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 keeping 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 white color. When we hear the loud sound of the engine, we can imagine an expensive sports car passing nearby.

Recently, in machine learning, multimodal learning methods – training the model 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 the contrastive approach and other approaches is that contrastive loss not only requires the learned representations 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.

Despite the empirical success of this contrastive learning-based method, 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 \( f_A, f_B \) from modules \( A \) and \( B \) respectively. Due to the existence of negative pairs that emphasize negative correlations of \( f_A \) and \( f_B \), it is unclear that the model still has incentives to align the features from different modules.
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 $x_A = Az_A + A\xi_A$ and $x_B = Bz_B + B\xi_B$, respectively, where $z_A, z_B$ are considered as the hidden signals, $A, B$ linear transformations from the signal to the observation and $A\xi_A, B\xi_B$ 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 ($A = B$). The positive pair of the data shares the same signal $z_A = z_B$, and has different noises $\xi_A, \xi_B$ and transformations $A, B$. The goal is to learn features $f_A, f_B$ that align positive pairs while keeping representations of negative pairs away from each other.

Under this setting, we make the following contributions:

1. We consider the challenging (but more practical) setting where the features in $A$ and $B$ 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 $A$ and $B$ are exactly column orthonormal matrices even in the simpler single-modal setting ($A = B$).
2. We consider feature learners $f_A, f_B$ with normalization, meaning that $f_A, f_B$ are always normalized to have (expected) norm one during training. Output normalization plays a critical role 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 $f_A$ and $f_B$.
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 $f_A$ and $f_B$, the features learned by contrastive learning with negative pairs is much more uniform, meaning that $f_A, f_B$ can recover all the singular vectors of $A$ and $B$ and normalize them. On the other hand, without negative pairs, the learned representation is close to a rank one solution, meaning that $f_A, f_B$ will only focus on the top singular direction of $A$ and $B$.
5. We also perform simulations and more practical experiments 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 representations only span 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. Though these works do consider the optimization problem, they focus on the case where the features are uniform, and only Wen & Li (2021) considers output normalization. We compare our results 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 \( z \), 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, and 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

Similar to previous theoretical works (e.g., Lee et al. (2021); Wen & Li (2021)), we consider a linear data-generating model. Formally, we assume that the contrastive pairs \((x_A^+, x_B^-)\) are constructed as

\[
x_A^+ = A z^+ + A \xi_A^+, \quad x_B^- = B z^- + B \xi_B^-.
\]

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

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

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

\[
f_A(x_A) := \frac{W_A^\top x_A}{\sqrt{\mathbb{E}_{x_A} \left\| W_A^\top x_A \right\|^2}}, \quad f_B(x_B) := \frac{W_B^\top x_B}{\sqrt{\mathbb{E}_{x_B} \left\| W_B^\top x_A \right\|^2}},
\]

where \( W_A, W_B \in \mathbb{R}^{d \times m} \) are the trainable parameters. In words, we first map the inputs \((x_A, x_B)\) into the embedding space \( \mathbb{R}^m \) using \( W_A \) and \( W_B \), and then apply batch normalization to the outputs. By saying the learned representations are aligned, we mean that \( f_A \) and \( f_B \) 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 \( z \) does not change the representation dramatically. See Section 4 for formal definitions.

One can easily verify that, in the population case, we have \( \mathbb{E}_{x_A} \left\| W_A^\top x_A \right\|^2 = \left\| W_A^\top A \right\|^2_F / d + \left\| W_A^\top A \xi_{\cdot} \right\|^2_F / d \). For notational simplicity, we write \( K_A = W_A^\top A \), \( K_B = W_B^\top B \), \( K_A \xi_{\cdot} = W_A^\top A \xi_{\cdot} \).
In the non-contrastive case, we always use gradient descent over the following contrastive loss \( L \). First, we define2

\[
S_A(x_A^+, x_B^+) = \frac{\exp(\tau_t f_A^+ : f_B^+)}{\exp(\tau_t f_A^+ : f_B^+) + K E x_B^+ \exp(\tau_t f_A^- : f_B^-)},
\]

\[
S_B(x_A^+, x_B^+) = \frac{\exp(\tau_t f_A^+ : f_B^+)}{\exp(\tau_t f_A^+ : f_B^+) + K E x_B^- \exp(\tau_t f_A^- : f_B^-)},
\]

where \( K \) is a 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

\[
\mathcal{L} := \mathcal{L}_A + \mathcal{L}_B := - E \log S_A(x_A^+, x_B^+) - E \log S_B(x_A^+, x_B^+).
\]

By the non-contrastive loss, we mean

\[
\hat{\mathcal{L}} := - E \langle f_A^+, f_B^- \rangle.
\]

Formally, the training algorithm is defined as follows.

**Algorithm 3.1** (Training algorithm). Let \( \hat{\mathcal{L}} \) 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 pairs \( \{(x_A^+, x_B^+, x_A^-, x_B^-)\}_{i=1}^N \), use them to compute the empirical version of \( \hat{\mathcal{L}} \), and update the weight matrices using \( W_A \leftarrow W_A - \tau_t^{-2} \eta \nabla W_A \hat{\mathcal{L}} \) and \( W_B \leftarrow W_B - \tau_t^{-2} \eta \nabla W_B \hat{\mathcal{L}} \).

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

\[
\Gamma_{\text{Align}} := \frac{1}{2} \mathbb{E}_{x_A \sim x_A^+, x_B \sim x_B^+} \left\{ \mathbb{I} \left\{ \| f_A^+ - f_B^+ \| < \| f_A^+ - f_B^- \| \right\} + \mathbb{I} \left\{ \| f_A^- - f_B^- \| < \| f_A^+ - f_B^- \| \right\} \right\}.
\]

Namely, \( \Gamma_{\text{Align}} \) is the accuracy of classifying whether the input pair \( (x_A, x_B) \) is a positive pair. We say that the learned representations are aligned if \( \Gamma_{\text{Align}} \approx 1 \).

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

\(^1\)See Section D for discussions on the sample complexity.

\(^2\)We use \( f_A \) as a shorthand for \( f_A(x_A^+) \), similarly for other combinations of \( A, B \) and \( \pm \).

\(^3\)Note that, in the non-contrastive case, changing \( \tau_t \) only changes the learning rate.
Definition 4.2 (Balance). We define the balance score as
\[ \Gamma_{\text{Balance}} := \frac{\| \Sigma_f \|_F^2}{\| \Sigma \|_2^2} \text{ where } \Sigma_f := \mathbb{E}_X \{ f_A f_A^T \}. \]

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

We make some short remarks on Definition 4.2. First, we only use $f_A$ in this definition because if the embeddings are well-aligned, $f_A f_A^T$ and $f_B f_B^T$ should be approximately the same. Second, the stable rank is usually used as an alternative to the actual rank because it is less sensitive to small singular values (Rudelson & Vershynin (2007)). To see why this makes sense, note that the stable rank is usually used as an alternative to the actual rank because it is less sensitive to small singular values (Rudelson & Vershynin (2007)).

In Jing et al. (2022), the authors also use rank to measure the degree of “dimensional collapse”. Unlike their argument, here we only require $\Gamma_{\text{Balance}}$ to be at least $\alpha r$, instead of $\alpha m$, 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 to span the entire embedding space. Finally, note that a sufficient condition for the rank is still at most $\Gamma \leq \Gamma_{\text{Balance}}$, instead of $\alpha m$, 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 to span the entire embedding space.

