymbol{A}}+\boldsymbol{K}_{\boldsymbol{B}}\|_F}.$$

Note that we always have

$$\hat{\rho}_{N/S}^2 \geq \Theta(1) \frac{\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\|_F^2}{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2} \geq \Theta(1) \frac{(d-r)\,\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\kappa_0^2 r\,\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}$$

In other words, $\rho_{N/S}$ can be bounded by $\rho_{N/S} \leq O\left(\frac{d\kappa_0}{r}\hat{\rho}_{N/S}\right)$.

Lemma B.11 (Convergence rate of $\rho_{N/S}$)**.** *In Stage 1, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\rho}_{N/S}^2 \leq -\frac{4K}{1+K}\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\hat{\rho}_{N/S}^2 + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\hat{\rho}_{N/S}^2\right).$$

Lemma B.12 (Growth rate of the condition number)**.** *Define* $\rho_{p/q} = \|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 / \|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|^2$. *In Stage 1, we have*

$$\dot{\rho}_{p/q} \leq \frac{16K}{1+K}\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\rho_- + \min\left\{\hat{\rho}_{N/S}, 1\right\}\right)\rho_{p/q} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}} + \tau_t^2 d\right)\rho_{p/q}\right).$$

OMITTED PROOFS OF THIS SUBSECTION

Proof of Lemma B.10. By Corollary B.8, we have

$$\begin{aligned}
\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle^2\right) \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right) \\
&= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right)\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Hence, by symmetry, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle &= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right)\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

□

Proof of Lemma B.11. Similar to the proof of Corollary B.7 and Corollary B.8, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A}} &= \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\boldsymbol{Q}_1\,\mathrm{diag}(\boldsymbol{\sigma}^2) + \frac{\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}\,\mathrm{diag}(\boldsymbol{\sigma}^2) + \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d}Q_0\,\mathrm{diag}(\boldsymbol{\sigma}^2) \\
&= \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\frac{2K}{1+K}\boldsymbol{I}_d \pm O\left(d\tau_t^2\right)\right)\mathrm{diag}(\boldsymbol{\sigma}^2) \\
&\quad \pm \frac{\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}O\left(\tau_t^2 d^2\rho_{N/S}\delta_{\xi,\perp}\right)\mathrm{diag}(\boldsymbol{\sigma}^2) \\
&\quad + \frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}^2 d}\left(-\frac{2K}{1+K}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}}, \boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \pm O\left(d\tau_t^2\right)\right)\mathrm{diag}(\boldsymbol{\sigma}^2) \\
&= \frac{2K}{1+K}\frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\,\mathrm{diag}(\boldsymbol{\sigma}^2) - \frac{2K}{1+K}\frac{\boldsymbol{K}_{\boldsymbol{A}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}}, \boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\,\mathrm{diag}(\boldsymbol{\sigma}^2) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\|\boldsymbol{K}_{\boldsymbol{A}}\|_F\right).
\end{aligned}$$

Define $\boldsymbol{K} = \boldsymbol{K}_{\boldsymbol{A}} + \boldsymbol{K}_{\boldsymbol{B}}$. Then, by symmetry, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K} = \frac{2K}{1+K}\frac{1}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1 - \frac{\langle \boldsymbol{K}_{\boldsymbol{A}}, \boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\boldsymbol{K}\,\mathrm{diag}(\boldsymbol{\sigma}^2) \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\|\boldsymbol{K}\|_F\right).$$

Hence,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|\boldsymbol{K}\right\|_F^2 &= \frac{4K}{1+K}\frac{1}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1-\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\left\langle \boldsymbol{K},\boldsymbol{K}\operatorname{diag}(\boldsymbol{\sigma}^2)\right\rangle \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2+\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\left\|\boldsymbol{K}\right\|_F^2\right)\\ &\geq \frac{4K}{1+K}\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1-\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\left\|\boldsymbol{K}\right\|_F^2 - O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2+\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\left\|\boldsymbol{K}\right\|_F^2\right).\end{aligned}$$

Similarly, we also have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} &= \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top\sigma_\xi^2 + \frac{\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\boldsymbol{Q}_2\sigma_\xi^2 + \frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2\\ &= \frac{\boldsymbol{K}_{\boldsymbol{B}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}O\left(\tau_t^2 d^2\rho_{N/S}\delta_{\xi,\perp}\right)\sigma_\xi^2 + \frac{\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}O\left(\tau_t^2 d^2\rho_{N/S}\delta_{\xi,\perp}\right)\sigma_\xi^2\\ &\quad + \frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d}\left(-\frac{2K}{1+K}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\pm O\left(d\tau_t^2\right)\right)\sigma_\xi^2\\ &= -\frac{2K}{1+K}\frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}^2 d}d\tau_t^2\left\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\right\|_F\right) \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^2\delta_{\xi,\perp}\frac{d\kappa_0}{r}\left\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\right\|_F\right)\\ &= -\frac{2K}{1+K}\frac{\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}}{N_{\boldsymbol{A}}^2 d}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}^2 d}d\tau_t^2\left\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\right\|_F\right).\end{aligned}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\right\|_F^2 = -\frac{4K}{1+K}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}^2 d}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\right\|_F^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}^2 d}d\tau_t^2\left\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\right\|_F^2\right).$$

Then, we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\hat{\rho}_{N/S}^2 &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\right\|_F^2}{\left\|\boldsymbol{K}\right\|_F^2} - \hat{\rho}_{N/S}^2\frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|\boldsymbol{K}_{\boldsymbol{A}}\right\|_F^2}{\left\|\boldsymbol{K}\right\|_F^2}\\ &\leq -\frac{4K}{1+K}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}^2 d}\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\hat{\rho}_{N/S}^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}^2 d}d\tau_t^2\hat{\rho}_{N/S}^2\right)\\ &\quad - \hat{\rho}_{N/S}^2\left(\frac{4K}{1+K}\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1-\frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right) - O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2+\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right)\right)\\ &\leq -\frac{4K}{1+K}\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\hat{\rho}_{N/S}^2 + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2+\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\hat{\rho}_{N/S}^2\right).\end{aligned}$$

□

Proof of Lemma B.12. For notational simplicity, define $\rho_{p/q} := \left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 / \left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2$. By Corollary B.7, we have

$$\begin{aligned}\dot{\rho}_{p/q} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} - \rho_{p/q}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2}\\ &= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2} - \frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_{p/q} \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}+\tau_t^2 d\right)\rho_{p/q}\right)\\ &\quad - \rho_{p/q}\left(\frac{4K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_q,[\boldsymbol{K}_{\boldsymbol{B}}]_q\rangle}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} - \frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right) \pm O\left(\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}+\tau_t^2 d\right)\right)\right)\\ &= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle - \frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_{p/q}\\ &\quad - \frac{4K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle - \frac{\langle \boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_{p/q}\\ &\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}+\tau_t^2 d\right)\rho_{p/q}\right).\end{aligned}$$

Now we bound $\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p} \right\rangle - \langle \boldsymbol{K_A}, \boldsymbol{K_B} \rangle / (N_{\boldsymbol{A}} N_{\boldsymbol{B}} d)$. Clear that this term is bounded by 2. Meanwhile, by definition, we have $\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p} \right\rangle = 1 \pm \rho_-$. For the second term, we have

$$\begin{aligned}
\frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B} \rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} &= \sum_{k=1}^{r} \left\langle \overline{[\boldsymbol{K_A}]_k}, \overline{[\boldsymbol{K_B}]_k} \right\rangle \frac{\|\boldsymbol{K_A}\|_k \|\boldsymbol{K_B}\|_k}{\|\boldsymbol{K_A}\|_F \|\boldsymbol{K_B}\|_F} \frac{\|\boldsymbol{K_A}\|_F \|\boldsymbol{K_B}\|_F}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \\
&= \sum_{k=1}^{r} (1 \pm \rho_-) \left(1 \pm \min\left\{\hat{\rho}_{N/S}, 1\right\}\right) \frac{\kappa_k^2}{\|\boldsymbol{\kappa}\|^2} \left(1 \pm \sqrt{\delta_{A/\boldsymbol{B}}}\right) \\
&= \left(1 \pm \rho_- \pm \min\left\{\hat{\rho}_{N/S}, 1\right\}\right) \left(1 \pm \sqrt{\delta_{A/\boldsymbol{B}}}\right).
\end{aligned}$$

Combine these together and we obtain

$$\left| \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p} \right\rangle - \frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B} \rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \right| \leq 2\rho_- + 2 \min\left\{\hat{\rho}_{N/S}, 1\right\} \pm \sqrt{\delta_{A/\boldsymbol{B}}}.$$

The same is also true for q. Thus,

$$\dot{\rho}_{p/q} \leq \frac{16K}{1+K} \frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \left(\rho_- + \min\left\{\hat{\rho}_{N/S}, 1\right\}\right) \rho_{p/q} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \left(\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}} + \tau_t^2 d\right) \rho_{p/q}\right).$$

□

B.3 Controlling the discretization error

In this subsection, we estimate the growth rate of the errors described in (10). As in the previous subsections, the proofs are deferred to the end of this subsection.

First, we consider the relative difference between $\|[\boldsymbol{K_A}]_p\|^2$ and $\|[\boldsymbol{K_B}]_p\|^2$. Instead of directly control the difference, we define

$$\rho_{\boldsymbol{A}/\boldsymbol{B},p} := \frac{\|[\boldsymbol{K_A}]_p\|^2}{\|[\boldsymbol{K_B}]_p\|^2} \quad \text{and} \quad \rho_{\boldsymbol{B}/\boldsymbol{A},p} := \frac{\|[\boldsymbol{K_B}]_p\|^2}{\|[\boldsymbol{K_A}]_p\|^2}.$$

Note that $\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \rho_{\boldsymbol{B}/\boldsymbol{A},p} \geq 2$, with equality obtained iff $\|[\boldsymbol{K_A}]_p\|^2 = \|[\boldsymbol{K_B}]_p\|^2$. Meanwhile, at initialization, this quantity can be made arbitrarily close to 2. Hence, it suffices to control the growth of this quantity. The reason we consider $\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \rho_{\boldsymbol{B}/\boldsymbol{A},p}$ is to leverage the symmetry. Similarly, we also define

$$\rho_{\boldsymbol{A}/\boldsymbol{B},\xi,q} := \frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|^2} \quad \text{and} \quad \rho_{\boldsymbol{B}/\boldsymbol{A},\xi,q} := \frac{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2},$$

and analyze $\rho_{\boldsymbol{A}/\boldsymbol{A},\xi} + \rho_{\boldsymbol{B}/\boldsymbol{A},\xi}$.

Lemma B.13 (Difference of diagonal terms)**.** *In Stage 1, we have*

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) &\leq O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(\delta_{\boldsymbol{A}/\boldsymbol{B}}^2 + \tau_t^2 d\right)\right), \\
\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\boldsymbol{A}/\boldsymbol{B},\xi,q} + \rho_{\boldsymbol{B}/\boldsymbol{A},\xi,q}\right) &\leq O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(\delta_{\boldsymbol{A}/\boldsymbol{B}}^2 + \tau_t^2 d\right)\right).
\end{aligned}$$

Now we consider the condition number of $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}$. Unlike $\boldsymbol{K_A}$, for the noises, the σ's for different coordinates are the same. Hence, we have the following simple bound on the growth rate of δ_{ξ,κ_0}.

Lemma B.14 (Condition number of $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}$)**.** *Define* $\rho_{\xi,p/q} := \|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p\|^2 / \|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2$. *In Stage 1, we have*

$$\dot{\rho}_{\xi,p/q} \leq O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).$$

Then, we consider the orthogonality conditions.

Lemma B.15 (Orthogonality between signals)**.** *For any $p \neq q$, define*

$$\hat{\delta}_{\perp,p,q} := \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_A}]_q} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_A}]_q} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_B}]_q} \right\rangle^2 .$$

In Stage 1, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\perp,p,q} \leq O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\hat{\delta}_{\perp,p,q} + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).$$

Recall that ρ_- converges to 0 at a sufficiently rate so that, by Lemma B.2, $\hat{\delta}_{\perp,p,q}$ will not blow up. Meanwhile, since $\hat{\delta}_{\perp,p,q}$ can be made arbitrarily small at initialization, this implies that we can make sure it is still small at the end of Stage 1. Finally, we consider the orthogonality conditions between the signals and noises and between noises. The proof follows the same spirit.

Lemma B.16 (Orthogonality between signals and noises)**.** *For any $p \in [r]$ and $q \in [d-r]$, define*

$$\hat{\delta}_{\perp,\xi_{\boldsymbol{A}},p,q} = \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_{A,\xi}}]_q} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_{A,\xi}}]_q} \right\rangle^2 ,$$

and define $\hat{\delta}_{\perp,\xi_{\boldsymbol{B}},p,q}$ similarly. In Stage 1, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\perp,\xi_{\boldsymbol{A}},p,q} \leq O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\hat{\delta}_{\perp,\xi_{\boldsymbol{A}},p,q} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).$$

For any $q, s \in [d-r]$, define

$$\hat{\delta}_{\perp,\xi,p,q} = \left\langle \overline{[\boldsymbol{K_{A,\xi}}]_q}, \overline{[\boldsymbol{K_{B,\xi}}]_s} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K_{B,\xi}}]_q}, \overline{[\boldsymbol{K_{A,\xi}}]_s} \right\rangle^2 ,$$

In Stage 1, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\perp,\xi,p,q} \leq O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\kappa_0\delta_{\xi,\perp}\sqrt{\hat{\delta}_{\perp,\xi,p,q}}\right).$$

OMITTED PROOFS OF THIS SUBSECTION

Proof of Lemma B.13 . Note that we cannot directly use Corollary B.7 as the error term contains $\delta_{\boldsymbol{A}/\boldsymbol{B}}$, the quantity we wish to control. However, by the proof of it, we still have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K_A}]_p\right\|^2 = \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle[\boldsymbol{K_A}]_p, [\boldsymbol{K_B}]_p\right\rangle - \frac{4K}{1+K}\frac{\left\langle \boldsymbol{K_A}, \boldsymbol{K_B}\right\rangle\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\left\|[\boldsymbol{K_A}]_p\right\|^2}{N_{\boldsymbol{A}}^2 d} \\ \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d\kappa_p^2\right).\end{aligned}$$

Interchange the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$ and we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K_B}]_p\right\|^2 = \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle[\boldsymbol{K_A}]_p, [\boldsymbol{K_B}]_p\right\rangle - \frac{4K}{1+K}\frac{\left\langle \boldsymbol{K_A}, \boldsymbol{K_B}\right\rangle\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\left\|[\boldsymbol{K_B}]_p\right\|^2}{N_{\boldsymbol{B}}^2 d} \\ \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d\kappa_p^2\right).\end{aligned}$$

For notational simplicity, define $\rho_{\boldsymbol{A}/\boldsymbol{B},p} = \|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 / \|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2$. Then, we compute

$$
\begin{aligned}
\dot{\rho}_{\boldsymbol{A}/\boldsymbol{B},p} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2} - \rho_{\boldsymbol{A}/\boldsymbol{B},p}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2} \\
&= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}\rho_{\boldsymbol{A}/\boldsymbol{B},p} - \frac{4K}{1+K}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\rho_{\boldsymbol{A}/\boldsymbol{B},p}}{N_{\boldsymbol{A}}^2 d} \\
&\quad - \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \frac{4K}{1+K}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\rho_{\boldsymbol{A}/\boldsymbol{B},p}}{N_{\boldsymbol{B}}^2 d} \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d\right) \\
&= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle\left(\frac{1}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2} - \frac{1}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\right)\rho_{\boldsymbol{A}/\boldsymbol{B},p} \\
&\quad - \frac{4K}{1+K}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\frac{1}{N_{\boldsymbol{A}}^2 d} - \frac{1}{N_{\boldsymbol{B}}^2 d}\right)\rho_{\boldsymbol{A}/\boldsymbol{B},p} \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d\right).
\end{aligned}
$$

Then, by symmetry, we have

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) \\
&= \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle\left(\frac{1}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2} - \frac{1}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\right)\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p} - \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) \\
&\quad - \frac{4K}{1+K}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\frac{1}{N_{\boldsymbol{A}}^2 d} - \frac{1}{N_{\boldsymbol{B}}^2 d}\right)\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p} - \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d\right) \\
&\le O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\delta_{\boldsymbol{A}/\boldsymbol{B}}^2 + \tau_t^2 d\right)\right).
\end{aligned}
$$