Why balance representations are important? In our setting, one simple example of aligned but unbalanced representations is $W_A^F = \text{diag}(\nu)A^{-1}$ and $W_B^F = \text{diag}(\nu)B^{-1}$ with $\nu_1 = 1$ and $\nu_2 = \cdots = \nu_r = \exp(-d)$. This model maps inputs whose latent vector is $z$ to $\text{diag}(\mu)z$, up to some normalization, for both modules, whence it has $\Gamma_{\text{Align}} = 1$. Meanwhile, one can verify that this model has $\Gamma_{\text{Balance}} / r \leq 2 / r \approx 0$. The problem of this model is that it overly emphasizes $z_1$ and is sensitive to small changes in $z_1$. This model can be made balanced by replacing $\text{diag}(\mu)$ with $I_r$, in which case the model directly recovers the latent vector $z$. As a result, all changes in $z$ will be reflected in the final representation in a faithful way. Similar notions have 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 $m = \text{poly}(d)$ is sufficiently large, the learning rate $\eta = 1 / \text{poly}(d)$ is sufficiently small, and we generate sufficiently, but still polynomially, many samples at each step to compute the loss$^4$. Suppose that, for some small constant $c > 0$,
\[ \frac{\sigma_{\max}^2}{\sigma_{\min}^2} \max \left\{ 1, \frac{(d - r)\sigma_{\max}^2}{r\sigma_{\min}^2} \right\} \leq c \log d. \]

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

(b) Contrastive loss. There exists $\hat{\tau}_0^2 = 1 / \text{poly}(d)$ and $T_1 = \text{poly}(d)$ such that, for any valid $\sigma$, Algorithm 3.1 will reach a point after $\text{poly}(d)$ iterations at which the learned representations are aligned and balanced, that is, $\Gamma_{\text{Align}} \approx 1$ and $\Gamma_{\text{Balance}} / r \approx 1$.

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 $\Gamma_{\text{Align}} \approx 1$ is to learn $\| f_A - f_B \| \approx \| K_0 (z_A - z_B) \|$ for some $K_0 \in \mathbb{R}^{m \times r}$ with $K_0^T K_0 > 0$.

$^4$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.
In this case, the distance between positive pairs is always close to 0 while the distance between negative pairs is positive. Recall that the output of our model is \( f_A = (K_A x_A + K_A \xi A) / N_A \) and \( f_B = (K_B z_B + K_B \xi B) / N_B \). Hence, we have the following sufficient condition.

**Lemma 4.4** (Sufficient condition for aligned representations). If \( \|K_A\|_F \gg \|K_A \xi\|_F, \|K_B\|_F \gg \|K_B \xi\|_F, K_A \approx K_B, \) and the condition number of \( K_A \) is upper bounded by \( \text{poly}(d) \), 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 \( K_0 \), and the final condition makes sure the smallest singular value \( K_0^\top K_0 \) of is still at least \( 1/\text{poly}(d) \) 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

\[
\Sigma_f = \mathbb{E}_{x_A} \{ f_A f_A^\top \} = \mathbb{E}_z \left\{ \frac{K_A z z^\top K_A^\top}{\|K_A\|_F^2 / d} \right\} = \frac{K_A K_A^\top}{\|K_A\|_F^2}.
\]

Recall the relationship between \( \Gamma_{\text{Balance}} \) and the effective condition number of \( \Sigma_f \). 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 \( K_A K_A \) is close to 1, then the learned representations are balanced.

Note that, since the columns of \( A \) are not orthonormal, even at initialization, the condition number of \( K_A K_A \) 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 \( \tau_t^2 = 1/\text{poly}(d) \) and run gradient descent for \( \text{poly}(d) \) iterations. This stage is called Stage 1. Then, we set \( \tau_t^2 = 1 \) and run gradient descent for another \( \text{poly}(d) \) iterations. This stage is called Stage 2. The use of different \( \tau_t \) is mainly for technical purposes since this gives cleaner separation between stages. We observe similar stage-wise behavior when using a uniform \( \tau_t^2 \). 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 \( W_A \) and \( W_B \) directly, we will track \( K_A, K_A \xi, K_B, K_B \xi \), 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 \( K_A \) are governed by the following equation

\[
K_A = \mathbb{E}_{x_A^+, x_B^+} \left\{ \left( 2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+) \right) \left( \frac{f_B^+(z^+)^\top}{N_A} - \frac{f_A^+ f_B^+}{N_A^2 d} \right) \right\} \text{diag}(\sigma^2)
- K \mathbb{E}_{x_A^+, x_B^+} \left\{ \frac{S_A(x_A^+, x_B^+) \exp(\tau_t^2 f_A^+ f_B^+)}{\exp(\tau_t^2 f_A^+ f_B^+)} \left( \frac{f_B^+(z^+)^\top}{N_A} - \frac{f_A^+ f_B^+}{N_A^2 d} \right) \right\} \text{diag}(\sigma^2)
- K \mathbb{E}_{x_A^+, x_B^+} \left\{ \frac{S_B(x_A^+, x_B^+) \exp(\tau_t^2 f_A^- f_B^+)}{\exp(\tau_t^2 f_A^+ f_B^+)} \left( \frac{f_B^-(z^-)^\top}{N_A} - \frac{f_A^- f_B^+}{N_A^2 d} \right) \right\} \text{diag}(\sigma^2).
\]

(5)

See Lemma A.3 for the calculation. The first term comes from the positive pairs and the second and third terms from the negative pairs. Within each term, the second part comes from the normalization. The equations for the other \( K \)-matrices can be derived similarly. We rescale the gradients by \( 1/\tau_t^2 \) so that \( \frac{d}{dt} K_A \) does not shrink with \( \tau_t^2 \). For the non-contrastive case, the equation is

\[
\frac{d}{dt} K_A = \mathbb{E}_{x_A^+, x_B^+} \left\{ \frac{f_B^+(z^+)^\top}{N_A} - \frac{f_A^+ f_B^+}{N_A^2 d} \right\} \text{diag}(\sigma^2).
\]

(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 $\tau_t$ 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 $\tau^2$ (cf. Section 5.3 and Lemma B.9).

5.2 THE INFINITE-WIDTH DYNAMICS

The overall proof strategy is to first characterize the 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 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 $W_A$ and $W_B$. Hence, in the $m \to \infty$ limit, different columns of $K_A$ and $K_B$ 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 $A$ and $B$ are always the same. Namely, when $m \to \infty$, we have $K_A K_A^T = K_B K_B^T = \text{diag}(\kappa^2)$ and $K_A^T K_B = \text{diag}(\hat{\kappa}^2)$ for some $\kappa, \hat{\kappa} \in \mathbb{R}^r$. As a result, in order to characterize the dynamics of $K_A, K_B$, it suffices to look at $\kappa^2$ and $\hat{\kappa}^2$. One can show that, in this noiseless infinite-width limit, we have (cf. Lemma A.7)

$$\frac{d}{dt} \kappa^2 = 4 \left(1 - \tilde{S} \right) \left( \frac{\kappa^2}{\|\kappa\|^2} - \frac{\|\kappa\|^2}{\|\kappa\|^2} \right) \sigma_p^2 - 4 \left(1 - \tilde{S} \right) \left( \frac{\kappa^2}{\|\kappa\|^2} T_p - \frac{\|\kappa\|^2}{\|\kappa\|^2} \right) \sigma_p^2,$$

$$\frac{d}{dt} \hat{\kappa}^2 = 4 \left(1 - \tilde{S} \right) \left( \frac{\hat{\kappa}^2}{\|\kappa\|^2} - \frac{\|\kappa\|^2}{\|\kappa\|^2} \right) \sigma_p^2 - 4 \left(1 - \tilde{S} \right) \left( \frac{\hat{\kappa}^2}{\|\kappa\|^2} T_p - \frac{\|\kappa\|^2}{\|\kappa\|^2} \right) \sigma_p^2,$$

(7)
where \( \hat{S} \) is a \( \Theta(1) \) quantity depending on \( \hat{\kappa} \) and \( \kappa \), \( T_p = \tanh(\tau_p^2 \hat{\kappa}^2 / \|\kappa\|^2) \), and \( \hat{T} = \sum_{\kappa=1}^{\kappa}(\hat{\kappa}_p^2 / \|\kappa\|^2)T_p \). The first term comes from the first term of (5) and the second term from the second and third terms of (5). When \( \kappa_p^2 \approx \hat{\kappa}_p^2 \) for all \( p \in [r] \), the model is aligned. When all \( \kappa_p^2 \) are roughly the same, the model is balanced.

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. For ease of presentation, we will mostly use the infinite-width dynamics (7) instead of the finite-width one (5).

Equivalence of Stage 1 and non-contrastive methods

Recall that we use a small \( \tau_p^2 \) in Stage 1, \( T_p \approx 0 \) for all \( p \in [r] \), whence \( \hat{T} \) is also approximately 0. As a result, the second terms of (7) are approximately 0. In other words, only the positive pairs matter. Meanwhile, one can check that, in the infinite-width limit, (6) corresponds to the first term of (7), up to some multiplicative factor. This gives the equivalence of the dynamics of Stage 1 and the non-contrastive method. See Lemma B.9 for a more formal proof in the finite-width setting.