The above proof, *mutatis mutandis*, yields the result for $\rho_{\boldsymbol{A}/\boldsymbol{B},\xi,q} + \rho_{\boldsymbol{B}/\boldsymbol{A},\xi,q}$. □

Proof of Lemma B.14. By Corollary B.7, we have

$$
\begin{aligned}
\dot{\rho}_{\xi,p/q} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2} - \rho_{\xi,p/q}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2} \\
&= -\frac{4K}{1+K}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\rho_{\xi,p/q} \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right) \\
&\quad - \rho_{\xi,p/q}\left(-\frac{4K}{1+K}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle\boldsymbol{K}_{\boldsymbol{A}},\boldsymbol{K}_{\boldsymbol{B}}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right)\right) \\
&= \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}
$$

□

Proof of Lemma B.15. By Corollary B.8, we have

$$\begin{aligned}
\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p} \right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right) \\
&= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle\right) \\
&\quad + \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p} \right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Interchange the roles of p, q and $\boldsymbol{A}, \boldsymbol{B}$ and we obtain

$$\begin{aligned}
\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle\right) \\
&\quad + \frac{2K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle\right) \\
&\quad + \frac{2K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Interchange the roles of p and q and we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle\right) \\
&\quad + \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Similarly, we compute

$$\begin{aligned}
\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right) \\
&= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle\right) \\
&\quad + \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right) \\
&= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Then, by interchanging the roles of p and q, we obtain

$$\begin{aligned}
\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Add them together and we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle\right) \\
&\quad + \frac{2K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Interchange the roles of p and q and we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle &= \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle\right) \\
&\quad + \frac{2K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle\right) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

For notational simplicity, define $Z_1 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle$, $Z_2 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle$, $Z_3 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle$, and $Z_4 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle$. Also define $G_p = \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}$. Then, we can summarize the above equations as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} &= \begin{bmatrix} -G_p - G_q & 0 & G_q & G_p \\ 0 & -G_p - G_q & G_p & G_q \\ G_q & G_p & -G_p - G_q & 0 \\ G_p & G_q & 0 & -G_p - G_q \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

Note that the eigenvalues of the first matrix is $-2G_p - 2G_q$, $-2G_p$, $-2G_q$ and 0. Namely, it is negative semi-definite. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|\boldsymbol{Z}\right\|^2 \leq O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\|\boldsymbol{Z}\right\|^2 + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).$$

□

Proof of Lemma B.16. By Corollary B.8,

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle = \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right),$$

and

$$\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle = \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\kappa_0\delta_{\xi,\perp}\right).$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle = \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).$$

Similarly, we also have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle = \frac{2K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).$$

Thus,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2\right)\\
=& -\frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2\right)^2\\
&\pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2\right)\\
&\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right)\\
\leq& O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_-\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle^2\right)\\
&\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right).
\end{aligned}$$

For the orthogonality between noises, by Corollary B.8, we have

$$\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_s}\right\rangle = \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\kappa_0\delta_{\xi,\perp}\right)$$

Clear that this bound holds for all other combinations. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\perp,\xi,p,q} \leq O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 d^3\kappa_0\delta_{\xi,\perp}\sqrt{\hat{\delta}_{\perp,\xi,p,q}}\right).$$

□

B.4 PROOF OF THE MAIN LEMMA OF STAGE 1

Proof of Lemma B.1. First, we recap the estimations we have derived in previous subsections and introduce some notations. By Lemma B.10 and Lemma B.11, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_- \le -\Omega(1)\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\rho_- + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\rho}_{N/S} \le -\Omega(1)\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\hat{\rho}_{N/S} + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)d\right).$$

Define $\rho := \max\left\{\rho_-, \hat{\rho}_{N/S}\right\}$ to be the indicator of the progress we have made. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho \le -\Omega(1)\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\rho + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right).\right). \tag{11}$$

Let $\hat{\kappa}_0 := \max_{p,q}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 / \left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2$ be the condition number at time t. By Lemma B.12, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\kappa}_0 \le O(1)\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\rho\hat{\kappa}_0. \tag{12}$$

Now we consider the discretization errors, define

$$\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}} = \max_{p\in[r],q\in[d-r]}\left\{\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \rho_{\boldsymbol{B}/\boldsymbol{A},p} - 2, \rho_{\boldsymbol{A}/\boldsymbol{B},\xi,q} + \rho_{\boldsymbol{B}/\boldsymbol{A},\xi,q} - 2\right\}.$$

Note that at time t, the first condition of (10) holds with $\delta_{\boldsymbol{A}/\boldsymbol{B}}$ replaced by $O(\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}})$. Meanwhile, by Lemma B.13, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}} \le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\delta_{\boldsymbol{A}/\boldsymbol{B}}^2 + \tau_t^2 d\right)\right). \tag{13}$$

Let $\hat{\delta}_{\xi,\kappa_0}(t)$ be the smallest number such that the second condition of (10) holds at time t. By Lemma B.14, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\xi,\kappa_0} \le O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\tau_t^2 + \sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right). \tag{14}$$

Then, define

$$\hat{\delta}_\perp := \max\left\{\hat{\delta}_{\perp,p,q}, \hat{\delta}_{\perp,\xi_{\boldsymbol{A}},p,k}, \hat{\delta}_{\perp,\xi_{\boldsymbol{B}},p,k}, \hat{\delta}_{\perp,\xi,k,l}, : p \ne q \in [d], k, l \in [d-r]\right\}$$

Clear that the last two conditions hold at time t when $\delta_{\boldsymbol{AB},\perp}$ and $\delta_{\xi,\perp}$ are replaced by $\sqrt{\hat{\delta}_\perp(t)}$. By Lemma B.15 and Lemma B.16, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_\perp \le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho\hat{\delta}_\perp + O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tau_t^2 d^2\kappa_0 + d\sqrt{\delta_{\boldsymbol{A}/\boldsymbol{B}}}\right)\right). \tag{15}$$

Now, we are ready to show that the errors do not blow up in Stage 1. Note that for all these δ's, we can make them arbitrarily inverse-polynomially small by choosing a sufficiently large m.

First, we consider $\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}}$. Note that the dependence of the RHS of (13) on $\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}}$ is quadratic. Hence, by making the initial value of $\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}}$, the RHS can be made to be dominated the τ_t^2-related terms. Hence, $\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}} \le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tau_t^2 dT_1\right)$. As we will see later, $T_1 = \mathrm{poly}(d)$. Therefore, by choosing a sufficiently small τ_t^2, we can make $\hat{\delta}_{\boldsymbol{A}/\boldsymbol{B}}$ remain small throughout Stage 1.

Then, we consider $\hat{\delta}_{\xi,\kappa_0}$. As we have argued earlier, the RHS of (14) can be made arbitrarily small, so that $\hat{\delta}_{\xi,\kappa_0}$ remains small in Stage 1.

Now, we consider the condition number $\hat{\kappa}_0$ and $\hat{\delta}_\perp$. For $\hat{\delta}_\perp$, by our previous argument, the second term of the RHS of (15) can be merged into the first term, by choosing a sufficiently large m and a sufficiently small τ_t^2. The same is also true for (11). Hence, for these quantities, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho \le -\Omega(1)\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\rho, \quad \frac{\mathrm{d}}{\mathrm{d}t}\hat{\kappa}_0 \le O(1)\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\rho\hat{\kappa}_0, \quad \frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_\perp \le O(1)\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\rho\hat{\delta}_\perp.$$

Hence, by Lemma B.2, we have

$$\hat{\kappa}_0 \le \hat{\kappa}_0(0) \exp\left(O\left(\frac{\sigma_{\max}^2}{\sigma_{\min}^2}\right)\rho(0)\right), \quad \hat{\delta}_\perp \le \hat{\delta}_\perp(0) \exp\left(O\left(\frac{\sigma_{\max}^2}{\sigma_{\min}^2}\right)\rho(0)\right).$$

Note that $\rho_-(0) = O(1)$ and $\hat{\rho}_{N/S} \le \frac{(d-r)\sigma_\xi^2}{r\sigma_{\min}^2}$. Therefore,

$$\exp\left(O\left(\frac{\sigma_{\max}^2}{\sigma_{\min}^2}\right)\rho(0)\right) \le \exp\left(O(1)\frac{\sigma_{\max}^2}{\sigma_{\min}^2}\max\left\{1, \frac{(d-r)\sigma_\xi^2}{r\sigma_{\min}^2}\right\}\right) \le \exp\left(\frac{1}{2}\log d\right) = \sqrt{d}.$$

In other words, both $\hat{\kappa}_0$ and $\hat{\delta}_\perp$ can at most grow $\sqrt{d}$ times.

Finally, we derive an upper bound on T_1 to complete the proof. Similar to the proof for the condition number, one can show that $\|[\boldsymbol{K_A}]_p\|^2$ can at most grow $\sqrt{d}$ times in Stage 1. As a result, $1/(N_{\boldsymbol{A}}N_{\boldsymbol{B}}d)$ is lower bounded by some $1/\operatorname{poly}(d)$. Thus, by (11), $T_1 \le \operatorname{poly}(d)$. □

B.5 Negative results

Lemma B.17. *There exists a $\boldsymbol{\sigma} \in \mathbb{R}^r$ satisfying the assumptions of Theorem 4.3 such that, at the end of Stage 1, the condition number of $\boldsymbol{K_A}$ is $d^{\Omega(1)}$.*

Proof. We choose $d = r$ and $\sigma_1^2 = c\log d$, $\sigma_2^2 = \cdots = \sigma_d^2 = 1$. Clear that this satisfies the condition of Theorem 4.3. Note that it suffices to consider the infinite-width case, since, as we have proved earlier, the finite-width trajectory tracks the infinite-width one. By Lemma B.10, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle \approx \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1 + \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle\right)\left(1 - \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle\right).$$

By the proof of Lemma B.12, we have

$$\begin{aligned}\dot{\rho}_{p/q} \approx{}& \frac{4K}{1+K}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle - \frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_{p/q} \\ &- \frac{4K}{1+K}\frac{\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K_A}]_q}, \overline{[\boldsymbol{K_B}]_q}\right\rangle - \frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_{p/q}.\end{aligned}$$

Note that, in the infinite-width case, we have

$$\left\langle \overline{[\boldsymbol{K_A}]_1}, \overline{[\boldsymbol{K_B}]_1}\right\rangle \ge \left\langle \overline{[\boldsymbol{K_A}]_2}, \overline{[\boldsymbol{K_B}]_2}\right\rangle = \cdots = \left\langle \overline{[\boldsymbol{K_A}]_d}, \overline{[\boldsymbol{K_B}]_d}\right\rangle.$$

Therefore, $\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p}\right\rangle - \frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \ge 0$a and $\left\langle \overline{[\boldsymbol{K_A}]_q}, \overline{[\boldsymbol{K_B}]_q}\right\rangle - \frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \le 0$ for any $q \ge 2$. Hence,

$$\begin{aligned}\dot{\rho}_{1/2} \ge{}& \frac{4K}{1+K}\frac{\sigma_1^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left\langle \overline{[\boldsymbol{K_A}]_1}, \overline{[\boldsymbol{K_B}]_1}\right\rangle - \frac{\langle \boldsymbol{K_A}, \boldsymbol{K_B}\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\rho_{1/2} \\ \ge{}& \frac{4K}{1+K}\frac{\sigma_1^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left(1 - \frac{\kappa_1^2}{\|\boldsymbol{\kappa}\|^2}\right)\left\langle \overline{[\boldsymbol{K_A}]_1}, \overline{[\boldsymbol{K_B}]_1}\right\rangle - \frac{(d-1)\kappa_1^2}{\|\boldsymbol{\kappa}\|^2}\left\langle \overline{[\boldsymbol{K_A}]_2}, \overline{[\boldsymbol{K_B}]_2}\right\rangle\right)\rho_{1/2}.\end{aligned}$$

For notational simplicity, define $X_1 = 1 - \left\langle \overline{[\boldsymbol{K_A}]_1}, \overline{[\boldsymbol{K_B}]_1}\right\rangle$, $X_2 = 1 - \left\langle \overline{[\boldsymbol{K_A}]_2}, \overline{[\boldsymbol{K_B}]_2}\right\rangle$, $Y = \rho_{1/2}$, $A = \frac{4K}{1+K}\frac{1}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}$. Then we have

$$\dot{X}_1 \le -\sigma_1^2 A X_1, \quad \dot{X}_2 \ge -2AX_2.$$

First, by Gronwall's lemma, we have $X_1(T) \le \exp\left(-\sigma_1^2\int_0^T A\right)$ and

$$X_2(T) \ge \exp\left(-2\int_0^T A\right) \ge X_1^{2/\sigma_1^2}(T).$$

As a result, when X_1 reaches $1/d$, we have $X_2 = \Omega(1)$. Let T_1 be the time X_1 reaches $1/d$ and T_2 the time $X_2(T_2) = X_2(T_1)/2$. On $[T_1, T_2]$, we have

$$\dot{\rho}_{1/2} \geq \Omega(1)\sigma_1^2 A \rho_{1/2}.$$

By Gronwall's lemma, in order for X_2 to half, we need $\exp(-2\int_{T_1}^{T_2} A) = 1/2$. Hence,

$$\rho_{1/2}(T_2) \geq \rho_{1/2}(T_1) \exp\left(\Omega(1)\sigma_1^2 \int_{T_1}^{T_2} A\right) \geq 2^{\Omega(\sigma_1^2)} = d^{\Omega(1)}.$$

□

C STAGE 2

In this section, we show that, throughout Stage 2, the discretization error and the noise-signal ratio still remain small, and, at the end of Stage 2, the condition number is close to 1. Formally, we prove the following.

Lemma C.1 (Stage 2)**.** *Suppose that at the beginning of Stage 2, we have $\kappa_0 \leq \sqrt{d}$ and all errors mentioned in (16) are sufficiently small*[5]*. Let $c_{\texttt{Target}} > 1$ be a constant. Let T_1 be the earliest time that*

$$\frac{\|[\boldsymbol{K_A}]_p\|^2}{\|[\boldsymbol{K_A}]_q\|^2} \leq c_{\texttt{Target}}, \quad \forall p, q \in [r].$$

We have $T_1 \leq \mathrm{poly}(d)$. Moreover, throughout Stage 2, we have

$$\begin{aligned}
&\max\left\{1 - \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p} \right\rangle, \left|1 - \frac{\|[\boldsymbol{K_A}]_p\|^2}{\|[\boldsymbol{K_B}]_p\|^2}\right|\right\} \leq \delta_- && \forall p \in [r], \\
&\max\left\{\left|1 - \frac{\|[\boldsymbol{K_{A,\xi}}]_p\|^2}{\|[\boldsymbol{K_{B,\xi}}]_p\|^2}\right|, \left|1 - \frac{\|[\boldsymbol{K_{A,\xi}}]_p\|^2}{\|[\boldsymbol{K_{A,\xi}}]_q\|^2}\right|\right\} \leq \delta_- && \forall p, q \in [d-r], \\
&\max\left\{\frac{\|[\boldsymbol{K_{C,\xi}}]_q\|}{\|[\boldsymbol{K_D}]_p\|} : \boldsymbol{C}, \boldsymbol{D} \in \{\boldsymbol{A}, \boldsymbol{B}\}\right\} \leq \delta_{N/S}, && \forall p \in [r], q \in [d-r], \\
&\max\left\{\left|\left\langle \overline{[\boldsymbol{K_C}]_p}, \overline{[\boldsymbol{K_D}]_q} \right\rangle\right| : \boldsymbol{C}, \boldsymbol{D} \in \{\boldsymbol{A}, \boldsymbol{B}\}\right\} \leq \delta_{\boldsymbol{AB},\perp}, && \forall p \neq q \in [r], \\
&\max\left\{\left|\left\langle \overline{[\boldsymbol{K_C}]_p}, \overline{[\boldsymbol{K_{D,\xi}}]_q} \right\rangle\right|, \left|\left\langle \overline{[\boldsymbol{K_{A,\xi}}]_s}, \overline{[\boldsymbol{K_{B,\xi}}]_q} \right\rangle\right| : \boldsymbol{C}, \boldsymbol{D} \in \{\boldsymbol{A}, \boldsymbol{B}\}\right\} \leq \delta_{\xi,\perp}, && \forall p \in [r], q, s \in [d-r],
\end{aligned} \tag{16}$$

where the δ's are some small $1/\mathrm{poly}(d)$ values.

The rest of this section is organized as follows. We derive estimations for the $\boldsymbol{Q}$-matrices in Section C.1. In Section C.2, we maintain the last two conditions of (16). In Section C.3, we handle the first two conditions of (16). In Section C.4, we deal with the noise-signal ratio. We estimate the convergence rate in Section C.5. Finally, we prove Lemma C.1 in Section C.6.

C.1 ESTIMATIONS FOR Q

As in Stage 1, we estimate the $\boldsymbol{Q}$-matrices in this subsection. The analysis here will be more complicated than the one in Section B.1 since now τ_t^2 is no longer close to 0, and we cannot simply approximate $S_{\boldsymbol{A}}$ and $S_{\boldsymbol{B}}$ with $(1+K)^{-1}$. First, we need the following lemma which gives closed-form formula for some expectations we will encounter later.

[5] By Lemma B.1, this condition indeed holds.

Lemma C.2. *Define* $\langle \boldsymbol{z}^+, \boldsymbol{z}^- \rangle_{\hat{\boldsymbol{\kappa}}^2} := \sum_{k=1}^r \hat{\kappa}_k^2 z_k^+ z_k^-$ *and* $T_p := \tanh\left(\frac{\hat{\kappa}_p^2/d}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right)$, $p \in [r]$. *For any* $p \neq q \in [r]$, *we have*

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\exp\left(\frac{\langle \boldsymbol{z}^+, \boldsymbol{z}^- \rangle_{\hat{\boldsymbol{\kappa}}^2}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right)\right\} = \prod_{k=1}^r \cosh\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right) =: Z_c,$$

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\exp\left(\frac{\langle \boldsymbol{z}^+, \boldsymbol{z}^- \rangle_{\hat{\boldsymbol{\kappa}}^2}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right) z_p^-\right\} = Z_c T_p z_p^+,$$

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\exp\left(\frac{\langle \boldsymbol{z}^+, \boldsymbol{z}^- \rangle_{\hat{\boldsymbol{\kappa}}^2}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right) z_p^- z_q^-\right\} = Z_c T_p T_q z_p^+ z_q^+.$$

Then, we derive estimations for $S_{\boldsymbol{A}}$ and $S_{\boldsymbol{B}}$. There are two types of errors we need to consider. The first one comes from the noises and the second one from the non-diagonalness of $\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B}}$. Similar to Lemma B.4, we deal with them separately. The next lemma handles the first type of error.

Lemma C.3 (Estimations for S)**.** *Define*

$$E_{+,-} := \exp\left(\frac{\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^+, \boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{z}^- \rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right), \quad \tilde{E}_{+,-} := \exp\left(\frac{\langle \boldsymbol{z}^+, \boldsymbol{z}^- \rangle_{\hat{\boldsymbol{\kappa}}^2}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right),$$

$$\delta_{+,\xi_-} := \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{z}^+, \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{B}}^- \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}, \quad \delta_{\xi+,T+} := \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{K}_{\boldsymbol{B}} \operatorname{diag}([T_k]_{k\in[r]}) \boldsymbol{z}^+ \right\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}},$$

$$Z_{\boldsymbol{A},c} := \mathop{\mathbb{E}}_{\boldsymbol{z}^-} E_{+,-}, \quad Z_{\boldsymbol{B},c} := \mathop{\mathbb{E}}_{\boldsymbol{z}^-} E_{-,+},$$

$$\tilde{S}_{\boldsymbol{A}} := \frac{E_{+,+}}{E_{+,+} + K Z_{\boldsymbol{A},c}}, \quad \tilde{S}_{\boldsymbol{B}} := \frac{E_{+,+}}{E_{+,+} + K Z_{\boldsymbol{B},c}}, \quad \tilde{S} := \frac{\tilde{E}_{+,+}}{\tilde{E}_{+,+} + K Z_c}.$$

In Stage 2, we have

$$S_{\boldsymbol{A}} = \tilde{S}_{\boldsymbol{A}} + \tilde{S}(1-\tilde{S})\left(\delta_{+,\xi+} + \delta_{\xi+,+} - \delta_{\xi+,T+}\right) \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right),$$

$$S_{\boldsymbol{B}} = \tilde{S}_{\boldsymbol{B}} + \tilde{S}(1-\tilde{S})\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+} - \delta_{T+,\boldsymbol{\xi}+}\right) \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right),$$

and

$$\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} = \frac{\tilde{S}_{\boldsymbol{A}} E_{+,-}}{E_{+,+}} + \frac{\tilde{S}\tilde{E}_{+,-}}{\tilde{E}_{+,+}}\left(\delta_{+,\boldsymbol{\xi}-} + \delta_{\boldsymbol{\xi}+,-} - \tilde{S}\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+}\right) - (1-\tilde{S})\delta_{\boldsymbol{\xi}+,T+}\right)$$
$$\pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right),$$

$$\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} = \frac{\tilde{S}_{\boldsymbol{B}} E_{-,+}}{E_{+,+}} + \frac{\tilde{S}\tilde{E}_{-,+}}{\tilde{E}_{+,+}}\left(1 + \delta_{\boldsymbol{\xi}-,+} + \delta_{-,\boldsymbol{\xi}+} - \tilde{S}\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+}\right) - (1-\tilde{S})\delta_{T+,\boldsymbol{\xi}+}\right)$$
$$\pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Then, we consider the error comes from the non-diagonalness of $\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B}}$.

Lemma C.4 (Further estimations for S)**.** *Define*

$$\tilde{\delta}_{+,-} = \sum_{i\neq j} \frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_i, [\boldsymbol{K}_{\boldsymbol{B}}]_j \rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} z_i^+ z_j^-, \quad \tilde{\delta}_{+,T+} = \sum_{i\neq j} \frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_i, [\boldsymbol{K}_{\boldsymbol{B}}]_j \rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} z_i^+ T_j z_j^+,$$

$$\tilde{E}_0 := \tilde{E}_{+,+} = \tilde{E}_{-,-} = \exp\left(\frac{\|\hat{\boldsymbol{\kappa}}\|^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right).$$

In Stage 2, we have

$$\tilde{S}_{\boldsymbol{A}}(\boldsymbol{z}^+) = \tilde{S}\left(1 + (1-\tilde{S})(\tilde{\delta}_{+,+} - \tilde{\delta}_{+,T+})\right) \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right),$$

$$\tilde{S}_{\boldsymbol{B}}(\boldsymbol{z}^+) = \tilde{S}\left(1 + (1-\tilde{S})(\tilde{\delta}_{+,+} - \tilde{\delta}_{T+,+})\right) \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right),$$

$$\frac{\tilde{S}_{\boldsymbol{A}} E_{+,-}}{E_{+,+}} = \frac{\tilde{S}\tilde{E}_{+,-}}{\tilde{E}_0}\left(1 - \tilde{S}\tilde{\delta}_{+,+} - (1-\tilde{S})\tilde{\delta}_{+,T+} + \tilde{\delta}_{+,-}\right) \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right),$$

$$\frac{\tilde{S}_{\boldsymbol{B}} E_{-,+}}{E_{+,+}} = \frac{\tilde{S}\tilde{E}_{-,+}}{\tilde{E}_0}\left(1 - \tilde{S}\tilde{\delta}_{+,+} - (1-\tilde{S})\tilde{\delta}_{T+,+} + \tilde{\delta}_{-,+}\right) \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).$$

With the above two lemmas in hand, we can now derive estimations for the $\boldsymbol{Q}$-matrices.

Lemma C.5 (Estimations for $\boldsymbol{Q}_1$)**.** *Define $\boldsymbol{K_{AB}} = \boldsymbol{K_A}^\top \boldsymbol{K_B}$ and $\boldsymbol{K_{BA}} = \boldsymbol{K_B}^\top \boldsymbol{K_A}$. In Stage 2, for any $p \neq q \in [r]$, we have*

$$
\begin{aligned}
[\boldsymbol{Q}_1]_{p,p} &= 2(1-\tilde{S})(1-T_p) \pm O\left(d^2 \delta_{\boldsymbol{AB},\perp}^2\right), \\
[\boldsymbol{Q}_1]_{p,q} &= -\tilde{S}(1-\tilde{S}) \frac{[\boldsymbol{K_{AB}}]_{p,q} + [\boldsymbol{K_{BA}}]_{q,p}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} (2 - T_p - T_q) \\
&\quad - (1-\tilde{S}) \frac{[\boldsymbol{K_{AB}}]_{p,q}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \tilde{S} \left(2T_pT_q - T_p - T_q\right) \\
&\quad - (1-\tilde{S}) \frac{[\boldsymbol{K_{AB}}]_{q,p}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \left(2 - \tilde{S}(T_p + T_q) - (1-\tilde{S})(T_p^2 + T_q^2)\right) \pm O\left(d^2 \delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}
$$

In particular, we have

$$
|[\boldsymbol{Q}_1]_{p,q}| \leq O\left(\frac{\kappa_0 \delta_{\boldsymbol{AB},\perp}}{d}\right) \leq O\left(\frac{\delta_{\boldsymbol{AB},\perp}}{\sqrt{d}}\right).
$$

Lemma C.6 (Estimations for $\boldsymbol{Q}_{1,\xi}$)**.** *In Stage 2, for any $p \in [d-r]$ and $q \in [r]$, we have*

$$
\begin{aligned}
[\boldsymbol{Q}_{1,\boldsymbol{\xi_A}}]_{p,q} &= -(1-\tilde{S})\left(1 + \tilde{S} + (1-\tilde{S})T_q^2\right) \frac{\langle [\boldsymbol{K_B}]_q, [\boldsymbol{K_{A,\xi}}]_p \rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(d^3 \left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right) \delta_{\xi,\perp} \delta_{N/S}\right), \\
[\boldsymbol{Q}_{1,\boldsymbol{\xi_B}}]_{p,q} &= -(1-\tilde{S})\left(1 + \tilde{S} + (1-\tilde{S})T_q^2\right) \frac{\langle [\boldsymbol{K_A}]_q, [\boldsymbol{K_{B,\xi}}]_p \rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(d^3 \left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right) \delta_{\xi,\perp} \delta_{N/S}\right).
\end{aligned}
$$

In particular, we have

$$
\max\left\{[\boldsymbol{Q}_{1,\boldsymbol{\xi_A}}]_{p,q}, [\boldsymbol{Q}_{1,\boldsymbol{\xi_B}}]_{p,q}\right\} \leq O\left(\delta_{N/S} \delta_{\xi,\perp}\right).
$$

Lemma C.7 (Estimations for $\boldsymbol{Q}_2$)**.** *In Stage 2, for any $p, q \in [d-r]$, we have*

$$
[\boldsymbol{Q}_2]_{p,q} = \pm O\left(d^3 \left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right) \delta_{\xi,\perp} \delta_{N/S}\right).
$$

Lemma C.8 (Estimations for Q_0)**.** *In Stage 2, we have*

$$
Q_0 = -\sum_{k=1}^{r} \frac{\kappa_k^2}{\|\boldsymbol{\kappa}\|^2} [\boldsymbol{Q}_1]_{k,k} \pm O\left(d^2 \delta_{\boldsymbol{AB},\perp}^2 + \delta_- + \kappa_0 d \delta_{N/S} \delta_{\xi,\perp}\right).
$$

Finally, we use these estimations to simplify the formulas for the norms. We do not consider the tangent movement here since the situation is trickier there, and we will handle them in later subsections.

Corollary C.9 (Dynamics of the norms)**.** *In Stage 2, we have*

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \|[\boldsymbol{K_A}]_p\|^2 &= \frac{2\sigma_p^2 \|[\boldsymbol{K_A}]_p\| \|[\boldsymbol{K_B}]_p\|}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_p} \right\rangle [\boldsymbol{Q}_1]_{p,p} + 2\frac{\|[\boldsymbol{K_A}]_p\|^2}{N_{\boldsymbol{A}}^2 d} Q_0 \sigma_p^2 \\
&\quad \pm O\left(\frac{\sigma_p^2 \kappa_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \kappa_0 d \delta_{\boldsymbol{AB},\perp}^2\right) \\
\frac{\mathrm{d}}{\mathrm{d}t} \|[\boldsymbol{K_{A,\xi}}]_q\|^2 &= \frac{2\|[\boldsymbol{K_{A,\xi}}]_q\|^2}{N_{\boldsymbol{A}}^2 d} Q_0 \sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2 \|[\boldsymbol{K_{A,\xi}}]_q\|^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d \delta_{\xi,\perp}^2\right).
\end{aligned}
$$

The formulas for $\|[\boldsymbol{K_B}]_p\|$ and $\|[\boldsymbol{K_{B,\xi}}]_q\|$ can be obtained by interchanging the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$.

OMITTED PROOFS OF THIS SUBSECTION

Proof of Lemma C.2. First, we compute

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\exp\left(\frac{\langle \boldsymbol{z}^+, \boldsymbol{z}^- \rangle_{\hat{\boldsymbol{\kappa}}^2}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right)\right\} &= \prod_{k=1}^{r} \mathop{\mathbb{E}}_{z_k^-}\left\{\exp\left(\frac{\hat{\kappa}_k^2 z_k^+ z_k^-}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right)\right\} = \prod_{k=1}^{r} \frac{1}{2}\left(\exp\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right) + \exp\left(-\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right)\right) \\
&= \prod_{k=1}^{r} \cosh\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\right).
\end{aligned}
$$

Similarly, we also have

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\exp\left(\frac{\langle \boldsymbol{z}^+, \boldsymbol{z}^-\rangle_{\hat{\boldsymbol{\kappa}}^2}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) z_p^-\right\} = \mathop{\mathbb{E}}_{z_p^-}\left\{\exp\left(\frac{\hat{\kappa}_k^2 z_p^+ z_p^-}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) z_p^-\right\} \prod_{k\neq p} \mathop{\mathbb{E}}_{z_k^-}\left\{\exp\left(\frac{\hat{\kappa}_k^2 z_k^+ z_k^-}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right)\right\}.$$

Each factor in $\prod_{k\neq p}$ is still $\cosh\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right)$. For the first term, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{z_p^-}\left\{\exp\left(\frac{\hat{\kappa}_k^2 z_p^+ z_p^-}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) z_p^-\right\} &= \frac{1}{2\sqrt{d}}\mathop{\mathbb{E}}_{z_p^-}\left\{\exp\left(\frac{\hat{\kappa}_k^2 z_p^+/\sqrt{d}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right)\right\} - \frac{1}{2\sqrt{d}}\mathop{\mathbb{E}}_{z_p^-}\left\{\exp\left(\frac{-\hat{\kappa}_k^2 z_p^+/\sqrt{d}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) z_p^-\right\} \\ &= \frac{1}{\sqrt{d}}\sinh\left(\frac{\hat{\kappa}_k^2 z_p^+/\sqrt{d}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) = \frac{\operatorname{sgn} z_o^+}{\sqrt{d}}\sinh\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) = \sinh\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) z_p^+.\end{aligned}$$

Therefore,

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\exp\left(\frac{\langle \boldsymbol{z}^+, \boldsymbol{z}^-\rangle_{\hat{\boldsymbol{\kappa}}^2}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) z_p^-\right\} = z_p^+\sinh\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right)\prod_{k\neq p}\cosh\left(\frac{\hat{\kappa}_k^2/d}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) = Z_c T_p z_p^+.$$

The above calculation, *mutatis mutandis*, also yields the last identity. □

Proof of Lemma C.3. First, we write

$$\begin{aligned}\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^-\rangle &= \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{z}^+ + \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{z}^- + \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{B}}^-\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} \\ &= \frac{\langle \boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{z}^+, \boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{z}^-\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} + \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{z}^+, \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{B}}^-\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} + \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{z}^-\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right).\end{aligned}$$

Also note that the middle terms are $O(d^2\delta_{\xi,\perp}\delta_{N/S})$. Then, we compute

$$\exp\left(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-\right) = \exp\left(1 + \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{z}^+, \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{B}}^-\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} + \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{z}^-\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right) \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right).$$