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

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

\[
\hat{\kappa}_p^2 / \kappa_p^2 \approx 1, \quad \| [K_A]_s \|_2^2 / \kappa_q^2 \approx 0, \quad \kappa_p^2 / \kappa_q^2 \leq O \left( \sqrt{d} \right) .
\]

Moreover, there exists a \( \sigma \in \mathbb{R}^r \) 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, \( K_A \approx K_B \) 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 \( \hat{X}_t \leq A_t X_t, \hat{Y}_t \leq \beta A_t X_t Y_t, \) with \( X_0, Y_0, \beta \) being positive. Then, for any \( T \geq 0 \), we have \( Y_T \leq Y_0 \exp(\beta X_0) \).

Here, \( X_t \) represents the progress we have made and \( Y_t \) the error we wish to control. In our case, \( X_t \) is the maximum between \( 1 - \kappa_p^2 / \kappa_p^2 \) and the noise-signal ratio, and \( Y_t \) the discretization error and condition number. This lemma says that, if the error growth rate decreases as we make more progress, then by coupling these two processes, we can make sure the error does not blow up.

5.4 Stage 2

In Stage 2, \( \tau_p^2 \) 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 \( K_A \) 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-width one throughout Stage 2, which again takes at most poly(\( d \)) iterations, and, at the end of Stage 2, we have, for any \( p, q \in [r] \) and \( s \in [d - r] \),

\[
\hat{\kappa}_p^2 / \kappa_p^2 \approx 1, \quad \| [K_A]_s \|_2^2 / \kappa_q^2 \approx 0, \quad \kappa_p^2 / \kappa_q^2 \approx 1 .
\]

The way to control \( \hat{\kappa}_p^2 / \kappa_p^2 \) and the noise-signal ratio is similar to Stage 1. For the condition number, note that when \( \tau_p^2 = 1 \) and \( \hat{\kappa} \approx \kappa \), the first term of (7) becomes close to 0 while the second term becomes \( -4(1 - \hat{S}) \kappa_p^2 / \|\kappa\|^2 (T_p - \hat{T}) \sigma_p^2 \). Since \( T_p = \tanh(\tau_p^2 \hat{\kappa}^2 / \|\kappa\|^2) \) is positively correlated with \( \hat{\kappa}_p^2 / \|\kappa\|^2 \) and \( \hat{T} \) is a weight average of these \( T_p \)'s, this term will push \( \hat{\kappa}_p^2 / \|\kappa\|^2 \) towards their average and, consequently, reduces the condition number. To obtain a satisfactory convergence rate, it suffices to consider the ratio \( \kappa_p^2 / \kappa_q^2 \) directly. See Appendix C.5 for details. The discretization error is handled in Appendix C.2.
6 Experimental results

![Figure 2: Results of the MSCOCO experiments. The top row plots the training loss, alignment scores, and balance scores during training, respectively. The bottom row figures plot the downstream accuracies and the largest 25 singular value of $\Sigma_f$ at different epochs, normalized so that the largest one has value 1. One can see that the alignment score quickly reaches near 100%, and the balance score, as well as the downstream accuracy, increases gradually during training, which matches our theoretical analysis.](image)

Besides the simulation results reported in Figure 1, we also conduct experiments on the MSCOCO-2014 dataset (Lin et al., 2014) using more practical models. See Figure 2 for the results. For the text part, we use a pre-trained RoBERTa (Liu et al., 2019), followed by a 3-layer fully-connected network with batch norm between layers. For the image part, we use a pre-trained ResNet101 (He et al., 2015), followed by the same layers. In both parts, the width of the fully-connected layers and the output dimension are 768. We freeze the pre-trained parts of the model and only train the fully-connected parts.

We measure the quality of the learned representation using its zero-shot performance on the MSCOCO-2014 validation set. Unlike common image classification datasets, images in the MSCOCO dataset usually have multiple labels, each corresponding to an object that appears in the image, and there are 80 categories in total. We regard a prediction to be correct if it matches one label. The zero-shot classification is done in the same way as in Radford et al. (2021). Namely, we compute the image embedding and the embeddings of prompts “This is a [LABEL\_NAME]”, and use the prompt with the highest correlation with the image embedding as the prediction.

7 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 include 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


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

The results in this subsection are mostly brute-force calculations. First, we prove the following auxiliary lemma.

**Lemma A.1.** For any $x_A^+$ and $x_B^+$, we have

$$
\nabla_{W_A} \exp(\tau_f x_A^+ \cdot f_B) = \frac{\tau^2_f}{N_A} \left( x_A^+(f_B)^\top - \langle f_A^+, f_B^+ \rangle \frac{A K_A^+ + A_\xi K_A^+}{N_A} \right).
$$

Then we compute the gradients.

**Lemma A.2.** We have

$$
\nabla_{W_A} \mathcal{L} = -\tau^2_f \mathbb{E}_{x_A^+, x_B^+} \left\{ \left(2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+)\right) \left( \frac{x_A^+(f_B)^\top}{N_A} - \langle f_A^+, f_B^+ \rangle \frac{A K_A^+ + A_\xi K_A^+}{N_A} \right) \right\}

+ \frac{K \tau^2_f}{N_A} \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau_f^2 f_A^+ \cdot f_B^+) \left( \frac{x_A^+(f_B)^\top}{N_A} - \langle f_A^+, f_B^+ \rangle \frac{A K_A^+ + A_\xi K_A^+}{\sqrt{d}} \right) \right\}

+ \frac{K \tau^2_f}{N_A} \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau_f^2 f_A^+ \cdot f_B^+) \left( \frac{x_A^+(f_B)^\top}{N_A} - \langle f_A^+, f_B^+ \rangle \frac{A K_A^+ + A_\xi K_A^+}{\sqrt{d}} \right) \right\}

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

Instead of tracking $W_A$ and $W_B$ directly, we are going to track $K_A$, $K_B$, $K_A_\xi$ and $K_B_\xi$. 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

$$
d \frac{d}{dt} K_A = \tau^2_f \mathbb{E}_{x_A^+, x_B^+} \left\{ \left(2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+)\right) \left( \frac{(K_B z^+ + K_B_\xi x_B^+)(z^+)^\top}{N_A N_B} - \langle f_A^+, f_B^+ \rangle \frac{K_A}{N_A d} \right) \right\} \text{diag}(\sigma^2)

- \tau^2_f \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau_f^2 f_A^+ \cdot f_B^+) \left( \frac{f_B(z^+)^\top}{N_A} - \langle f_A^+, f_B^+ \rangle \frac{K_A}{N_A d} \right) \right\} \text{diag}(\sigma^2)

- \tau^2_f \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau_f^2 f_A^+ \cdot f_B^+) \left( \frac{f_B(z^+)^\top}{N_A} - \langle f_A^+, f_B^+ \rangle \frac{K_A}{N_A d} \right) \right\} \text{diag}(\sigma^2),

and

$$
d \frac{d}{dt} K_A_\xi = \tau^2_f N_A \mathbb{E}_{x_A^+, x_B^+} \left\{ \left(2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+)\right) \left( f_B^+(\xi_A)^\top - \frac{\langle f_A^+, f_B^+ \rangle K_A_\xi}{\sqrt{d}} \right) \right\} \sigma^2_\xi

- \tau^2_f N_A \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau_f^2 f_A^+ \cdot f_B^+) \left( f_B^+(\xi_A)^\top - \frac{\langle f_A^+, f_B^+ \rangle K_A_\xi}{\sqrt{d}} \right) \right\} \sigma^2_\xi

- \tau^2_f N_A \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau_f^2 f_A^+ \cdot f_B^+) \left( f_B^+(\xi_A)^\top - \frac{\langle f_A^+, f_B^+ \rangle K_A_\xi}{\sqrt{d}} \right) \right\} \sigma^2_\xi.
We can rewrite the above result as follows.