Similar results also hold for other combinations of $\pm$. With the notations defined in this lemma, we can write these results as

$$\begin{aligned}\exp\left(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-\right) &= E_{+,-}\left(1 + \delta_{+,\xi-} + \delta_{\xi+,-}\right) \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right), \\ \exp\left(\boldsymbol{f}_{\boldsymbol{A}}^-\cdot\boldsymbol{f}_{\boldsymbol{B}}^+\right) &= E_{-,+}\left(1 + \delta_{-,\xi+} + \delta_{\xi-,+}\right) \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right), \\ \exp\left(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+\right) &= E_{+,+}\left(1 + \delta_{+,\xi+} + \delta_{\xi+,+}\right) \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right).\end{aligned}$$

To compute $S_{\boldsymbol{A}}$ and $S_{\boldsymbol{B}}$, we then need to take expectations over the negative examples. Note that by taking expectation over $\boldsymbol{\xi}_{\boldsymbol{B}}^-$, the term $E_{+,-}\delta_{+,\xi-}$ becomes 0. Therefore, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{B}}^-}\exp\left(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-\right) = \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{B}}^-}\left\{E_{+,-}\left(1 + \delta_{\xi+,-}\right)\right\} \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right).$$

Unfortunately, the same argument does not apply to $\delta_{\xi+,-}$ since both $E_{+,-}$ and $\delta_{\xi+,-}$ depend on $\boldsymbol{z}^-$. However, it is still possible to further simplify the expression. First, we write

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{E_{+,-}\delta_{\xi+,-}\right\} = \mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\tilde{E}_{+,-}\left(1 \pm O(d\delta_{\boldsymbol{AB},\perp})\right)\delta_{\xi+,-}\right\} = \mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\tilde{E}_{+,-}\delta_{\xi+,-}\right\} \pm O(d^3\delta_{\boldsymbol{AB},\perp}\delta_{\xi,\perp}\delta_{N/S}).$$

Recall Lemma C.2. Then, we compute

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\tilde{E}_{+,-}\delta_{\xi+,-}\right\} = \mathop{\mathbb{E}}_{\boldsymbol{z}^-}\left\{\tilde{E}_{+,-}\frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{z}^-\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\right\} &= \frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{K}_{\boldsymbol{B}}\,\mathbb{E}_{\boldsymbol{z}^-}\left\{\tilde{E}_{+,-}\boldsymbol{z}^-\right\}\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} \\ &= Z_c\frac{\left\langle \boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{\xi}_{\boldsymbol{A}}^+, \boldsymbol{K}_{\boldsymbol{B}}\operatorname{diag}([T_k]_{k\in[r]})\boldsymbol{z}^+\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} = Z_c\delta_{\xi+,T+}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{B}}^-} \exp\left(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-\right) &= \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{B}}^-} \{E_{+,-}\} + \mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{B}}^-} \{E_{+,-}\delta_{\xi+,-}\} \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right) \\ &= Z_{\boldsymbol{A},c} + Z_c\delta_{\xi+,T+} \pm O(d^3\delta_{\boldsymbol{AB},\perp}\delta_{\xi,\perp}\delta_{N/S}) \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right) \\ &= Z_{\boldsymbol{A},c} + Z_c\delta_{\xi+,T+} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).\end{aligned}$$

Similarly, we also have

$$\mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^-} \exp\left(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-\right) = Z_{\boldsymbol{B},c} + Z_c\delta_{T+,\xi+} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Recall that $\exp\left(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+\right) = E_{+,+}\left(1 + \delta_{+,\xi+} + \delta_{\xi+,+}\right) \pm O\left(d^2\delta_{\xi,\perp}\delta_{N/S}^2\right)$. Hence, we have

$$\begin{aligned}S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) &= \frac{E_{+,+}(1 + \delta_{+,\xi+} + \delta_{\xi+,+})}{E_{+,+}(1 + \delta_{+,\xi+} + \delta_{\xi+,+}) + KZ_{\boldsymbol{A},c} + KZ_c\delta_{\xi+,T+}} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right) \\ &= \tilde{S}_{\boldsymbol{A}}\left(1 - \tilde{S}_{\boldsymbol{A}}(\delta_{+,\xi+} + \delta_{\xi+,+}) - (1 - \tilde{S}_{\boldsymbol{A}})\delta_{\xi+,T+}\right) \\ &\qquad + \tilde{S}_{\boldsymbol{A}}(\delta_{+,\xi+} + \delta_{\xi+,+}) \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right) \\ &= \tilde{S}_{\boldsymbol{A}} + \tilde{S}(1 - \tilde{S})\left(\delta_{+,\xi+} + \delta_{\xi+,+} - \delta_{\xi+,T+}\right) \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).\end{aligned}$$

Similarly, we also have

$$S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) = \tilde{S}_{\boldsymbol{B}} + \tilde{S}(1 - \tilde{S})\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+} - \delta_{T+,\boldsymbol{\xi}+}\right) \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Then, we compute

$$\begin{aligned}\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} &= \left(\tilde{S}_{\boldsymbol{A}} + \tilde{S}(1 - \tilde{S})\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+} - \delta_{\boldsymbol{\xi}+,T+}\right)\right)\frac{E_{+,-}\left(1 + \delta_{+,\boldsymbol{\xi}-} + \delta_{\boldsymbol{\xi}+,-}\right)}{E_{+,+}\left(1 + \delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+}\right)} \\ &\qquad \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right) \\ &= \frac{\tilde{S}_{\boldsymbol{A}}E_{+,-}}{E_{+,+}}\left(1 + \delta_{+,\boldsymbol{\xi}-} + \delta_{\boldsymbol{\xi}+,-} - \tilde{S}\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+}\right) - (1 - \tilde{S})\delta_{\boldsymbol{\xi}+,T+}\right) \\ &\qquad \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right) \\ &= \frac{\tilde{S}_{\boldsymbol{A}}E_{+,-}}{E_{+,+}} + \frac{\tilde{S}\tilde{E}_{+,-}}{\tilde{E}}\left(\delta_{+,\boldsymbol{\xi}-} + \delta_{\boldsymbol{\xi}+,-} - \tilde{S}\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+}\right) - (1 - \tilde{S})\delta_{\boldsymbol{\xi}+,T+}\right) \\ &\qquad \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right),\end{aligned}$$

and

$$\begin{aligned}\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} &= \frac{\tilde{S}_{\boldsymbol{B}}E_{-,+}}{E_{+,+}} + \frac{\tilde{S}\tilde{E}_{-,+}}{\tilde{E}}\left(1 + \delta_{\boldsymbol{\xi}-,+} + \delta_{-,\boldsymbol{\xi}+} - \tilde{S}\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+}\right) - (1 - \tilde{S})\delta_{T+,\boldsymbol{\xi}+}\right) \\ &\qquad \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).\end{aligned}$$

□

Proof of Lemma C.4. We write

$$\frac{\langle \boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{z}^+, \boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{z}^-\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} = \sum_{k=1}^{r}\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_k, [\boldsymbol{K}_{\boldsymbol{B}}]_k\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}z_k^+z_k^- + \sum_{i\neq j}\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_i, [\boldsymbol{K}_{\boldsymbol{B}}]_j\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}z_i^+z_j^- =: I_{+,-} + \tilde{\delta}_{+,-}.$$

Note that, as a special case, we have $I_{+,+} = I_{-,-} = I_0$. In other words, $I_{+,+}$ and $I_{-,-}$ do not depend on the actual value of $\boldsymbol{z}^{\pm}$. Also note that $I_{+,-}$ is bounded by $O(d\delta_{\boldsymbol{AB},\perp})$. Then, we compute

$$E_{+,-} = \exp(I_{+,-})\left(1 + \tilde{\delta}_{+,-} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right)\right) \quad \text{and} \quad E_{+,+} = \exp(I_0)\left(1 + \tilde{\delta}_{+,+} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right)\right).$$

Take expectation over $\boldsymbol{z}^-$ and we obtain

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-} E_{+,-} = \mathop{\mathbb{E}}_{\boldsymbol{z}^-} \exp(I_{+,-}) + \mathop{\mathbb{E}}_{\boldsymbol{z}^-} \left\{ \exp(I_{+,-}) \tilde{\delta}_{+,-} \right\} \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right).$$

By Lemma C.2, the first term is Z_c. For the second term, we compute

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{z}^-} \left\{ \exp(I_{+,-}) \tilde{\delta}_{+,-} \right\} &= \sum_{i \neq j} \frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{i,j}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} \mathop{\mathbb{E}}_{\boldsymbol{z}^-} \left\{ \exp(I_{+,-}) z_j^- \right\} z_i^+ \\
&= Z_c \sum_{i \neq j} \frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{i,j}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} T_j z_j^+ z_i^+ \\
&= Z_c \tilde{\delta}_{+,T+},
\end{aligned}$$

where the second line again comes from Lemma C.2. Hence, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^-} E_{+,-} = Z_c + Z_c \tilde{\delta}_{+,T+} \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right).$$

Then, for $\tilde{S}_{\boldsymbol{A}}$, we have

$$\begin{aligned}
\tilde{S}_{\boldsymbol{A}}(\boldsymbol{z}^+) = \frac{E_{+,+}}{E_{+,+} + K \mathbb{E}_{\boldsymbol{z}^-} E_{+,-}} &= \frac{\exp(I_0)\left(1 + \tilde{\delta}_{+,+}\right)}{\exp(I_0)\left(1 + \tilde{\delta}_{+,+}\right) + K Z_c + K Z_c \tilde{\delta}_{+,T+}} \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right) \\
&= \tilde{S}\left(1 - \tilde{S}\tilde{\delta}_{+,+} - (1 - \tilde{S})\tilde{\delta}_{+,T+}\right) + \tilde{S}\tilde{\delta}_{+,+} \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right) \\
&= \tilde{S}\left(1 + (1 - \tilde{S})(\tilde{\delta}_{+,+} - \tilde{\delta}_{+,T+})\right) \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right).
\end{aligned}$$

Similarly, we also have

$$\tilde{S}_{\boldsymbol{B}}(\boldsymbol{z}^+) = \tilde{S}\left(1 + (1 - \tilde{S})(\tilde{\delta}_{+,+} - \tilde{\delta}_{T+,+})\right) \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right).$$

Then, we compute

$$\begin{aligned}
\frac{\tilde{S}_{\boldsymbol{A}} E_{+,-}}{E_{+,+}} &= \tilde{S}\left(1 + (1 - \tilde{S})(\tilde{\delta}_{+,+} - \tilde{\delta}_{+,T+})\right) \frac{\exp(I_{+,-})\left(1 + \tilde{\delta}_{+,-}\right)}{\exp(I_{+,-})\left(1 + \tilde{\delta}_{+,+}\right)} \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right) \\
&= \frac{\tilde{S}\exp(I_{+,-})}{\exp(I_0)} \left(1 - \tilde{S}\tilde{\delta}_{+,+} - (1 - \tilde{S})\tilde{\delta}_{+,T+} + \tilde{\delta}_{+,-}\right) \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right).
\end{aligned}$$

Similarly, we also have

$$\frac{\tilde{S}_{\boldsymbol{B}} E_{-,+}}{E_{+,+}} = \frac{\tilde{S}\exp(I_{-,+})}{\exp(I_0)} \left(1 - \tilde{S}\tilde{\delta}_{+,+} - (1 - \tilde{S})\tilde{\delta}_{T+,+} + \tilde{\delta}_{-,+}\right) \pm O\left(d^2 \delta^2_{\boldsymbol{AB},\perp}\right).$$

□

Proof of Lemma C.5. Recall that

$$\begin{aligned}
\boldsymbol{Q}_1 := \mathbb{E}\left\{ \left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \boldsymbol{z}^+ (\boldsymbol{z}^+)^\top d \right\} \\
- K \mathbb{E}\left\{ \frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{z}^- (\boldsymbol{z}^+)^\top d \right\} - K \mathbb{E}\left\{ \frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) \exp(\boldsymbol{f}_{\boldsymbol{A}}^- \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} \boldsymbol{z}^+ (\boldsymbol{z}^-)^\top d \right\}.
\end{aligned}$$

Note that there is no $\boldsymbol{\xi}$ here other than the ones in the coefficient. As a result, all terms contain $\delta_{+,\boldsymbol{\xi}+}$ and alike are 0. Hence, we have

$$\begin{aligned}
\boldsymbol{Q}_1 &:= \mathbb{E}\left\{ \left(2 - \tilde{S}_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - \tilde{S}_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) \boldsymbol{z}^+ (\boldsymbol{z}^+)^\top d \right\} \\
&\quad - K \mathbb{E}\left\{ \frac{\tilde{S}_{\boldsymbol{A}} E_{+,-}}{E_{+,+}} \boldsymbol{z}^- (\boldsymbol{z}^+)^\top d \right\} - K \mathbb{E}\left\{ \frac{\tilde{S}_{\boldsymbol{B}} E_{-,+}}{E_{+,+}} \boldsymbol{z}^+ (\boldsymbol{z}^-)^\top d \right\} \\
&=: \mathrm{T}_1(\boldsymbol{Q}_1) + \mathrm{T}_2(\boldsymbol{Q}_1) + \mathrm{T}_3(\boldsymbol{Q}_1).
\end{aligned}$$

Now we estimate each of these three terms separately. We also deal with the diagonal and off-diagonal terms separately. By Lemma C.4, we have

$$\begin{aligned}
\mathrm{T}_1([\boldsymbol{Q}_1]_{p,p}) &= \mathbb{E}\left\{2 - \tilde{S}_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - \tilde{S}_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right\} \\
&= \mathbb{E}\left\{2 - \tilde{S}\left(1 + (1-\tilde{S})(\tilde{\delta}_{+,+} - \tilde{\delta}_{+,T+})\right) - \tilde{S}\left(1 + (1-\tilde{S})(\tilde{\delta}_{+,+} - \tilde{\delta}_{T+,+})\right)\right\} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= 2(1-\tilde{S}) \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}$$

Note that we use the fact that all these δ's have mean 0. Also by Lemma C.4, we have

$$\begin{aligned}
\mathrm{T}_2([\boldsymbol{Q}_1]_{p,p}) &= -K\,\mathbb{E}\left\{\left(\frac{\tilde{S}\tilde{E}_{+,-}}{\tilde{E}_0}\left(1 - \tilde{S}\tilde{\delta}_{+,+} - (1-\tilde{S})\tilde{\delta}_{+,T+} + \tilde{\delta}_{+,-}\right)\right) z_p^- z_p^+ d\right\} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -\frac{\tilde{S}K}{\tilde{E}_0}\,\mathbb{E}\left\{\tilde{E}_{+,-} z_p^- z_p^+ d\right\} \\
&\quad - \frac{\tilde{S}K}{\tilde{E}_0}\,\mathbb{E}\left\{\left(\tilde{E}_{+,-}\left(-\tilde{S}\tilde{\delta}_{+,+} - (1-\tilde{S})\tilde{\delta}_{+,T+} + \tilde{\delta}_{+,-}\right)\right) z_p^- z_p^+ d\right\} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}$$

Note that $\tilde{\delta}_{+,+}$, $\tilde{\delta}_{+,-}$ and $\tilde{\delta}_{+,T+}$ only contain cross terms of form $z_i^+ z_j^\pm$ with $i \neq j$. Hence, the second term is 0. Meanwhile, by Lemma C.2, the first term is

$$-\frac{\tilde{S}K}{\tilde{E}_0}\,\mathbb{E}\left\{\tilde{E}_{+,-} z_p^- z_p^+ d\right\} = -\frac{\tilde{S}KZ_cT_p}{\tilde{E}_0} = -(1-\tilde{S})T_p.$$

As a result,

$$\mathrm{T}_2([\boldsymbol{Q}_1]_{p,p}) = -(1-\tilde{S})T_p \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).$$

Similarly, one can show that $\mathrm{T}_3([\boldsymbol{Q}_1]_{p,p}) = -(1-\tilde{S})T_p \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right)$ also holds. Thus,

$$[\boldsymbol{Q}_1]_{p,p} = 2(1-\tilde{S})(1-T_p) \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).$$

Now, we consider the off-diagonal terms. For notational simplicity, we define $\boldsymbol{K}_{\boldsymbol{AB}} = \boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B}}$. For any $p \neq q$, we compute

$$\begin{aligned}
\mathrm{T}_1([\boldsymbol{Q}_1]_{p,q}) &= -\mathbb{E}\left\{\left(\tilde{S}_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) + \tilde{S}_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right) z_p^+ z_q^+ d\right\} \\
&= -\tilde{S}(1-\tilde{S})\,\mathbb{E}\left\{\left(2\tilde{\delta}_{+,+} - \tilde{\delta}_{+,T+} - \tilde{\delta}_{T+,T}\right) z_p^+ z_q^+ d\right\} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -\tilde{S}(1-\tilde{S})\sum_{i\neq j}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{i,j}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}(2 - T_i - T_j)\,\mathbb{E}\left\{z_i^+ z_j^+ z_p^+ z_q^+ d\right\} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}$$

Clear that the summand is nonzero only if $(i,j) = (p,q)$ or $(i,j) = (q,p)$. Hence,

$$\mathrm{T}_1([\boldsymbol{Q}_1]_{p,q}) = -\tilde{S}(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q} + [\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}(2 - T_p - T_q) \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).$$

Then, for T_2, we compute

$$\begin{aligned}
\mathrm{T}_2([\boldsymbol{Q}_1]_{p,q}) &= -K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+ \cdot \boldsymbol{f}_{\boldsymbol{B}}^+)} z_p^- z_q^+ d\right\} \\
&= -K\,\mathbb{E}\left\{\left(\frac{\tilde{S}\tilde{E}_{+,-}}{\tilde{E}_0}\left(1 - \tilde{S}\tilde{\delta}_{+,+} - (1-\tilde{S})\tilde{\delta}_{+,T+} + \tilde{\delta}_{+,-}\right)\right) z_p^- z_q^+ d\right\} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -K\,\mathbb{E}\left\{\left(\frac{\tilde{S}\tilde{E}_{+,-}}{\tilde{E}_0}\left(-\tilde{S}\tilde{\delta}_{+,+} - (1-\tilde{S})\tilde{\delta}_{+,T+} + \tilde{\delta}_{+,-}\right)\right) z_p^- z_q^+ d\right\} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -K\sum_{i\neq j}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{i,j}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\,\mathbb{E}\left\{\left(\frac{\tilde{S}\tilde{E}_{+,-}}{\tilde{E}_0}\left(-\tilde{S}z_i^+ z_j^+ - (1-\tilde{S})z_i^+ z_j^+ T_j + z_i^+ z_j^-\right)\right) z_p^- z_q^+ d\right\} \\
&\quad \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}$$

Again, the summand is nonzero only if $(i,j)=(p,q)$ or $(i,j)=(q,p)$. By Lemma C.2, we have

$$
\begin{aligned}
&\mathrm{T}_2([\boldsymbol{Q}_1]_{p,q}) \\
&= -\frac{\tilde{S}K}{\tilde{E}_0}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\left(\left(-\tilde{S}-(1-\tilde{S})T_q\right)\mathbb{E}\left\{\exp(I_{+,-})z_p^+z_p^-\right\}+\mathbb{E}\left\{\exp(I_{+,-})z_p^+z_q^+z_p^-z_q^-d\right\}\right) \\
&\quad -\frac{\tilde{S}K}{\tilde{E}_0}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\left(\left(-\tilde{S}-(1-\tilde{S})T_p\right)\mathbb{E}\left\{\exp(I_{+,-})z_p^+z_p^-\right\}+d^{-1}\,\mathbb{E}\left\{\exp(I_{+,-})\right\}\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -\frac{\tilde{S}KZ_c}{\tilde{E}_0}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left(-\tilde{S}-(1-\tilde{S})T_q\right)T_p+T_pT_q\right) \\
&\quad -\frac{\tilde{S}KZ_c}{\tilde{E}_0}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\left(-\tilde{S}-(1-\tilde{S})T_p\right)T_p+1\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tilde{S}T_p(T_q-1)-(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(-\tilde{S}T_p-(1-\tilde{S})T_p^2+1\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}
$$

Similarly, for T_3, we have

$$
\begin{aligned}
&\mathrm{T}_3([\boldsymbol{Q}_1]_{p,q}) \\
&= -K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+,\boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^-\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}z_p^+z_q^-d\right\} \\
&= -K\,\mathbb{E}\left\{\frac{\tilde{S}\tilde{E}_{-,+}}{\tilde{E}_0}\left(1-\tilde{S}\tilde{\delta}_{+,+}-(1-\tilde{S})\tilde{\delta}_{T+,+}+\tilde{\delta}_{-,+}\right)z_p^+z_q^-d\right\}\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -\frac{\tilde{S}K}{\tilde{E}_0}\sum_{i\neq j}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{i,j}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\,\mathbb{E}\left\{\tilde{E}_{-,+}\left(\left(-\tilde{S}-(1-\tilde{S})T_i\right)z_i^+z_j^++z_i^-z_j^+\right)z_p^+z_q^-d\right\}\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -\frac{\tilde{S}K}{\tilde{E}_0}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\left(\left(-\tilde{S}-(1-\tilde{S})T_p\right)Z_cT_q/d+Z_cT_pT_q/d\right) \\
&\quad -\frac{\tilde{S}K}{\tilde{E}_0}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}\left(\left(-\tilde{S}-(1-\tilde{S})T_q\right)Z_cT_q/d+Z_c/d\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= -(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tilde{S}T_q(T_p-1)-(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(1-\tilde{S}T_q-(1-\tilde{S})T_q^2\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}
$$

Combine these together and we obtain

$$
\begin{aligned}
[\boldsymbol{Q}_1]_{p,q} = &-\tilde{S}(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}+[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}(2-T_p-T_q) \\
&-(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tilde{S}\left(2T_pT_q-T_p-T_q\right) \\
&-(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(2-\tilde{S}(T_p+T_q)-(1-\tilde{S})(T_p^2+T_q^2)\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}
$$

□

Proof of Lemma C.6. We write

$$
\begin{aligned}
\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}} &:= \mathbb{E}\left\{\left(2-S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+,\boldsymbol{x}_{\boldsymbol{B}}^+)-S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+,\boldsymbol{x}_{\boldsymbol{B}}^+)\right)\boldsymbol{\xi}_{\boldsymbol{B}}^+(\boldsymbol{z}^+)^\top d\right\} \\
&\quad -K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+,\boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}\boldsymbol{\xi}_{\boldsymbol{B}}^-(\boldsymbol{z}^+)^\top d\right\}-K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+,\boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^-\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}\boldsymbol{\xi}_{\boldsymbol{B}}^+(\boldsymbol{z}^-)^\top d\right\} \\
&=: \mathrm{T}_1\left(\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}\right)+\mathrm{T}_2\left(\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}\right)+\mathrm{T}_3\left(\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}\right).
\end{aligned}
$$

We will use the fact that if some quantity X does not depend on $\boldsymbol{\xi}$, then $\mathbb{E}\{X\boldsymbol{\xi}\} = 0$ to simplify these terms. For T_1, by Lemma C.3, we have

$$\begin{aligned}
\mathrm{T}_1\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right) &= -\mathbb{E}\left\{\left(S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) + S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^+]_p z_q^+ d\right\} \\
&= -\tilde{S}(1-\tilde{S})\,\mathbb{E}\left\{\left(2\delta_{+,\boldsymbol{\xi}+} + 2\delta_{\boldsymbol{\xi}+,+} - \delta_{\boldsymbol{\xi}+,T+} - \delta_{T+,\boldsymbol{\xi}+}\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^+]_p z_q^+ d\right\} \\
&\quad \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right) \\
&= -\tilde{S}(1-\tilde{S})\,\mathbb{E}\left\{\left(2\delta_{+,\boldsymbol{\xi}+} - \delta_{T+,\boldsymbol{\xi}+}\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^+]_p z_q^+ d\right\} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).
\end{aligned}$$

Recall that

$$\delta_{+,\xi+} = \sum_{i,j} \frac{[\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_{i,j}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} z_i^+ [\boldsymbol{\xi}_{\boldsymbol{B}}^+]_j, \quad \delta_{T+,\boldsymbol{\xi}+} = \sum_{i,j} \frac{[\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_{i,j}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} T_i z_i^+ [\boldsymbol{\xi}_{\boldsymbol{B}}^+]_j.$$

Hence, we can further rewrite $\mathrm{T}_1\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right)$ as

$$\begin{aligned}
\mathrm{T}_1\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right) = -\tilde{S}(1-\tilde{S}) \sum_{i,j} \frac{[\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_{i,j}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}} (2 - T_i)\,\mathbb{E}\left\{z_i^+ [\boldsymbol{\xi}_{\boldsymbol{B}}^+]_j [\boldsymbol{\xi}_{\boldsymbol{B}}^+]_p z_q^+ d\right\} \\
\pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).
\end{aligned}$$

Note that the summand is nonzero only if $i = q$ and $j = p$. Thus,

$$\mathrm{T}_1\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right) = -\tilde{S}(1-\tilde{S})(2 - T_q)\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_q, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_p\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Similarly, we also have

$$\begin{aligned}
\mathrm{T}_2\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right) &= -\frac{K\tilde{S}}{\tilde{E}}\,\mathbb{E}\left\{\tilde{E}_{+,-}\delta_{+,\boldsymbol{\xi}-}[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_p z_q^+ d\right\} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right), \\
\mathrm{T}_3\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right) &= -\frac{K\tilde{S}}{\tilde{E}}\,\mathbb{E}\left\{\tilde{E}_{-,+}\left(\delta_{-,\boldsymbol{\xi}+} - \tilde{S}\delta_{+,\boldsymbol{\xi}+} - (1-\tilde{S})\delta_{T+,\boldsymbol{\xi}+}\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_p z_q^+ d\right\} \\
&\quad \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).
\end{aligned}$$

Then, we write

$$\mathbb{E}\left\{\tilde{E}_{+,-}\delta_{+,\boldsymbol{\xi}-}[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_p z_q^+ d\right\} = \sum_{i,j} \frac{[\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_{i,j}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\,\mathbb{E}\left\{\tilde{E}_{+,-} z_i^+ [\boldsymbol{\xi}_{\boldsymbol{B}}^-]_j [\boldsymbol{\xi}_{\boldsymbol{B}}^-]_p z_q^+ d\right\}.$$

If $j \neq p$, clear that that summand is 0. If $i \neq q$, then by flipping the sign of $z_q^\pm$ simultaneously, we flip the sign of $\tilde{E}_{+,-} z_i^+ z_q^+$. Therefore, when $i \neq q$, the summand is also 0. Thus,

$$\begin{aligned}
\mathrm{T}_2\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right) &= -\frac{\tilde{S}KZ_c}{\tilde{E}}\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_q, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_p\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right) \\
&= -(1-\tilde{S})\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_q, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_p\rangle}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).
\end{aligned}$$

Similarly, for T_3, we compute

$$\begin{aligned}
&\mathbb{E}\left\{\tilde{E}_{-,+}\left(\delta_{-,\boldsymbol{\xi}+} - \tilde{S}\delta_{+,\boldsymbol{\xi}+} - (1-\tilde{S})\delta_{T+,\boldsymbol{\xi}+}\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_p z_q^+ d\right\} \\
&= \sum_{i,j} \frac{[\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_{i,j}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\,\mathbb{E}\left\{\tilde{E}_{-,+}\left(z_i^-[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_j - \tilde{S}z_i^+[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_j - (1-\tilde{S})(1-T_i)z_i^-[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_j\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_p z_q^+ d\right\} \\
&= \frac{[\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_{q,p}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}}}\,\mathbb{E}\left\{\tilde{E}_{-,+}\left(z_q^- - \tilde{S}z_q^+ - (1-\tilde{S})(1-T_i)z_q^-\right)z_q^+\right\} \\
&= Z_c\frac{[\boldsymbol{K}_{\boldsymbol{A}}^\top \boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_{q,p}}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}\left(\tilde{S}(T_q - 1 - T_q^2) + T_q^2\right).
\end{aligned}$$

Thus,

$$\mathrm{T}_3\left([\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{p,q}\right) = -(1-\tilde{S})\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_q, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\tilde{S}(T_q - 1 - T_q^2) + T_q^2\right) \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Combine these together, rearrange terms and we obtain

$$[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}]_{p,q} = -(1-\tilde{S})\left(1+\tilde{S}+(1-\tilde{S})T_q^2\right)\frac{\langle [\boldsymbol{K}_{\boldsymbol{A}}]_q, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

To obtain the formula for $\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}$, it suffices to interchange the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$. □

Proof of Lemma C.7. Recall that

$$\begin{aligned}\boldsymbol{Q}_2 &:= \mathbb{E}\left\{\left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right)\boldsymbol{\xi}_{\boldsymbol{B}}^+(\boldsymbol{\xi}_{\boldsymbol{A}}^+)^\top d\right\} \\ &\quad - K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}\boldsymbol{\xi}_{\boldsymbol{B}}^-(\boldsymbol{\xi}_{\boldsymbol{A}}^+)^\top d\right\} - K\,\mathbb{E}\left\{\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\boldsymbol{f}_{\boldsymbol{A}}^-\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}\boldsymbol{\xi}_{\boldsymbol{B}}^+(\boldsymbol{\xi}_{\boldsymbol{A}}^-)^\top d\right\} \\ &=: \mathrm{T}_1(\boldsymbol{Q}_2) + \mathrm{T}_2(\boldsymbol{Q}_2) + \mathrm{T}_3(\boldsymbol{Q}_2).\end{aligned}$$

We now estimate each of these three terms. Again, the strategy is to leverage the symmetry of the distribution of $\boldsymbol{\xi}$ to argue that some part of the expectation is 0. For T_1, we have

$$\mathrm{T}_1([\boldsymbol{Q}_2]_{p,q}) = -\tilde{S}(1-\tilde{S})\,\mathbb{E}\left\{\left(2\delta_{+,\boldsymbol{\xi}+} + 2\delta_{\boldsymbol{\xi}+,+} - \delta_{\boldsymbol{\xi}+,T+} - \delta_{T+,\boldsymbol{\xi}+}\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^+]_p[\boldsymbol{\xi}_{\boldsymbol{A}}^+]_q d\right\} \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Note that none of these δ's depends on both $\boldsymbol{\xi}_{\boldsymbol{A}}^+$ and $\boldsymbol{\xi}_{\boldsymbol{B}}^+$. Therefore, the first term is 0. Similarly, for T_2, we have

$$\begin{aligned}\mathrm{T}_2([\boldsymbol{Q}_2]_{p,q}) &= -\frac{K\tilde{S}}{\tilde{E}}\,\mathbb{E}\left\{\left(\tilde{E}_{+,-}\left(\delta_{+,\boldsymbol{\xi}-} + \delta_{\boldsymbol{\xi}+,-} - \tilde{S}\left(\delta_{+,\boldsymbol{\xi}+} + \delta_{\boldsymbol{\xi}+,+}\right) - (1-\tilde{S})\delta_{\boldsymbol{\xi}+,T+}\right)\right)[\boldsymbol{\xi}_{\boldsymbol{B}}^-]_p[\boldsymbol{\xi}_{\boldsymbol{A}}^+]_q d\right\} \\ &\quad \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right) \\ &= \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).\end{aligned}$$

The same is also true for T_3. Thus,

$$[\boldsymbol{Q}_2]_{p,q} = \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

□

Proof of Lemma C.8. Recall from Corollary A.4 that Q_0 is defined as

$$\begin{aligned}Q_0 := &-\mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^+,\boldsymbol{x}_{\boldsymbol{B}}^+}\left\{\left(2 - S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+) - S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\right)\left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^+\right\rangle\right\} \\ &+ K\mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^+,\boldsymbol{x}_{\boldsymbol{B}}^\pm}\left\{\frac{S_{\boldsymbol{A}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^-)}{\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}\left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^-\right\rangle\right\} \\ &+ K\mathop{\mathbb{E}}_{\boldsymbol{x}_{\boldsymbol{A}}^\pm,\boldsymbol{x}_{\boldsymbol{B}}^+}\left\{\frac{S_{\boldsymbol{B}}(\boldsymbol{x}_{\boldsymbol{A}}^+, \boldsymbol{x}_{\boldsymbol{B}}^+)\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^-\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}{\exp(\tau_t^2\boldsymbol{f}_{\boldsymbol{A}}^+\cdot\boldsymbol{f}_{\boldsymbol{B}}^+)}\left\langle \boldsymbol{f}_{\boldsymbol{A}}^-, \boldsymbol{f}_{\boldsymbol{B}}^+\right\rangle\right\}.\end{aligned}$$