**Corollary A.4.** Define

\[
Q_0 := - \mathbb{E}_{x_A^+, x_B^+} \left\{ (2 - S_A(x_A^+, x_B^+)) \langle f_A^+, f_B^+ \rangle \right\} + K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\} + K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\},
\]

and

\[
Q_1 := \mathbb{E}_{x_A^+, x_B^+} \left\{ (2 - S_A(x_A^+, x_B^+)) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\},
\]

\[
Q_{1, \xi_B} := \mathbb{E}_{x_A^+, x_B^+} \left\{ (2 - S_A(x_A^+, x_B^+)) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\},
\]

\[
Q_{1, \xi_A} := \mathbb{E}_{x_A^+, x_B^+} \left\{ (2 - S_A(x_A^+, x_B^+)) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\},
\]

\[
Q_2 := \mathbb{E}_{x_A^+, x_B^+} \left\{ (2 - S_A(x_A^+, x_B^+)) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_A(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\} - K \mathbb{E}_{x_A^+, x_B^+} \left\{ S_B(x_A^+, x_B^+) \exp(\tau^2 f_A^+ \cdot f_B^+) \langle f_A^+, f_B^+ \rangle \right\}.
\]

We have

\[
\frac{d}{dt} K_A = \frac{K_B}{N_A N_B d} Q_{1, \xi_B} \text{diag}(\sigma^2) + \frac{K_B \xi}{N_A N_B d} Q_{1, \xi_B} \text{diag}(\sigma^2) + \frac{K_A}{N_A^2 d} Q_0 \text{diag}(\sigma^2),
\]

\[
\frac{d}{dt} K_B = \frac{K_A}{N_A N_B d} Q_{1, \xi_A} \text{diag}(\sigma^2) + \frac{K_A \xi}{N_A N_B d} Q_{1, \xi_A} \text{diag}(\sigma^2) + \frac{K_B}{N_B^2 d} Q_0 \text{diag}(\sigma^2),
\]

\[
\frac{d}{dt} K_{A, \xi} = \frac{K_B}{N_A N_B d} Q_{1, \xi_B} \sigma^2 + \frac{K_B \xi}{N_A N_B d} Q_{2, \xi_B} \sigma^2 + \frac{K_A \xi}{N_A^2 d} Q_0 \sigma^2,
\]

\[
\frac{d}{dt} K_{B, \xi} = \frac{K_A}{N_A N_B d} Q_{1, \xi_A} \sigma^2 + \frac{K_A \xi}{N_A N_B d} Q_{2, \xi_A} \sigma^2 + \frac{K_B \xi}{N_B^2 d} Q_0 \sigma^2.
\]

We are interested in each column of $K_A$ and $K_B$, 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

\[
\frac{d}{dt} \|K_A\|_p^2 = 2 \left( \|K_A\|_p, \|K_BQ_1\|_p \right) \sigma_p^2 + 2 \left( \|K_A\|_p, \|K_BQ_1\|_p \right) \sigma_p^2 + 2 \left( \|K_A\|_p, \|K_BQ_1\|_p \right) \sigma_p^2 + 2 \left( \|K_B\|_p, \|K_AQ_1\|_p \right) \sigma_p^2 + 2 \left( \|K_B\|_p, \|K_AQ_1\|_p \right) \sigma_p^2 + 2 \left( \|K_B\|_p, \|K_AQ_1\|_p \right) \sigma_p^2 + 2 \left( \|K_B\|_p, \|K_AQ_1\|_p \right) \sigma_p^2,
\]

and

\[
\frac{d}{dt} \|K_A\|_q^2 = 2 \left( \|K_A\|_q, \|K_BQ_1\|_q \right) \sigma_q^2 + 2 \left( \|K_A\|_q, \|K_BQ_1\|_q \right) \sigma_q^2 + 2 \left( \|K_A\|_q, \|K_BQ_1\|_q \right) \sigma_q^2 + 2 \left( \|K_B\|_q, \|K_AQ_1\|_q \right) \sigma_q^2 + 2 \left( \|K_B\|_q, \|K_AQ_1\|_q \right) \sigma_q^2 + 2 \left( \|K_B\|_q, \|K_AQ_1\|_q \right) \sigma_q^2 + 2 \left( \|K_B\|_q, \|K_AQ_1\|_q \right) \sigma_q^2.
\]

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

\[
\nabla_{W_A} \hat{L} = \text{E} \left\{ \frac{\sigma^2}{N_A} (f_B^+)^T \right\} \frac{\left( AK_T + A\xi K_T A\xi \right)}{N_A^2 d}
\]

As a corollary, we have, in this case,

\[
\frac{d}{dt} K_A = \text{E} \left\{ \frac{f_B^+ (\xi_a^+)^T}{N_A} - \langle f_A^+, f_B^+ \rangle \frac{K_A}{N_A^2 d} \right\} \text{diag}(\sigma^2),
\]

\[
\frac{d}{dt} K_A \cdot \xi = \text{E} \left\{ \frac{f_B^+ (\xi_a^+)^T}{N_A} - \langle f_A^+, f_B^+ \rangle \frac{K_A \xi}{N_A^2 d} \right\} \sigma^2.
\]

Omitted proof of this subsection

Proof of Lemma A.1. We compute

\[
\nabla_{W_A} \exp(\tau_i^2 f_A \cdot f_B) = \tau_i^2 \exp(\tau_i^2 f_A \cdot f_B) \nabla_{W_A} \langle f_A^+, f_B^- \rangle
\]

\[
= \tau_i^2 \exp(\tau_i^2 f_A \cdot f_B) \nabla_{W_A} \frac{\langle W^T_A \sigma^+ A, f_B^- \rangle}{\sqrt{\|K_A\|^F + \|K_A\|_F}}
\]

\[
= \tau_i^2 \exp(\tau_i^2 f_A \cdot f_B) \nabla_{W_A} \left( \langle W^T_A \sigma^+ A, f_B^- \rangle \right) \sqrt{\|K_A\|^F + \|K_A\|_F} / \sqrt{d}
\]

\[
- \tau_i^2 \exp(\tau_i^2 f_A \cdot f_B) \langle f_A^+, f_B^- \rangle \nabla_{W_A} \sqrt{\|K_A\|^F + \|K_A\|_F} \frac{\langle f_A^+, f_B^- \rangle \sqrt{\|K_A\|^F + \|K_A\|_F}}{\sqrt{\|K_A\|^F + \|K_A\|_F}}.
\]
For the first term, we have $\nabla W_A \langle W_A^T x_A^+, f_B \rangle = x_A^+(f_B\uparrow)$. For the second term, we have

$$\nabla W_A \frac{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2} = \nabla W_A \|K_A\|_F^2 + \nabla W_A \|K_{A,\xi}\|_F^2 = \frac{AK_A^\top + A_{\xi}K_{A,\xi}^\top}{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}.$$

Hence,

$$\nabla W_A \exp(\tau_A^2 f_A^+ \cdot f_B) = \tau_A^2 \exp(\tau_A^2 f_A^+ \cdot f_B) \frac{x_A^+(f_B\uparrow)}{\sqrt{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}} - \tau_A^2 \exp(\tau_A^2 f_A^+ \cdot f_B) \langle f_A^+, f_B \rangle \frac{AK_A^\top + A_{\xi}K_{A,\xi}^\top}{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}$$

$$= \frac{\tau_A^2 \exp(\tau_A^2 f_A^+ \cdot f_B)}{\sqrt{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}} \left( x_A^+(f_B\uparrow) - \langle f_A^+, f_B \rangle \frac{AK_A^\top + A_{\xi}K_{A,\xi}^\top}{\sqrt{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}} \right).$$

Proof of Lemma A.2. We compute

$$\nabla W_A \mathcal{L}_A = -\mathbb{E} \frac{\nabla W_A S_A(x_A^+, x_B^+)}{S_A(x_A^+, x_B^+)}$$

$$= -\mathbb{E} \left\{ \frac{1}{S_A(x_A^+, x_B^+)} \nabla W_A \exp(\tau_A^2 f_A^+ \cdot f_B) \right\}$$

$$= -\mathbb{E} \left\{ \frac{1}{S_A(x_A^+, x_B^+)} \nabla W_A \exp(\tau_A^2 f_A^+ \cdot f_B) \right\}$$

$$+ \mathbb{E} \left\{ \frac{\nabla W_A \exp(\tau_A^2 f_A^+ \cdot f_B)}{\exp(\tau_A^2 f_A^+ \cdot f_B)} \frac{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}{\sqrt{\|K_A\|_F^2 + \|K_{A,\xi}\|_F^2}} \right\}.$$
Then, for $\nabla W_A\mathcal{L}_B$, we compute
\[
\nabla W_A\mathcal{L}_B = -\mathbb{E}\frac{\nabla W_A S_B(x_A^+, x_B^+)}{S_B(x_A^+, x_B^+)}
\]
\[
= -\mathbb{E}\left\{ \frac{1}{S_B(x_A^+, x_B^+)} \nabla W_A \exp(\tau_t^2 f_A^+ \cdot f_B^+) \right\}
\]
\[
+ \mathbb{E}\left\{ \nabla W_A \exp(\tau_t^2 f_A^+ \cdot f_B^+) + K \nabla W_A \mathbb{E}_{x_A} \exp(\tau_t^2 f_A^+ \cdot f_B^+) \right\}.
\]
Again, by Lemma A.1, the first term is
\[
-\mathbb{E}\left\{ \frac{1}{S_B(x_A^+, x_B^+)} \nabla W_A \exp(\tau_t^2 f_A^+ \cdot f_B^+) \right\}
\]
\[
= -\frac{\tau_t^2}{N_A} \mathbb{E}\left\{ x_A^+(f_B^+)^\top - \langle f_A^+, f_B^+ \rangle \frac{(AK_A^+ + A\xi K_A^+\xi) / \sqrt{d}}{\|K_A\|^2_F + \|K_A\xi\|^2_F} \right\},
\]
and the second term is
\[
\mathbb{E}\left\{ \frac{\nabla W_A \exp(\tau_t^2 f_A^+ \cdot f_B^+) + K \nabla W_A \mathbb{E}_{x_A} \exp(\tau_t^2 f_A^+ \cdot f_B^+)}{\exp(\tau_t^2 f_A^+ \cdot f_B^+) + K \mathbb{E}_{x_A} \exp(\tau_t^2 f_A^+ \cdot f_B^+)} \right\}
\]
\[
= \frac{\tau_t^2}{N_A} \mathbb{E}\left\{ S_B(x_A^+, x_B^+) \left( x_A^+(f_B^+)^\top - \langle f_A^+, f_B^+ \rangle \frac{(AK_A^+ + A\xi K_A^+\xi) / \sqrt{d}}{\|K_A\|^2_F + \|K_A\xi\|^2_F} \right) \right\}
\]
\[
+ \frac{K \tau_t^2}{N_A} \mathbb{E}\left\{ S_B(x_A^+, x_B^+) \exp(\tau_t^2 f_A^+ \cdot f_B^+) \left( x_A^-(f_B^+)^\top - \langle f_A^-, f_B^+ \rangle \frac{(AK_A^+ + A\xi K_A^+\xi) / \sqrt{d}}{\|K_A\|^2_F + \|K_A\xi\|^2_F} \right) \right\}.
\]
Thus,
\[
\nabla W_A\mathcal{L}_B = -\frac{\tau_t^2}{N_A} \mathbb{E}\left\{ (1 - S_B(x_A^+, x_B^+)) \left( x_A^+(f_B^+)^\top - \langle f_A^+, f_B^+ \rangle \frac{(AK_A^+ + A\xi K_A^+\xi) / \sqrt{d}}{\|K_A\|^2_F + \|K_A\xi\|^2_F} \right) \right\}
\]
\[
+ \frac{K \tau_t^2}{N_A} \mathbb{E}\left\{ S_B(x_A^+, x_B^+) \exp(\tau_t^2 f_A^+ \cdot f_B^+) \left( x_A^-(f_B^+)^\top - \langle f_A^-, f_B^+ \rangle \frac{(AK_A^+ + A\xi K_A^+\xi) / \sqrt{d}}{\|K_A\|^2_F + \|K_A\xi\|^2_F} \right) \right\}.
\]
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 $\kappa_p^2 = \|[K_A]_p\|^2$ and $\hat{\kappa}_p^2 = \langle [K_A]_p, [K_B]_p \rangle$.

Lemma A.7. In the noiseless infinite-width case, we have
\[
\frac{d}{dt} \kappa_p^2 = 4 \left( 1 - \hat{S} \right) \left( \frac{\kappa_p^2}{\|\kappa\|^2} - \frac{\hat{\kappa}_p^2}{\|\kappa\|^2} \right) \sigma_p^2 + 4 \left( 1 - \hat{S} \right) \left( \frac{\kappa_p^2}{\|\kappa\|^2} - \frac{\hat{\kappa}_p^2}{\|\kappa\|^2} \right) \sigma_p^2,
\]
\[
\frac{d}{dt} \hat{\kappa}_p^2 = 4 \left( 1 - \hat{S} \right) \left( \frac{\hat{\kappa}_p^2}{\|\kappa\|^2} - \frac{\hat{\kappa}_p^2}{\|\kappa\|^2} \right) \sigma_p^2 + 4 \left( 1 - \hat{S} \right) \left( \frac{\hat{\kappa}_p^2}{\|\kappa\|^2} - \frac{\hat{\kappa}_p^2}{\|\kappa\|^2} \right) \sigma_p^2.
\]
Proof. First, note that
\[
\langle f^+_A, f^-_B \rangle = \frac{\langle K_{BA} z^+, z^- \rangle}{\|\kappa\|^2 / d} = \frac{\langle z^+, z^- \rangle}{\|\kappa\|^2 / d}.
\]
This implies that (a) \( \langle f^+_A, f^-_B \rangle \) does not depend on the actual value of \( z^+ \), and (b) if we flip the signs of \( z^+_p \) simultaneously, then the value of \( \langle f^+_A, f^-_B \rangle \) remain unchanged. To compute \( S_A \), we then need to take expectation over \( z^- \). We compute
\[
\mathbb{E}_{z^-} \exp \left( \tau_1^2 f^+_A \cdot f^-_B \right) = \prod_{k=1}^r \mathbb{E}_{z^-} \exp \left( \tau_1^2 \frac{\kappa^2 z^+_k z^-_k}{\|\kappa\|^2 / d} \right) = \prod_{k=1}^r \cosh \left( \frac{\tau_1^2 \kappa^2_k}{\|\kappa\|^2} \right) =: Z_c.
\]
Again, it does not depend on the actual value of \( z^+ \). One can conduct similar calculation for \( S_B \) and, consequently, we have \( S_A \equiv S_B \equiv \hat{S} \) for some \( \hat{S} \) that depends on \( \kappa \) and \( \kappa \) but not on \( z^+ \).

Then, we can rewrite (5) as
\[
\frac{d}{dt} K_A = 2 \left( 1 - \hat{S} \right) \mathbb{E}_{x_A, x_B} \left\{ f^+_B (z^+)^\top \frac{\hat{K}}{N_A} - \frac{\langle f^+_A, f^-_B \rangle K_A}{N_A d} \right\} \text{diag}(\sigma^2)
\]
\[
- 2 K \hat{S} \mathbb{E}_{x_A, x_B} \left\{ \frac{\langle f^+_A, f^-_B \rangle K_A}{N_A d} - \frac{\langle f^+_A, f^-_B \rangle K_A}{N_A d} \right\} \text{diag}(\sigma^2)
\]
\[
= 2 \left( 1 - \hat{S} \right) \left( \frac{K_B}{\|\kappa\|^2} - \frac{\|\kappa\|^2 K_A}{\|\kappa\|^2} \right) \text{diag}(\sigma^2)
\]
\[
- 2 K \hat{S} \mathbb{E}_{x_A, x_B} \left\{ \exp \left( \frac{\tau_1^2 (z^+, z^-) \|\kappa\|^2}{\|\kappa\|^2 / d} \right) \left( \frac{K_B (z^+)^\top}{\|\kappa\|^2 / d} - \frac{\langle z^+, z^- \rangle}{\|\kappa\|^2 / d} \right) \right\} \text{diag}(\sigma^2).
\]

Note that
\[
\mathbb{E}_{z^-} \left\{ \exp \left( \frac{\tau_1^2 (z^+, z^-) \|\kappa\|^2}{\|\kappa\|^2 / d} \right) z^+_p z^-_q d \right\}
\]
\[
= \mathbb{I}_{\{p = q\}} \mathbb{E}_{z^-} \left\{ \exp \left( \frac{\tau_1^2 \kappa^2_k z^+_k z^-_k}{\|\kappa\|^2} \right) z^+_p z^-_p d \right\} \prod_{k \neq p} \mathbb{E}_{z^-} \left\{ \exp \left( \frac{\tau_1^2 \kappa^2_k z^+_k z^-_k}{\|\kappa\|^2} \right) \right\}
\]
\[
= \mathbb{I}_{\{p = q\}} \sinh \left( \frac{\tau_1^2 \kappa^2_k}{\|\kappa\|^2} \right) \prod_{k \neq p} \cosh \left( \frac{\tau_1^2 \kappa^2_k}{\|\kappa\|^2} \right)
\]
\[
= \mathbb{I}_{\{p = q\}} Z_c T_p.
\]