We write

$$\begin{aligned}\left\langle \boldsymbol{f}_{\boldsymbol{A}}^+, \boldsymbol{f}_{\boldsymbol{B}}^-\right\rangle &= \frac{\langle \boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{z}^+, \boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{z}^-\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}} \pm O\left(\kappa_0 d\delta_{N/S}\delta_{\xi,\perp}\right) \\ &= \sum_{i,j\in[r]}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{i,j}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}z_i^+z_j^- \pm O\left(\kappa_0 d\delta_{N/S}\delta_{\xi,\perp}\right).\end{aligned}$$

Therefore, we have

$$\begin{aligned}Q_0 &= -\sum_{i,j\in[r]}\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{i,j}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}}[\boldsymbol{Q}_1]_{i,j} \pm O\left(\kappa_0 d\delta_{N/S}\delta_{\xi,\perp}\right) \\ &= -\sum_{k=1}^{r}\frac{\kappa_k^2}{\|\boldsymbol{\kappa}\|^2}[\boldsymbol{Q}_1]_{k,k} \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + \kappa_0 d\delta_{N/S}\delta_{\xi,\perp}\right).\end{aligned}$$

□

Proof of Corollary C.9. Recall from Lemma A.5 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 = 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p,[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_p^2 + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_p^2$$
$$=: \sum_{i=1}^{3}\mathrm{T}_i\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\right).$$

By Lemma C.5, we have

$$\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\right) = 2\sum_{k=1}^{r}\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\right\|}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\right\rangle[\boldsymbol{Q}_1]_{k,p}\sigma_p^2$$
$$= \frac{2\sigma_p^2\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\|}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle[\boldsymbol{Q}_1]_{p,p}$$
$$\pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\kappa_0 d\delta_{\boldsymbol{AB},\perp}^2\right).$$

By Lemma C.6, we have

$$\mathrm{T}_2\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2\right) = 2\sum_{k=1}^{d-r}\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k\right\|}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k}\right\rangle[\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{k,p}\sigma_p^2$$
$$= O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{N/S}^2\delta_{\xi,\perp}^2\right).$$

Combine these together and we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 = \frac{2\sigma_p^2\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\|}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle[\boldsymbol{Q}_1]_{p,p} + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_p^2$$
$$\pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\kappa_0 d\delta_{\boldsymbol{AB},\perp}^2\right).$$

To get the formula for $\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\|^2$, it suffices to interchange the roles of $\boldsymbol{A}$ and $\boldsymbol{B}$. Similarly, for the noises, we write

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2 = \frac{2\left\langle[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q,[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top]_q\right\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + \frac{2\langle[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q,[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_2]_q\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\sigma_\xi^2 + \frac{2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|_F^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2$$
$$= \sum_{i=1}^{3}\mathrm{T}_i\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\right).$$

By Lemma C.6, we have

$$\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\right) = 2\sum_{k=1}^{r}\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\right\|}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\right\rangle[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,k}\sigma_\xi^2$$
$$= O\left(\frac{\sigma_\xi^2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).$$

By Lemma C.7, we have

$$\mathrm{T}_1\left(\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2\right) = 2\sum_{k=1}^{d-r}\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k\right\|}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q},\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k}\right\rangle[\boldsymbol{Q}_2]_{k,q}\sigma_\xi^2$$
$$= O\left(\frac{\sigma_\xi^2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d^4\left(\delta_{\boldsymbol{AB},\perp}+\delta_{N/S}\right)\delta_{\xi,\perp}^2\delta_{N/S}\right).$$

Combine these together and we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2 = \frac{2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).$$

□

C.2 MAINTAINING THE ORTHOGONALITY

In this subsection, we control the growth of $\delta_{\boldsymbol{AB},\perp}$ and $\delta_{\xi,\perp}$. First, we derive the equations that govern the evolution of the off-diagonal terms.

Lemma C.10. *In Stage 2, for any $p \neq q \in [r]$, we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle = {} & [\boldsymbol{Q}_1]_{q,p}\frac{\kappa_q/\kappa_p\sigma_p^2 + \kappa_p/\kappa_q\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ & - \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle \left(\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + \frac{[\boldsymbol{Q}_1]_{q,q}\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right) \\ & + \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_A}]_q}\right\rangle \frac{[\boldsymbol{Q}_1]_{q,q}\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle \frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ & \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} d\kappa_0\delta_{\boldsymbol{AB},\perp}\left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp}\right)\right).\end{aligned}$$

Lemma C.11. *In Stage 2, for any $p \neq q \in [r]$, we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_A}]_q}\right\rangle = {} & [\boldsymbol{Q}_1]_{q,p}\frac{\kappa_q/\kappa_p\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + [\boldsymbol{Q}_1]_{p,q}\frac{\kappa_p/\kappa_q\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ & - \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_A}]_q}\right\rangle \left(\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + \frac{[\boldsymbol{Q}_1]_{q,q}\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right) \\ & + \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle \frac{[\boldsymbol{Q}_1]_{q,q}\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_A}]_q}\right\rangle \frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ & \pm \left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} d\kappa_0\delta_{\boldsymbol{AB},\perp}\left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp}\right)\right),\end{aligned}$$

and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle = {} & [\boldsymbol{Q}_1]_{q,p}\frac{\kappa_q/\kappa_p\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + [\boldsymbol{Q}_1]_{p,q}\frac{\kappa_p/\kappa_q\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ & - \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle \left(\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + \frac{[\boldsymbol{Q}_1]_{q,q}\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right) \\ & + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_A}]_q}\right\rangle \frac{[\boldsymbol{Q}_1]_{q,q}\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} + \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle \frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ & \pm \left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} d\kappa_0\delta_{\boldsymbol{AB},\perp}\left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp}\right)\right).\end{aligned}$$

Now, we are ready to control the off-diagonalness. The proof is similar to the one of Lemma B.15.

Lemma C.12 (Orthogonality between signals)**.** *For any $p \neq q$, define*

$$\hat{\delta}_{\perp,p,q} := \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle^2 + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_A}]_q}\right\rangle^2 + \left\langle \overline{[\boldsymbol{K_A}]_p}, \overline{[\boldsymbol{K_A}]_q}\right\rangle^2 + \left\langle \overline{[\boldsymbol{K_B}]_p}, \overline{[\boldsymbol{K_B}]_q}\right\rangle^2.$$

In Stage 2, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\perp,p,q} \leq O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\kappa_0\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_{\boldsymbol{AB},\perp}\sqrt{\delta_-} + \delta_-\right)\sqrt{\hat{\delta}_{\perp,p,q}}\right).$$

Note that the LHS is of order $\delta_{\boldsymbol{AB},\perp}^2$ and for the RHS, the only term that can potentially have the same order is the $\left(\delta_-\sqrt{\hat{\delta}_{\perp,p,q}}\right)$-relatd term. We will show later that $\delta_- = o(\delta_{\boldsymbol{AB},\perp})$, whence this is also a higher order term.

Then, we consider the orthogonality between signals and noises.

Lemma C.13 (Orthogonality between signals and noises)**.** *For any $p \in [r]$ and $q \in [d-r]$, define*

$$\hat{\delta}_{\xi,\perp,p,q} := \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q} \right\rangle^2 + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q} \right\rangle^2 .$$

In Stage 2, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\xi,\perp,p,q} \le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\kappa_0 \delta_{\xi,\perp}^2 \left(\delta_{\xi,\perp} + \sqrt{\delta_-} \right) \right).$$

Finally, we deal with the orthogonality between the noises.

Lemma C.14 (Orthogonality between noises)**.** *In Stage 2, for any $p, q \in [d-r]$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t} \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q} \right\rangle = \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\delta_{\xi,\perp}^2 \right).$$

OMITTED PROOF OF THIS SUBSECTION

Proof of Lemma C.10. Recall that

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right)^\top \right) \left(\frac{[\boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} + \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \right) \frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d},$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \left(\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right)^\top \right) \left(\frac{[\boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{Q}_1^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} + \frac{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_q}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \right) \frac{\sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}.$$

Then, we write

$$\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle = \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \left(\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right)^\top \right) \frac{[\boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{Q}_1^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \frac{\sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \left(\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right)^\top \right) \frac{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_q}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \frac{\sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d}.$$

For the second term, by Lemma C.6, we have

$$\begin{aligned}
&\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \left(\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right)^\top \right) \frac{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}} \boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_q}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \\
&= \sum_{k=1}^{d-r} \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_k} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_k} \right\rangle \right) \frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_k\| \, [\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_{k,q}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \\
&= \pm O\left(d\delta_{\xi,\perp}^2 \delta_{N/S}^2 \right).
\end{aligned}$$

Hence,

$$\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle = \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \left(\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right)^\top \right) \frac{[\boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{Q}_1^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \frac{\sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\delta_{\xi,\perp}^2 \delta_{N/S}^2 \right).$$

For the first term, we have

$$\begin{aligned}
&\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \left(\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right)^\top \right) \frac{[\boldsymbol{K}_{\boldsymbol{A}} \boldsymbol{Q}_1^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \\
&\quad = \sum_{k=1}^{r} \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k} \right\rangle \right) \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\| \, [\boldsymbol{Q}_1]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|}.
\end{aligned}$$

When $k \notin \{p, q\}$, we have $\left|\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k} \right\rangle\right| \leq O(\delta_{\boldsymbol{AB},\perp})$, $\left|\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k} \right\rangle\right| \leq O(\delta_{\boldsymbol{AB},\perp})$, and $\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\| \,/\, \|[\boldsymbol{K}_{\boldsymbol{B}}]_q\| \leq \kappa_0$. Meanwhile, by Lemma C.5, we also have $|[\boldsymbol{Q}_1]_{q,k}| \leq O(\delta_{\boldsymbol{AB},\perp})$. Hence,

$$\left| \sum_{k \notin \{p,q\}} \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k} \right\rangle \right) \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\| \, [\boldsymbol{Q}_1]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \right| \leq O\left(d\kappa_0 \delta_{\boldsymbol{AB},\perp}^2\right).$$

Therefore,

$$\begin{aligned}
&\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t} \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left(\frac{\sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \right)^{-1} \\
= &\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right\rangle \right) \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\| \, [\boldsymbol{Q}_1]_{q,p}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \\
&+ \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \right) \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\| \, [\boldsymbol{Q}_1]_{q,q}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \\
&\pm O\left(d\kappa_0 \delta_{\boldsymbol{AB},\perp}^2\right) \pm O\left(d\delta_{\xi,\perp}^2 \delta_{N/S}^2\right).
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right\rangle \right) \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\| \, [\boldsymbol{Q}_1]_{q,p}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \\
= &\left(1 \pm \delta_{\boldsymbol{AB},\perp}^2\right) \frac{\kappa_p [\boldsymbol{Q}_1]_{q,p}}{\kappa_q} \left(1 \pm \sqrt{\delta_-}\right) \\
= &\frac{\kappa_p [\boldsymbol{Q}_1]_{q,p}}{\kappa_q} \pm O\left(\kappa_0 \delta_{\boldsymbol{AB},\perp} \sqrt{\delta_-}\right) \pm O\left(\kappa_0 \delta_{\boldsymbol{AB},\perp}^3\right).
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \right) \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\| \, [\boldsymbol{Q}_1]_{q,q}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|} \\
= &\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle (1 \pm \delta_-) \right) [\boldsymbol{Q}_1]_{q,q} \left(1 \pm \sqrt{\delta_-}\right) \\
= &\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \right) [\boldsymbol{Q}_1]_{q,q} \pm O\left(\delta_\perp \sqrt{\delta_-}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t} \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \left(\frac{\sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \right)^{-1} \\
= &\frac{\kappa_p [\boldsymbol{Q}_1]_{q,p}}{\kappa_q} \pm O\left(\kappa_0 \delta_{\boldsymbol{AB},\perp} \sqrt{\delta_-}\right) \pm O\left(\kappa_0 \delta_{\boldsymbol{AB},\perp}^3\right) \\
&+ \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \right) [\boldsymbol{Q}_1]_{q,q} \pm O\left(\delta_\perp \sqrt{\delta_-}\right) \\
&\pm O\left(d\kappa_0 \delta_{\boldsymbol{AB},\perp}^2\right) \pm O\left(d\delta_{\xi,\perp}^2 \delta_{N/S}^2\right) \\
= &\frac{\kappa_p [\boldsymbol{Q}_1]_{q,p}}{\kappa_q} + \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \right) [\boldsymbol{Q}_1]_{q,q} \\
&\pm O\left(d\kappa_0 \delta_{\boldsymbol{AB},\perp} \left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp}\right)\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t} \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle = &\frac{\kappa_p [\boldsymbol{Q}_1]_{q,p}}{\kappa_q} \frac{\sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \right) \frac{[\boldsymbol{Q}_1]_{q,q} \sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \\
&\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\kappa_0 \delta_{\boldsymbol{AB},\perp} \left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp}\right) \right).
\end{aligned}$$

Interchange the roles of $\boldsymbol{A}, \boldsymbol{B}, p, q$, replace $\boldsymbol{Q}_1$ with $\boldsymbol{Q}_1^\top$, and we obtain

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle = \frac{\kappa_q [\boldsymbol{Q}_1]_{q,p}}{\kappa_p} \frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \right) \frac{[\boldsymbol{Q}_1]_{p,p} \sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\kappa_0 \delta_{\boldsymbol{AB},\perp} \left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp} \right) \right).$$

Combine these together and we complete the proof. □

Proof of Lemma C.11. Similar to the proof of the previous lemma, we compute

$$\begin{aligned}\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle &= \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right)^\top \right) \left(\frac{[\boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} + \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}} \boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \right) \frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \\ &= \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right)^\top \right) \frac{[\boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \frac{\sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\delta_{\xi,\perp}^2 \delta_{N/S}^2 \right).\end{aligned}$$

Again, we have

$$\begin{aligned}&\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right)^\top \right) \frac{[\boldsymbol{K}_{\boldsymbol{B}} \boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \\ &= \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right)^\top \right) \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p} \frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\| \, [\boldsymbol{Q}_1]_{p,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \\ &\quad + \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right)^\top \right) \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\| \, [\boldsymbol{Q}_1]_{q,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \\ &\quad + \sum_{k \notin \{p,q\}} \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right)^\top \right) \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k} \frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\| \, [\boldsymbol{Q}_1]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} \\ &= \frac{\kappa_q [\boldsymbol{Q}_1]_{q,p}}{\kappa_p} + \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \right) [\boldsymbol{Q}_1]_{p,p} \pm \left(d\kappa_0 \delta_{\boldsymbol{AB},\perp} \left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp} \right) \right).\end{aligned}$$

Therefore,

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle = [\boldsymbol{Q}_1]_{q,p} \frac{\kappa_q/\kappa_p \sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \right) \frac{[\boldsymbol{Q}_1]_{p,p} \sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm \left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\kappa_0 \delta_{\boldsymbol{AB},\perp} \left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp} \right) \right).$$

Interchange the roles of p, q and we obtain

$$\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle = [\boldsymbol{Q}_1]_{p,q} \frac{\kappa_p/\kappa_q \sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} \right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \right) \frac{[\boldsymbol{Q}_1]_{q,q} \sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \pm \left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\kappa_0 \delta_{\boldsymbol{AB},\perp} \left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp} \right) \right).$$

Combine these together and we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle &= [\boldsymbol{Q}_1]_{q,p} \frac{\kappa_q/\kappa_p \sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + [\boldsymbol{Q}_1]_{p,q} \frac{\kappa_p/\kappa_q \sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \\ &\quad - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \left(\frac{[\boldsymbol{Q}_1]_{p,p} \sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + \frac{[\boldsymbol{Q}_1]_{q,q} \sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \right) \\ &\quad + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle \frac{[\boldsymbol{Q}_1]_{q,q} \sigma_q^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} + \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} \right\rangle \frac{[\boldsymbol{Q}_1]_{p,p} \sigma_p^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} \\ &\quad \pm \left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}} N_{\boldsymbol{B}} d} d\kappa_0 \delta_{\boldsymbol{AB},\perp} \left(\sqrt{\delta_-} + \delta_{\boldsymbol{AB},\perp} \right) \right).\end{aligned}$$

Interchange the roles of $\boldsymbol{A}, \boldsymbol{B}$, replace $\boldsymbol{Q}_1$ with $\boldsymbol{Q}_1^\top$, and we obtain the formula for $\frac{\mathrm{d}}{\mathrm{d}t} \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q} \right\rangle$. □

Proof of Lemma C.12. First, we consider the $[\boldsymbol{Q}_1]_{p,q}$-related terms, by Lemma C.5, we have

$$\begin{aligned}
[\boldsymbol{Q}_1]_{p,q} &= -\tilde{S}(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}+[\boldsymbol{K}_{\boldsymbol{BA}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}(2-T_p-T_q) \\
&\quad -(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\tilde{S}\left(2T_pT_q-T_p-T_q\right) \\
&\quad -(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{q,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(2-\tilde{S}(T_p+T_q)-(1-\tilde{S})(T_p^2+T_q^2)\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= O\left(\frac{\kappa_p\kappa_q\delta_{\boldsymbol{AB},\perp}}{N_A N_{\boldsymbol{B}}d}\right).
\end{aligned}$$

Hence,

$$[\boldsymbol{Q}_1]_{p,q}\frac{\kappa_q/\kappa_p\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} = O\left(\frac{\kappa_q^2}{N_A N_{\boldsymbol{B}}d}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\delta_{\boldsymbol{AB},\perp}\right) = O\left(\frac{1}{\sqrt{d}}\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\delta_{\boldsymbol{AB},\perp}\right).$$

The same bound also hold for other $[\boldsymbol{Q}_1]_{p,q}$-related terms. The important thing here is that we have and additional $1/\sqrt{d}$ factor.