Meanwhile, note that
\[
\mathbb{E}_{x_A, x_B} \left\{ \exp \left( \frac{\tau_1^2 (z^+, z^-) \|\kappa\|^2}{\|\kappa\|^2 / d} \right) \frac{\langle z^+, z^- \rangle}{\|\kappa\|^2 / d} \right\} = \sum_{k=1}^r \frac{\kappa^2_k}{\|\kappa\|^2 / d} \mathbb{E}_{z^-} \left\{ \exp \left( \frac{\tau_1^2 \kappa^2_k z^+_k z^-_k}{\|\kappa\|^2} \right) z^+_k z^-_k \right\}
\]
\[
= Z_c \sum_{k=1}^r \frac{\kappa^2_k}{\|\kappa\|^2} T_k
\]
\[
=: Z_c \hat{T}.
\]
Thus,
\[
\frac{d}{dt} K_A = 2 \left(1 - \tilde{S}\right) \left( \frac{K_B}{||\kappa||^2} - \frac{||\hat{\kappa}||^2}{||\kappa||^2} \frac{K_A}{||\kappa||^2} \right) \text{diag}(\sigma^2)
\]
\[
- \frac{2K \tilde{S} Z_c}{\exp\left(\frac{2\gamma^2||\kappa||^2}{||\kappa||^2}\right)} \left( \frac{K_B}{||\kappa||^2} \text{diag} \left( [T_k]_{k \in [r]} \right) - \frac{K_A}{||\kappa||^2} \tilde{T} \right) \text{diag}(\sigma^2)
\]
\[
= 2 \left(1 - \tilde{S}\right) \left( \frac{K_B}{||\kappa||^2} - \frac{||\hat{\kappa}||^2}{||\kappa||^2} \frac{K_A}{||\kappa||^2} \right) \text{diag}(\sigma^2)
\]
\[
- 2 \left(1 - \tilde{S}\right) \left( \frac{K_B}{||\kappa||^2} \text{diag} \left( [T_k]_{k \in [r]} \right) - \frac{K_A}{||\kappa||^2} \tilde{T} \right) \text{diag}(\sigma^2).
\]

As a corollary, we have
\[
\frac{d}{dt} [K_A]_p = 2 \left(1 - \tilde{S}\right) \left( \frac{[K_B]_p}{||\kappa||^2} - \frac{||\hat{\kappa}||^2}{||\kappa||^2} \frac{[K_A]_p}{||\kappa||^2} \right) \sigma_p^2
\]
\[
- 2 \left(1 - \tilde{S}\right) \left( \frac{[K_B]_p}{||\kappa||^2} \text{diag} \left( [T_k]_{k \in [r]} \right) - \frac{[K_A]_p}{||\kappa||^2} \tilde{T} \right) \sigma_p^2.
\]

Hence,
\[
\frac{d}{dt} \kappa_p^2 = 2 \left( [K_A]_p, \frac{d}{dt} [K_A]_p \right)
\]
\[
= 4 \left(1 - \tilde{S}\right) \left( \frac{\hat{\kappa}_p^2}{||\kappa||^2} - \frac{||\hat{\kappa}||^2}{||\kappa||^2} \frac{\kappa_p^2}{||\kappa||^2} \right) \sigma_p^2
\]
\[
- 2 \left(1 - \tilde{S}\right) \left( \frac{\hat{\kappa}_p^2}{||\kappa||^2} \text{diag} \left( [T_k]_{k \in [r]} \right) - \frac{\kappa_p^2}{||\kappa||^2} \tilde{T} \right) \sigma_p^2.
\]

By symmetry, for \( K_B \), we have
\[
\frac{d}{dt} [K_B]_p = 2 \left(1 - \tilde{S}\right) \left( \frac{[K_A]_p}{||\kappa||^2} - \frac{||\hat{\kappa}||^2}{||\kappa||^2} \frac{[K_B]_p}{||\kappa||^2} \right) \sigma_p^2
\]
\[
- 2 \left(1 - \tilde{S}\right) \left( \frac{[K_A]_p}{||\kappa||^2} \text{diag} \left( [T_k]_{k \in [r]} \right) - \frac{[K_B]_p}{||\kappa||^2} \tilde{T} \right) \sigma_p^2.
\]

Then, we can compute
\[
\frac{d}{dt} \hat{\kappa}_p^2 = 4 \left(1 - \tilde{S}\right) \left( \frac{\kappa_p^2}{||\kappa||^2} - \frac{||\hat{\kappa}||^2}{||\kappa||^2} \frac{\hat{\kappa}_p^2}{||\kappa||^2} \right) \sigma_p^2
\]
\[
- 4 \left(1 - \tilde{S}\right) \left( \frac{\kappa_p^2}{||\kappa||^2} \text{diag} \left( [T_k]_{k \in [r]} \right) - \frac{\hat{\kappa}_p^2}{||\kappa||^2} \tilde{T} \right) \sigma_p^2.
\]

\[\Box\]

## B \text{ STAGE 1}

In this section, we show that the following hold:

(a) \( K_A \approx K_B \) 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 \( \tau_k^2 \) that may depend on the \( \delta \)'s that appear in this lemma so that the following statement holds.

Let \( T_1 \) be the earliest time all the following hold:

\[
\left\langle [K_A]_p, [K_B]_p \right\rangle \geq 1 - \delta_-, \quad \forall p \in [r],
\]
\[
\frac{||[K_A]_q||}{||[K_A]_p||} \leq \delta_{K/S}, \quad \forall p \in [r], q \in [d-r],
\]

\[
\frac{||[K_A]_q||}{||[K_A]_p||} \geq 1 - \delta_+, \quad \forall p \in [r], q \in [d-r],
\]

\[
\frac{||K_A||}{||[K_A]_p||} \leq \delta_{K/S}, \quad \forall p \in [r],
\]

\[
\frac{||K_A||}{||[K_A]_p||} \geq 1 - \delta_+, \quad \forall p \in [r],
\]
where $\delta_{\bot}, \delta_{N/S} \in 1/\text{poly}(d)$ are two given parameters. We have $T_1 \leq \text{poly}(d)$. Moreover, at any time $t \in [0, T_1]$, we have $\kappa_0 := \max_{p,q \in [r]} \|K_A[p]\| / \|K_A[q]\| \leq O(\sqrt{d})$ and

$$
\|K_A[p]\|^2 = (1 \pm \delta_{A/B}) \|K_B[p]\|^2, \quad \|K_A[q]\|^2 = (1 \pm \delta_{A/B}) \|K_B[q]\|^2, \quad \forall p \in [r], q \in [d-r],
$$

$$
1 - \frac{\|K_A[p]\|}{\|K_A[q]\|} \leq \delta_{\pm}, \quad \forall p, q \in [d-r],
$$

$$
\max \left\{ \frac{\langle K_C[p], K_D[q] \rangle}{\langle K_C[p], K_D[q] \rangle} : C, D \in \{A, B\} \right\} \leq \delta_{AB}, \quad \forall p \neq q \in [r],
$$

$$
\max \left\{ \frac{\langle K_C[p], K_D[q] \rangle}{\langle K_C[p], K_D[q] \rangle} : C, D \in \{A, B\} \right\} \leq \delta_{\pm}, \quad \forall p \in [r], q, s \in [d-r],
$$

where $\delta_{A/B}, \delta_{\pm}, \delta_{AB, \pm}, \delta_{\pm, \bot} \in 1/\text{poly}(d)$ 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

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

For notational simplicity, we also define

$$
\rho_\pm = \max_{p \in [r]} \left(1 - \frac{\langle K_A[p], K_B[p] \rangle}{\langle K_A[p], K_B[p] \rangle} \right), \quad \rho_{N/S} = \max_{p \in [r]} \frac{\|K_A[q]\|}{\|K_A[p]\|}.
$$

Characterizing the dynamics in Stage 1 is relatively straightforward. We will see in Section B.1 that all those $Q$-matrices have simple forms. 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

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

with $X_0, Y_0, \alpha$ being positive. Then, for any $T \geq 0$, we have $Y_T \leq Y_0 \exp(\alpha X_0)$.