Now, we are ready to prove the result. For notational simplicity, define $Z_1 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle$, $Z_2 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle$, $Z_3 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q}\right\rangle$, and $Z_4 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle$. Also define $G_p := [\boldsymbol{Q}_1]_{p,p}\sigma_p^2/(N_{\boldsymbol{A}}N_{\boldsymbol{B}}d)$ and similarly for G_q. Then, we can write the results of Lemma C.10 and Lemma C.11 as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} &= \begin{bmatrix} -G_p-G_q & 0 & G_q & G_p \\ 0 & -G_p-G_q & G_p & G_q \\ G_q & G_p & -G_p-G_q & 0 \\ G_p & G_q & 0 & -G_p-G_q \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} \\
&\quad \pm O\left(\frac{1}{\sqrt{d}}\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\delta_{\boldsymbol{AB},\perp}\right) \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}\left(\sqrt{\delta_-}+\delta_{\boldsymbol{AB},\perp}\right)\right).
\end{aligned}$$

The eigenvalues of the first matrix are $-2G_p, -2G_q, -2G_p-2G_q$ and 0. For the first three eigenvalues, note that

$$G_p = \frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \geq \frac{1}{\sqrt{d}}\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}.$$

Hence, the signals will dominate the noises, in particular, the first term on the second line, and push $\|\boldsymbol{Z}\|$ toward 0. Now we consider the eigen-pair $(0,(1,1,1,1))$, for which we will use the actual form of $[\boldsymbol{Q}_1]_{p,q}$ we obtained in Lemma C.5. We have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^4 Z_i &= 2\left([\boldsymbol{Q}_1]_{p,q}+[\boldsymbol{Q}_1]_{q,p}\right)\frac{\kappa_q/\kappa_p\sigma_p^2+\kappa_p/\kappa_q\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}\left(\sqrt{\delta_-}+\delta_{\boldsymbol{AB},\perp}\right)\right).
\end{aligned}$$

By Lemma C.5, we have

$$\begin{aligned}
&[\boldsymbol{Q}_1]_{p,q}+[\boldsymbol{Q}_1]_{q,p} \\
&\quad = -(1-\tilde{S})\frac{[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}+[\boldsymbol{K}_{\boldsymbol{BA}}]_{p,q}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(2-T_p^2-T_q^2+\tilde{S}(2-T_p-T_q)^2\right)\pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}$$

Then, we write

$$\begin{aligned}
[\boldsymbol{K}_{\boldsymbol{AB}}]_{p,q}+[\boldsymbol{K}_{\boldsymbol{BA}}]_{p,q} &= \|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_q\|\,Z_1+\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|\,\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|\,Z_2 \\
&= (Z_1+Z_2)\kappa_p\kappa_q(1\pm\sqrt{\delta_-}).
\end{aligned}$$

Hence,

$$[\boldsymbol{Q}_1]_{p,q} + [\boldsymbol{Q}_1]_{q,p} = -(1-\tilde{S})\frac{\kappa_p\kappa_q}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}(Z_1+Z_2)\left(2 - T_p^2 - T_q^2 + \tilde{S}(2-T_p-T_q)^2\right) \\ \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_{\boldsymbol{AB},\perp}\sqrt{\delta_-}\right).$$

To convert $Z_1 + Z_2$ to $\sum_{i=1}^4 Z_i$. Note that we have

$$Z_1 + Z_2 - Z_3 - Z_4 = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_q} - \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_q}\right\rangle \le \delta_-.$$

Therefore,

$$[\boldsymbol{Q}_1]_{p,q} + [\boldsymbol{Q}_1]_{q,p} = -\frac{1-\tilde{S}}{2}\frac{\kappa_p\kappa_q}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(2 - T_p^2 - T_q^2 + \tilde{S}(2-T_p-T_q)^2\right)\sum_{i=1}^4 Z_i \\ \pm O\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_{\boldsymbol{AB},\perp}\sqrt{\delta_-} + \delta_-\right).$$

Thus,

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=1}^4 Z_i = -(1-\tilde{S})\frac{\kappa_q/\kappa_p\sigma_p^2 + \kappa_p/\kappa_q\sigma_q^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\kappa_p\kappa_q}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(2 - T_p^2 - T_q^2 + \tilde{S}(2-T_p-T_q)^2\right)\sum_{i=1}^4 Z_i \\ \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\kappa_0\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_{\boldsymbol{AB},\perp}\sqrt{\delta_-} + \delta_-\right)\right).$$

Note that the coefficient of the first term is negative. Combine this with the previous bound for $\|\boldsymbol{Z}\|$, and we complete the proof. □

Proof of Lemma C.13. Recall that

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\left(\frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} + \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\right)\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d},$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q} = \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right)\left(\frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|} + \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_2]_q}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}.$$

We now compute $\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle$ and $\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle$ separately. First, we write

$$\begin{aligned}\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle &= \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}^\top\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}^\top\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &=: \mathrm{T}_1\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) + \mathrm{T}_2\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right).\end{aligned}$$

Then, we compute

$$\begin{aligned}&\mathrm{T}_1\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) \\ &= \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right\rangle\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|\,[\boldsymbol{Q}_1]_{p,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \sum_{k\neq p}\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right\rangle\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\right\rangle\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\|\,[\boldsymbol{Q}_1]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &= \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right\rangle\right)\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\xi,\perp}\left(\delta_{\xi,\perp} + \sqrt{\delta_-}\right)\right),\end{aligned}$$

and, by Lemma C.6, we have

$$\mathrm{T}_2\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) = \sum_{k=1}^{d-r} \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k[\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}$$
$$= \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} d\delta_{N/S}^2\delta_{\xi,\perp}^2\right).$$

Combine these together and we get

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle = \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right\rangle\right)\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}$$
$$\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} d\kappa_0\delta_{\xi,\perp}\left(\delta_{\xi,\perp} + \sqrt{\delta_-}\right)\right).$$

Then, we compute $\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle$. We write

$$\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle = \sum_{k=1}^{r} \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B}}]_k[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}$$
$$+ \sum_{k=1}^{d-r} \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k[\boldsymbol{Q}_2]_{k,q}}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}$$
$$=: \mathrm{T}_1\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) + \mathrm{T}_2\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right).$$

For T_1, when $k = p$, by Lemma C.6, we have

$$\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B}}]_k[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}$$
$$= \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|}[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,p}$$
$$= -(1-\tilde{S})\left(1 + \tilde{S} + (1-\tilde{S})T_p^2\right)\left(1 \pm O\left(\delta_- + \delta_{\xi,\perp}^2\right)\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle$$
$$\pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right)$$
$$= -\Theta\left(\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle \pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

When $k \neq p$, we have

$$\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B}}]_k[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|} = O\left(\kappa_0\delta_{\boldsymbol{AB},\perp}\delta_{\xi,\perp}\right).$$

Hence,

$$\mathrm{T}_1\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) = -\Theta\left(\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle$$
$$\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d^3\kappa_0\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Now we consider T_2. By Lemma C.7, we have

$$\mathrm{T}_2\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle\right) = \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d^4\delta_{\xi,\perp}^2\delta_{N/S}\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\right).$$

Therefore,

$$\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle = -\Theta\left(\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle$$
$$\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d^3\kappa_0\left(\delta_{\boldsymbol{AB},\perp} + \delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).$$

Thus,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle &= \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right\rangle\right)\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad - \Theta\left(\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle \\ &\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\xi,\perp}\left(\delta_{\xi,\perp}+\sqrt{\delta_-}\right)\right).\end{aligned}$$

Similarly, one can show that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle &= \left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right)\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad - \Theta\left(\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle \\ &\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\xi,\perp}\left(\delta_{\xi,\perp}+\sqrt{\delta_-}\right)\right).\end{aligned}$$

For notational simplicity, define

$$X = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle, \quad Y = \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q}\right\rangle, \quad C_1 = \frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}, \quad C_2 = \frac{\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}.$$

Then, we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}X\\Y\end{bmatrix} = \begin{bmatrix}-C_1 & C_1-\Theta(C_2)\\ C_1-\Theta(C_2) & -C_1\end{bmatrix}\begin{bmatrix}X\\Y\end{bmatrix} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\xi,\perp}\left(\delta_{\xi,\perp}+\sqrt{\delta_-}\right)\right).$$

The first matrix is negative semi-definite, whence

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(X^2+Y^2\right) \le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\xi,\perp}^2\left(\delta_{\xi,\perp}+\sqrt{\delta_-}\right)\right).$$

Since the roles of $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}$ and $\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}$ are interchangeable, the same bound also holds for $\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle$ and $\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle$. □

Proof of Lemma C.14. We write

$$\begin{aligned}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle &= \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}^\top\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right)^\top\right)\left(\frac{[\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|} + \frac{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_2^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|}\right)\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &= \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}^\top\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}^\top\left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right)^\top\right)\frac{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_2^\top]_q}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &=: \mathrm{T}_1\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle\right) + \mathrm{T}_2\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle\right).\end{aligned}$$

Then, we compute

$$\begin{aligned}&\mathrm{T}_1\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle\right) \\ &= \sum_{k=1}^r\left(\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k}\right\rangle - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}, \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_k}\right\rangle\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\|\,[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}.\end{aligned}$$

Note that, the first part of each summand is bounded by $O(\delta_{\xi,\perp})$. For the second part, by Lemma C.6 we also have

$$\begin{aligned}\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\|\,[\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|} &= \pm\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\|}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|}\left(O(1)\frac{\langle[\boldsymbol{K}_{\boldsymbol{B}}]_p,[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\pm O\left(d^3\left(\delta_{\boldsymbol{AB},\perp}+\delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right)\right)\\ &= \pm O(\delta_{\xi,\perp}).\end{aligned}$$

Hence,

$$\mathrm{T}_1\left(\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p},\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle\right)=O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).$$

Now we consider T_2. By Lemma C.7, We have

$$\begin{aligned}&\mathrm{T}_2\left(\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p},\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle\right)\\ &=\sum_{k=1}^{d-r}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}^\top\left(\boldsymbol{I}-\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\left(\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right)^\top\right)\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_k}\frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_k\|\,[\boldsymbol{Q}_2]_{q,k}}{\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\\ &=\pm\sum_{k=1}^{d-r}O(1)[\boldsymbol{Q}_2]_{q,k}\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\\ &=\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d^4\left(\delta_{\boldsymbol{AB},\perp}+\delta_{N/S}\right)\delta_{\xi,\perp}\delta_{N/S}\right).\end{aligned}$$

Combine this together and we obtain

$$\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p},\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle=\pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).$$

Similarly, one can derive the same bound for $\left\langle\frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle$ and complete the proof. □

C.3 Maintaining $\mathbf{K_A}\approx\mathbf{K_B}$

In this subsection, we show that $\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\approx 1$, $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|\approx\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|$, and also $\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|\approx\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|$ throughout Stage 2.

Lemma C.15. *In Stage 2, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle=\Omega\left(\frac{\sigma_p^2[\boldsymbol{Q}_1]_{p,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\left(1-\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right)\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right).$$

Lemma C.16. *Define $\rho_{\boldsymbol{A}/\boldsymbol{B},p}:=\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2/\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2$ and $\rho_{\boldsymbol{B}/\boldsymbol{A},p}:=\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2/\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2$. In Stage 2, we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p}+\rho_{\boldsymbol{B}/\boldsymbol{A},p}\right)=&\frac{2[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p},\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\left(2-\rho_{\boldsymbol{A}/\boldsymbol{B},p}-\rho_{\boldsymbol{B}/\boldsymbol{A},p}\right)\\ &\pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\kappa_0\delta_{\boldsymbol{AB},\perp}^2+\delta_-^2\right)\right).\end{aligned}$$

Lemma C.17. *For any $p,q\in[d-r]$, define $\rho_{\xi,\boldsymbol{A}/\boldsymbol{A},p/q}=\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p\|^2/\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2$ and $\rho_{\xi,\boldsymbol{A}/\boldsymbol{B},p/q}=\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p\|^2/\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\|^2$. In Stage 2, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\xi,\boldsymbol{A}/\boldsymbol{A},p/q}=\pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right),$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\xi,\boldsymbol{A}/\boldsymbol{B},p/q}+\rho_{\xi,\boldsymbol{B}/\boldsymbol{A},q/p}\right)=\pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\delta_{\xi,\perp}^2+\delta_-^2\right)\right).$$

OMITTED PROOF OF THIS SUBSECTION

Proof of Lemma C.15. First, we write

$$\begin{aligned}\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle &= \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &=: \mathrm{T}_1\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) + \mathrm{T}_2\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right).\end{aligned}$$

For T_1, we compute

$$\begin{aligned}\mathrm{T}_1\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) &= \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right) \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|\,[\boldsymbol{Q}_1]_{p,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad + \sum_{k\neq p}\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right) \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_k}\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_k\|\,[\boldsymbol{Q}_1]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &= \left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle^2\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|\,[\boldsymbol{Q}_1]_{p,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right).\end{aligned}$$

For T_2, we compute

$$\begin{aligned}\mathrm{T}_2\left(\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) &= \sum_{k=1}^{d-r}\overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}^\top \left(\boldsymbol{I} - \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\left(\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}\right)^\top\right) \frac{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_k[\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_{k,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &= \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{N/S}^2\delta_{\xi,\perp}^2\right).\end{aligned}$$

Therefore,

$$\left\langle \frac{\mathrm{d}}{\mathrm{d}t}\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle = \left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle^2\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|\,[\boldsymbol{Q}_1]_{p,p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right).$$

Then, by symmetry, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle &= \left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle^2\right)[\boldsymbol{Q}_1]_{p,p}\left(\frac{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|} + \frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|}\right)\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d} \\ &\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right) \\ &= \Omega\left(\frac{\sigma_p^2[\boldsymbol{Q}_1]_{p,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\left(1 - \left\langle \overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right).\end{aligned}$$

□

Proof of Lemma C.16. Similar to Stage 1, we define $\rho_{\boldsymbol{A}/\boldsymbol{B},p} = \|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 / \|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2$. By Lemma A.5, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\boldsymbol{A}/\boldsymbol{B},p} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2} - \rho_{\boldsymbol{A}/\boldsymbol{B},p}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2} \\
&= 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2 + 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2 + 2\frac{\rho_{\boldsymbol{A}/\boldsymbol{B},p}}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_p^2 \\
&\quad - \rho_{\boldsymbol{A}/\boldsymbol{B},p}\left(2\frac{\langle[\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_1^\top]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2 + 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2 + 2\frac{1}{N_{\boldsymbol{B}}^2 d}Q_0\sigma_p^2\right) \\
&= 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2 + 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{B}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2 \\
&\quad - \rho_{\boldsymbol{A}/\boldsymbol{B},p}\left(2\frac{\langle[\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_1^\top]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2 + 2\frac{\langle[\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\xi_{\boldsymbol{A}}}]_p\rangle}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d\,\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\sigma_p^2\right) \\
&\quad + 2Q_0\sigma_p^2\left(\frac{1}{N_{\boldsymbol{A}}^2 d} - \frac{1}{N_{\boldsymbol{B}}^2 d}\right)\rho_{\boldsymbol{A}/\boldsymbol{B},p}.
\end{aligned}$$