**Remark.** Here, $X_t$ represent the progress we have made and $Y_t$ the error. In our case, $X_t$ is the maximum between $1 - \frac{\langle K_A[p], K_B[p] \rangle}{\langle K_A[p], K_B[p] \rangle}$ 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

$$
X_T = X_0 \exp \left(- \int_0^T A_t \, dt\right), \quad Y_T = Y_0 \exp \left(\alpha X_0 \int_0^T A_t \exp \left(- \int_0^t A_s \, ds\right) \, dt\right).
$$

Note that

$$
\int_0^T A_t \exp \left(- \int_0^t A_s \, ds\right) \, dt = - \int_0^T d \exp \left(- \int_0^t A_s \, ds\right) = 1 - \exp \left(- \int_0^T A_t \, dt\right).
$$

Hence,

$$
Y_T = Y_0 \exp \left(\alpha X_0 \left(1 - \exp \left(- \int_0^T A_t \, dt\right)\right)\right) \leq Y_0 \exp \left(\alpha X_0\right).
$$

$\square$
The organization of this section is as follows. In Section B.1, we derive estimations for the $Q$-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 $1 - \langle [K_A]_{1p}, [K_B]_{1p} \rangle$ 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 $Q$-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, $\tau_i^2$ is small. Hence, $\exp(\tau_i^2 \langle f_A, f_B \rangle) = 1 \pm O(\tau_i^2)$. With this approximation, one can derive the following estimation for $S_A$ and $S_B$.

**Lemma B.3** (Estimations for $S$). In Stage 1, we have
\[ S_A = \frac{1}{1 + K} \pm O_z(\tau_i^2) \pm O(\tau_i^2 \delta_{\xi, \perp} \rho_{N/S}), \]
and the same is also true for $S_B$. Here, $O_z(\tau_i^2)$ means $O(\tau_i^2)$ and this does not depend on $\xi_A$ nor $\xi_B$.

The proof of this lemma and all following lemma is deferred to the end of this subsection. Note that we derive a slightly finer estimation for the noise-related part. This additional $\rho_{N/S}$ will be used cancel with terms like $\rho_{N/S}$ that we derive a slightly finer estimation for the noise-related part. This additional $\rho_{N/S}$ will be used to simplify the equations governing the training dynamics. In Section B.2, we estimate the rate at which $d\tau_i^2$ and $\xi$ 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.

**Lemma A.5.** Recall that, in Stage 1, $\tau_i^2$ is small. Hence, $\exp(\tau_i^2 \langle f_A, f_B \rangle) = 1 \pm O(\tau_i^2)$. With this approximation, one can derive the following estimation for $S_A$ and $S_B$.

**Lemma B.3** (Estimations for $S$). In Stage 1, we have
\[ S_A = \frac{1}{1 + K} \pm O_z(\tau_i^2) \pm O(\tau_i^2 \delta_{\xi, \perp} \rho_{N/S}), \]
and the same is also true for $S_B$. Here, $O_z(\tau_i^2)$ means $O(\tau_i^2)$ and this does not depend on $\xi_A$ nor $\xi_B$.

The proof of this lemma and all following lemma is deferred to the end of this subsection. Note that we derive a slightly finer estimation for the noise-related part. This additional $\rho_{N/S}$ will be used cancel with terms like $\rho_{N/S}$ that we derive a slightly finer estimation for the noise-related part. This additional $\rho_{N/S}$ will be used to simplify the equations governing the training dynamics. In Section B.2, we estimate the rate at which $d\tau_i^2$ and $\xi$ 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.

**Lemma B.4** (Estimations for $Q_1$). In Stage 1, we have
\[ Q_1 = \frac{2K}{1 + K} I_d \pm O(\tau_i^2). \]

**Lemma B.5** (Estimations for $Q_{1, \xi}$ and $Q_2$). In Stage 1, we have
\[ \max\{\|Q_{1, \xi_A}\|_{F}, \|Q_{1, \xi_B}\|_{F}, \|Q_2\|_{F}\} = \pm O(\tau_i^2 \delta_{\xi, \perp} \rho_{N/S}). \]

The above two lemmas say that in Stage 1, $Q_1$ is approximately diagonal, and $Q_{1, \xi}$ and $Q_2$ can more or less be ignored.

**Lemma B.6** (Estimations for $Q_0$). In Stage 1, we have
\[ Q_0 = -\frac{2K}{1 + K} \langle K_A, K_B \rangle_{N_A N_B d} \pm O(\tau_i^2). \]

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
\[
\frac{d}{dt} \| [K_A]_{1p} \|^2 = \frac{4K}{1 + K} \frac{\sigma_p^2}{N_A N_B d} \left( \| [K_A]_{1p} \| \cdot \langle [K_A]_{1p}, [K_B]_{1p} \rangle - \frac{\langle K_A, K_B \rangle}{N_A N_B d} \right) \| [K_A]_{1p} \|^2
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \left( \sqrt{\delta_{A/B} + \tau_i^2 d} \right) \kappa_p \right),
\]
\[
\frac{d}{dt} \| [K_{A, \xi}]_{1q} \|^2 = -\frac{4K}{1 + K} \frac{\sigma_q^2}{N_A N_B d} \frac{\langle K_A, K_B \rangle}{N_A N_B d} \| [K_{A, \xi}]_{1q} \|^2 \pm O \left( \frac{\sigma_q^2}{N_A N_B d} \left( \tau_i^2 + \sqrt{\delta_{A/B}} \right) \kappa_q \right),
\]

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

21
Basically, the two parts of the first term of $\frac{d}{dt} \| [K_{A}]_{p} \|^2$ correspond to the signal and normalization, respectively. For $\frac{d}{dt} \| [K_{A,\xi}]_{q} \|^2$, the signal term is 0 because the noises in $x^+_{A}$ and $x^+_{B}$ are independent and have mean 0.

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

$$
\frac{d}{dt} [K_{A}]_{p} = \frac{2K}{1+K} \frac{\sigma_{p}^{2}}{N_{A}N_{B}d} \left( I - [K_{A}]_{p} [([K_{A}]_{p})^\top] \right) [K_{B}]_{p} + O \left( \frac{\sigma_{p}^{2}}{N_{A}N_{B}d} \tau_{\xi}^{2}d^{2}K_{0} + d\sqrt{\delta_{A/B}} \right),
$$

$$
\frac{d}{dt} [K_{A,\xi}]_{q} = \pm O \left( \frac{\sigma_{\xi}^{2}}{N_{A}N_{B}d} \tau_{\xi}^{2}d^{2}K_{0}\delta_{\xi,\perp} \right).
$$

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

Note that, after normalization, the normalization terms cancel with each other, and the signal terms drive $[K_{A}]_{p}$ and $[K_{A,\xi}]_{q}$ toward each other. Moreover, without the normalization terms, different $[K_{A}]_{p}$ are approximately independent of each other. This allows us to maintain the orthogonality between different columns.

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

$$
\langle f_{A}^{+} \cdot f_{B}^{+} \rangle = \frac{\langle [K_{A,z}^{+}, K_{B,z}^-] \rangle}{N_{A}N_{B}d} + \frac{\langle [K_{A,\xi}^{+}, K_{B,\xi}^-] \rangle}{N_{A}N_{B}d} + \frac{\langle [K_{A,\xi}^{+}, K_{B,\xi}^-] \rangle}{N_{A}N_{B}d}.
$$

For the second term, we write

$$
\frac{\langle [K_{A,z}^{+}, K_{B,\xi}^{-}] \rangle}{N_{A}N_{B}} = \sum_{i \in [r], j \in [d-1]} [K_{A}]_{i} [K_{B,\xi}]_{j} [K_{A}]_{i} [K_{B,\xi}]_{j} \langle [K_{A}]_{i}, [K_{B,\xi}]_{j} \rangle z_{i}^{+} \xi_{j}^{-} d.
$$

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

$$
\langle f_{A}^{+} \cdot f_{B}^{+} \rangle = \frac{\langle [K_{A,z}^{+}, K_{B,z}^-] \rangle}{N_{A}N_{B}} + O \left( d\delta_{\xi,\perp} \rho_{N/S} \right).
$$