Then, by Lemma C.5 and Lemma C.6, we have

$$\begin{aligned}
\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}\boldsymbol{Q}_1]_p\rangle &= \langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle\,[\boldsymbol{Q}_1]_{p,p} \pm O\left(d\kappa_p^2\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right), \\
\left\langle[\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A}}\boldsymbol{Q}_1^\top]_p\right\rangle &= \langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle\,[\boldsymbol{Q}_1]_{p,p} \pm O\left(d\kappa_p^2\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right), \\
\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{B}}}]_p\rangle &= \pm O\left(d\kappa_p^2\delta_{N/S}^2\delta_{\xi,\perp}\right), \\
\langle[\boldsymbol{K}_{\boldsymbol{B}}]_p, [\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}\boldsymbol{Q}_{1,\boldsymbol{\xi}_{\boldsymbol{A}}}]_p\rangle &= \pm O\left(d\kappa_p^2\delta_{N/S}^2\delta_{\xi,\perp}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\boldsymbol{A}/\boldsymbol{B},p} &= 2\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2}\left(1 - \rho_{\boldsymbol{A}/\boldsymbol{B},p}\right) \\
&\quad + 2Q_0\sigma_p^2\left(\frac{1}{N_{\boldsymbol{A}}^2 d} - \frac{1}{N_{\boldsymbol{B}}^2 d}\right)\rho_{\boldsymbol{A}/\boldsymbol{B},p} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}$$

Interchange the roles of $\boldsymbol{A}, \boldsymbol{B}$ and we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\boldsymbol{B}/\boldsymbol{A},p} &= 2\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\frac{\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}\left(1 - \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) \\
&\quad + 2Q_0\sigma_p^2\left(\frac{1}{N_{\boldsymbol{B}}^2 d} - \frac{1}{N_{\boldsymbol{A}}^2 d}\right)\rho_{\boldsymbol{B}/\boldsymbol{A},p} \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) &= 2\frac{[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\langle[\boldsymbol{K}_{\boldsymbol{A}}]_p, [\boldsymbol{K}_{\boldsymbol{B}}]_p\rangle\left(\frac{1 - \rho_{\boldsymbol{A}/\boldsymbol{B},p}}{\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\|^2} + \frac{1 - \rho_{\boldsymbol{B}/\boldsymbol{A},p}}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}\right) \\
&\quad + 2Q_0\sigma_p^2\left(\frac{1}{N_{\boldsymbol{A}}^2 d} - \frac{1}{N_{\boldsymbol{B}}^2 d}\right)\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p} - \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right) \\
&= \frac{2[\boldsymbol{Q}_1]_{p,p}\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\left(2 - \rho_{\boldsymbol{A}/\boldsymbol{B},p} - \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) \\
&\quad \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\kappa_0\delta_{\boldsymbol{AB},\perp}^2 + \delta_-^2\right)\right).
\end{aligned}$$

□

Proof of Lemma C.17. For notational simplicity, we drop the subscript ξ in the proof. Recall from Corollary C.9 that

$$\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2 = \frac{2\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).$$

Hence, for any $p, q \in [d-r]$, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\boldsymbol{A}/\boldsymbol{A},p/q} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2} - \rho_{\boldsymbol{A}/\boldsymbol{A},p/q}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2} \\
&= \frac{2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2\rho_{\boldsymbol{A}/\boldsymbol{A},p/q} \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right) \\
&\quad - \rho_{\boldsymbol{A}/\boldsymbol{A},p/q}\left(\frac{2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right)\right) \\
&= \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\boldsymbol{A}/\boldsymbol{B},p/q} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\right\|^2} - \rho_{\boldsymbol{A}/\boldsymbol{B},p/q}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q\right\|^2} \\
&= \frac{2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2\rho_{\boldsymbol{A}/\boldsymbol{B},p/q} \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right) \\
&\quad - \rho_{\boldsymbol{A}/\boldsymbol{B},p/q}\left(\frac{2}{N_{\boldsymbol{B}}^2 d}Q_0\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right)\right) \\
&= \left(\frac{2}{N_{\boldsymbol{A}}^2 d} - \frac{2}{N_{\boldsymbol{B}}^2 d}\right)Q_0\sigma_\xi^2\rho_{\boldsymbol{A}/\boldsymbol{B},p/q} \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).
\end{aligned}$$

By symmetry, we also have

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_{\boldsymbol{B}/\boldsymbol{A},q/p} = \left(\frac{2}{N_{\boldsymbol{B}}^2 d} - \frac{2}{N_{\boldsymbol{A}}^2 d}\right)Q_0\sigma_\xi^2\rho_{\boldsymbol{B}/\boldsymbol{A},q/p} \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).$$

Hence,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p/q} + \rho_{\boldsymbol{B}/\boldsymbol{A},q/p}\right) &= \left(\frac{2}{N_{\boldsymbol{A}}^2 d} - \frac{2}{N_{\boldsymbol{B}}^2 d}\right)Q_0\sigma_\xi^2\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p/q} - \rho_{\boldsymbol{B}/\boldsymbol{A},q/p}\right) \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right) \\
&= \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\delta_{\xi,\perp}^2 + \delta_-^2\right)\right).
\end{aligned}$$

□

C.4 Controlling the noise-signal ratio

In this subsection, we show that the noise-signal ratio remains small throughout Stage 2.

Lemma C.18. *Let $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|$ be the smallest among all $\{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\|\}_{k\in[r]}$. For any $q \in [r]$, in Stage 2, we have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2} \le O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\delta_{\xi,\perp}^2 + d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\right)\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}.$$

Proof. Recall from Corollary C.9 that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 &= \frac{2\sigma_p^2\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|\left\|[\boldsymbol{K}_{\boldsymbol{B}}]_p\right\|}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle[\boldsymbol{Q}_1]_{p,p} + 2\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_p^2 \\
&\quad \pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\kappa_0 d\delta_{\boldsymbol{AB},\perp}^2\right) \\
\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2 &= \frac{2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{N_{\boldsymbol{A}}^2 d}Q_0\sigma_\xi^2 \pm O\left(\frac{\sigma_\xi^2\left\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\right\|^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).
\end{aligned}$$

Since the condition number of $\boldsymbol{K}_{\boldsymbol{A}}$ is bounded by $\sqrt{d}$, it suffices to consider the smallest $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|$, for which we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 &= \frac{2\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}[\boldsymbol{Q}_1]_{p,p}\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 + \frac{2\sigma_p^2}{N_{\boldsymbol{A}}^2 d}Q_0\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 \\ &\quad \pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\kappa_0 d\delta_{\boldsymbol{AB},\perp}^2 + \delta_-\right)\right) \\ &= \frac{2\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}[\boldsymbol{Q}_1]_{p,p}\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 - \frac{2\sigma_p^2}{N_{\boldsymbol{A}}^2 d}\left(\sum_{k=1}^r \frac{\kappa_k^2}{\|\boldsymbol{\kappa}\|^2}[\boldsymbol{Q}_1]_{k,k}\right)\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 \\ &\quad \pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + d^2\delta_{N/S}\delta_{\xi,\perp} + \delta_-\right)\right),\end{aligned}$$

where the last line comes from Lemma C.8. By Lemma B.4, $[\boldsymbol{Q}_1]_{p,p}$ is negative correlated with κ_p^2. As a result, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 \geq -O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + d^2\delta_{N/S}\delta_{\xi,\perp} + \delta_-\right)\right).$$

For $\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}$, we simply have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2 \leq O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\delta_{\xi,\perp}^2 + d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\right)\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2.$$

Thus,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2} - \frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2} \\ &\leq O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\delta_{\xi,\perp}^2 + d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}.\end{aligned}$$

□

C.5 Estimating the convergence rate

In this subsection, we estimate how fast the condition number will become close to 1.

Lemma C.19. *Suppose that $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|$ is the largest and $\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|$ the smallest among all $\{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\|\}_{k\in[r]}$. In Stage 2, we have*

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|^2} &\leq -\frac{4(1-\tilde{S})\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(T_p - T_q\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|^2} \\ &\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\kappa_0^2\right).\end{aligned}$$

Corollary C.20 (Convergence rate)**.** *Suppose that $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|$ is the largest and $\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|$ the smallest among all $\{\|[\boldsymbol{K}_{\boldsymbol{A}}]_k\|\}_{k\in[r]}$. For any constant $c > 1$, it takes at most $\mathrm{poly}(d)$ amount of time for $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2 / \|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|^2$ to become smaller than c.*

OMITTED PROOF OF THIS SUBSECTION

Proof of Lemma C.19. By Corollary C.9, Lemma C.8 and Lemma C.5, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 &= \frac{2\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}[\boldsymbol{Q}_1]_{p,p}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 + \frac{2\sigma_p^2}{N_{\boldsymbol{A}}^2 d}Q_0\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 \\
&\quad \pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\kappa_0 d\delta_{\boldsymbol{AB},\perp}^2 + \delta_-\right)\right) \\
&= \frac{4(1-\tilde{S})\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}(1-T_p)\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 + \frac{4(1-\tilde{S})\sigma_p^2}{N_{\boldsymbol{A}}^2 d}\left(\sum_{k=1}^r \frac{\kappa_k^2}{\|\boldsymbol{\kappa}\|^2}(1-T_p)\right)\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 \\
&\quad \pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\right) \\
&= -\frac{4(1-\tilde{S})\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(T_p - \tilde{T}\right)\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 \pm O\left(\frac{\sigma_p^2\kappa_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\right),
\end{aligned}$$

where $\tilde{T} = \sum_{k=1}^r \frac{\kappa_k^2}{\|\boldsymbol{\kappa}\|^2}T_p$. Then, we compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} &= \frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} - \frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2}\frac{\frac{\mathrm{d}}{\mathrm{d}t}\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} \\
&= -\frac{4(1-\tilde{S})}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(\sigma_p^2\left(T_p - \tilde{T}\right) - \sigma_q^2\left(T_q - \tilde{T}\right)\right)\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\kappa_0^2\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\right).
\end{aligned}$$

Since $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|$ is the largest, $\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|$ is the smallest, T_p is positively correlated with $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|$, and $\tilde{T}$ is a weighted average of T_p, we have

$$\sigma_p^2\left(T_p - \tilde{T}\right) - \sigma_q^2\left(T_q - \tilde{T}\right) \geq \sigma_q^2\left(T_p - T_q\right).$$

Thus,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} &\leq -\frac{4(1-\tilde{S})\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(T_p - T_q\right)\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2} \\
&\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\kappa_0^2\right).
\end{aligned}$$

□

Proof of Corollary C.20. Recall that

$$T_p = \tanh\left(\frac{\hat{\kappa}_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right) = \tanh\left(\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2}{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2}\left(1 \pm O(\delta_- + \delta_{N/S})\right)\right).$$

Since $\kappa_0 \leq \sqrt{d}$, we have $\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 / \left\|[\boldsymbol{K}_{\boldsymbol{A}}]_F\right\|^2 \leq 1/2$. Note that $\tanh'(z) = 1 - \tanh^2(z) = \Omega(1)$ for any $z \leq 1.1/2$. Therefore,

$$T_p - T_q \geq \Omega\left(\frac{\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\right\|^2 - \left\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\right\|^2}{\|\boldsymbol{K}_{\boldsymbol{A}}\|_F^2}\right) \pm O(\delta_- + \delta_{N/S}) \geq \Omega\left(\frac{1}{d}\right).$$

Then, by Lemma C.19, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|^2} &\le -\frac{4(1-\tilde{S})\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(T_p - T_q\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|^2} \\ &\quad \pm O\left(\frac{\sigma_p^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\kappa_0^2\right) \\ &\le -\Omega\left(\frac{\sigma_{\min}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|^2}.\end{aligned}$$

By the proof of Lemma C.19, the largest $\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2$ is non-increasing. Hence, $N_{\boldsymbol{A}}N_{\boldsymbol{B}}d$ is upper bounded by some $\mathrm{poly}(d)$. Thus, it takes at most $\mathrm{poly}(d)$ for $\left\|[\boldsymbol{K}_{\boldsymbol{A}}]_p^2\right\| / \|[\boldsymbol{K}_{\boldsymbol{A}}]_q\|$ to become smaller than c. □

C.6 Proof of the main lemma of Stage 2

Proof of Lemma C.1. The polynomial bound on the convergence time has been proved in Corollary C.20. For the errors, recall from Lemma C.12, Lemma C.13 and Lemma C.14 that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\perp,p,q} &\le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\kappa_0\left(d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_{\boldsymbol{AB},\perp}\sqrt{\delta_-} + \delta_-\right)\sqrt{\hat{\delta}_{\perp,p,q}}\right), \\ \frac{\mathrm{d}}{\mathrm{d}t}\hat{\delta}_{\xi,\perp,p,q} &\le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\xi,\perp}^2\left(\delta_{\xi,\perp} + \sqrt{\delta_-}\right)\right), \\ \frac{\mathrm{d}}{\mathrm{d}t}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B},\boldsymbol{\xi}}]_q}\right\rangle &= \pm O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right).\end{aligned}$$

Recall from Lemma C.15, Lemma C.16, and Lemma C.17 that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle &= \Omega\left(\frac{\sigma_p^2[\boldsymbol{Q}_1]_{p,p}}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\right)\left(1 - \left\langle\overline{[\boldsymbol{K}_{\boldsymbol{A}}]_p}, \overline{[\boldsymbol{K}_{\boldsymbol{B}}]_p}\right\rangle\right) \pm O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\kappa_0\delta_{\boldsymbol{AB},\perp}^2\right), \\ \frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\boldsymbol{A}/\boldsymbol{B},p} + \rho_{\boldsymbol{B}/\boldsymbol{A},p}\right) &\le O\left(\frac{\sigma_{\max}^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\kappa_0\delta_{\boldsymbol{AB},\perp}^2 + \delta_-^2\right)\right), \\ \frac{\mathrm{d}}{\mathrm{d}t}\rho_{\xi,\boldsymbol{A}/\boldsymbol{A},p/q} &\le O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}d\delta_{\xi,\perp}^2\right), \\ \frac{\mathrm{d}}{\mathrm{d}t}\left(\rho_{\xi,\boldsymbol{A}/\boldsymbol{B},p/q} + \rho_{\xi,\boldsymbol{B}/\boldsymbol{A},q/p}\right) &\le O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\delta_{\xi,\perp}^2 + \delta_-^2\right)\right).\end{aligned} \tag{17}$$

By Lemma C.18, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2} \le O\left(\frac{\sigma_\xi^2}{N_{\boldsymbol{A}}N_{\boldsymbol{B}}d}\left(d\delta_{\xi,\perp}^2 + d^2\delta_{\boldsymbol{AB},\perp}^2 + \delta_- + d^2\delta_{N/S}\delta_{\xi,\perp}\right)\right)\frac{\|[\boldsymbol{K}_{\boldsymbol{A},\boldsymbol{\xi}}]_q\|^2}{\|[\boldsymbol{K}_{\boldsymbol{A}}]_p\|^2}.$$

Note that on the RHS of these equations, the only terms whose order may potentially be smaller than or equal to LHS are the δ_--reltaed terms. However, (17), we can make sure δ_- is at most $\delta_{\boldsymbol{AB},\perp}^{1.5}$. As a result, the orders of the RHS are all greater than the orders of the LHS, which implies that these errors can at most double within $\mathrm{poly}(d)$ time if they are sufficiently small at the beginning of Stage 2. □

D From Gradient Flow to Gradient Descent

Converting the above gradient flow argument to a gradient descent one is standard. All our estimations can tolerate an inverse polynomially large error. Since the all quantities of interest here are polynomially and inverse polynomially bounded, at each step of gradient descent, one can always make the GF-to-GD discretization error sufficiently (inverse polynomially) small by choose

a sufficiently (inverse polynomially) small learning rate and generating sufficiently (polynomially) many samples. Since the times need for Stage 1 and Stage 2 are both polynomial, this also implies a polynomial sample complexity.