Then, we compute

$$
\exp \left( \tau_{\xi}^{2} \langle f_{A}^{+} \cdot f_{B}^{+} \rangle \right) = \exp \left( \tau_{\xi}^{2} \langle [K_{A,z}^{+}, K_{B,z}^-] \rangle \frac{N_{A}N_{B}}{d} \right) \left( 1 + O \left( \tau_{\xi}^{2}d^{2}\delta_{\xi,\perp} \rho_{N/S} \right) \right)
$$

$$
= 1 + O_{2} \left( \tau_{\xi}^{2}d^{2}\delta_{\xi,\perp} \rho_{N/S} \right).
$$

Thus,

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

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

**Proof of Lemma B.4.** Recall that

$$
Q_{1} := \mathbb{E} \left\{ ( 2 - S_{A}(x_{A}^{+}, x_{B}^{+}) - S_{B}(x_{A}^{+}, x_{B}^{+}) ) z^{+} (z^{+})^\top d \right\}
$$

$$
- K \mathbb{E} \left\{ S_{A}(x_{A}^{+}, x_{B}^{+}) \exp(\tau_{i}^{2} f_{A}^{+} \cdot f_{B}^{+} ) z^{+} (z^{+})^\top d \right\}
$$

$$
- K \mathbb{E} \left\{ S_{B}(x_{A}^{+}, x_{B}^{+}) \exp(\tau_{i}^{2} f_{A}^{+} \cdot f_{B}^{+} ) z^{+} (z^{+})^\top d \right\}. \tag{22}
$$
Hence, we now estimate these terms one-by-one. For \( T \),

\[
2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+) = 2 - \frac{2}{1 + K} \pm O(\tau^2_t) = 2K \pm O(\tau^2_t).
\]

As a result,

\[
Q_1 = \frac{2K}{1 + K} E \{ z^+(z^+)\T d \} - \frac{K}{1 + K} E \{ z^-(z^-)\T d \} - \frac{K}{1 + K} E \{ z^+(z^-)\T d \} \pm O(d\tau^2_t).
\]

Note that \( E \{ z^-(z^+)\T \} = 0 \) and \( E \{ z^+(z^+)\T \} = I_d/d. \) Therefore,

\[
Q_1 = \frac{2K}{1 + K} I_d \pm O(d\tau^2_t).
\]

**Proof of Lemma B.5.** Recall that

\[
Q_{1,\xi} := E \{ (2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+)) \xi_A^+(z^+)\T d \}
\]

\[
- K E \left\{ \frac{S_A(x_A^+, x_B^+)}{\exp(\tau^2_t f_A \cdot f_B)} \xi_A^+(z^-)\T d \right\}
\]

\[
- K E \left\{ \frac{S_B(x_A^+, x_B^+)}{\exp(\tau^2_t f_A \cdot f_B)} \xi_A^+(z^+)\T d \right\}.
\]

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

\[
Q_{1,\xi} = \pm O(\tau^2_t d^2 \rho_{N/S} \delta_{\xi,\perp}).
\]

The proof for \( Q_{1,\xi} \) and \( Q_2 \) is essentially the same. \( \square \)

**Proof of Corollary B.7.** Recall that

\[
\frac{d}{dt} \| [K_A]_p \|^2 = 2\left( \frac{[K_A]_p \cdot [K_B Q_1]_p}{N_A N_B d} \sigma_p^2 + 2 \frac{[K_B]_p \cdot [K_B]_p Q_1} {N_A N_B d} \sigma_p^2 + 2 \frac{[K_A]_p \|^2}{N_A d} Q_0 \sigma_p^2 \right)
\]

\[
= \sum_{i=1}^{3} T_i \left( \frac{d}{dt} \| [K_A]_p \|^2 \right).
\]

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

\[
\langle [K_A]_p, [K_B]_p \rangle = \langle [K_A]_p, [K_B]_p \rangle \| Q_1 \|_{p,p} + \sum_{k \neq p} \langle [K_A]_p, [K_B]_k \rangle \| Q_1 \|_{k,p}
\]

\[
= \frac{2K}{1 + K} \| [K_A]_p \| \| [K_B]_p \| \pm O(\tau^2_t d^2 \sigma_p^2) \pm O(\tau^2_t d^2 \sigma_p^2 \rho_{N/S} \delta_{\xi,\perp})
\]

\[
= \frac{2K}{1 + K} \| [K_A]_p \| \| [K_B]_p \| \pm O(\tau^2_t d^2 \sigma_p^2) .
\]

Hence,

\[
T_1 \left( \frac{d}{dt} \| [K_A]_p \|^2 \right) = \frac{4K}{1 + K} \frac{\sigma_p^2}{N_A N_B d} \| [K_A]_p \| \| [K_B]_p \| \pm O(\frac{\sigma_p^2}{N_A N_B d} \tau^2_t d^2 \sigma_p^2).
\]

For \( T_2 \), we compute

\[
T_2 \left( \frac{d}{dt} \| [K_A]_p \|^2 \right) = \frac{2 \sigma_p^2}{N_A N_B d} \sum_{k=1}^{d-r} \| [K_A]_p \| \| [K_B]_k \| \| [Q_1]_{k,p} \|
\]

\[
= \pm O(1) \frac{\sigma_p^2}{N_A N_B d} \sum_{k=1}^{d-r} \| [K_A]_p \| \| [K_B]_k \| \| [Q_1]_{k,p} \| \| [K_A]_p \| \| [K_B]_k \| \| [Q_1]_{k,p} \|
\]

\[
= \pm O\left( \frac{\sigma_p^2}{N_A N_B d} \tau^2_t d^3 \rho_{N/S} \delta_{\xi,\perp} \sigma_p^2 \right).
\]
Finally, for $T_3$, we compute

$$
T_3 \left( \frac{d}{dt} \| [K_A]_p \| ^2 \right) = 2 \left\| [K_A]_p \right\|^2 \left( -\frac{2K}{1 + K} \langle [K_A, K_B] \rangle \pm O (d\tau^2) \right) \sigma_p^2 \\
= -\frac{4K}{1 + K} \frac{\sigma_p^2}{N_A N_B d} \left( 1 \pm \sqrt{\delta_{A/B}} \right) \frac{\langle [K_A, K_B] \rangle}{N_A N_B d} \| [K_A]_p \|^2 \pm O \left( \frac{\kappa_p^2}{N_A N_B d} \tau_t^2 \sigma_p^2 \right)
$$

Combine these together and we obtain

$$
\frac{d}{dt} \| [K_A]_p \|^2 = 2 \frac{\sigma_p^2}{N_A N_B d} \left( \frac{\langle [K_A, [K_B]_p] \rangle}{N_A N_B d} - \frac{\langle [K_A, K_B] \rangle}{N_A N_B d} \| [K_A]_p \|^2 \right) \pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \left( \sqrt{\delta_{A/B} + \tau_t^2} \right) \kappa_p^2
$$

Interchange the roles of $A$ and $B$ and we obtain the formula for $\frac{d}{dt} \| [K_B]_p \|^2$. Now we consider $K_{A, \xi}$. We write

$$
\frac{d}{dt} \| [K_{A, \xi}]_q \|^2 = 2 \frac{\langle [K_{A, \xi}]_q, [K_B]_q \rangle}{N_A N_B d} \sigma_{\xi}^2 + 2 \frac{\langle [K_{A, \xi}]_q, [K_B \xi Q_2]_q \rangle}{N_A N_B d} \sigma_{\xi}^2 + 2 \| [K_{A, \xi}]_q \|^2 \| Q_0 \| \sigma_{\xi}^2
$$

For $T_1$, we compute

$$
T_1 \left( \frac{d}{dt} \| [K_{A, \xi}]_q \|^2 \right) = 2 \sum_{k=1}^{r} \frac{\langle [K_{A, \xi}]_q, [K_B]_k \rangle \langle Q_1, [K_{A, \xi}]_q \rangle}{N_A N_B d} \sigma_{\xi}^2
$$

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

$$
T_3 \left( \frac{d}{dt} \| [K_{A, \xi}]_q \|^2 \right) = 2 \| [K_{A, \xi}]_q \|^2 \left( 1 \pm \sqrt{\delta_{A/B}} \right) \left( -\frac{2K}{1 + K} \frac{\langle [K_A, K_B] \rangle}{N_A N_B d} \pm O (d\tau^2) \right) \sigma_{\xi}^2
$$

$$
= -\frac{4K}{1 + K} \frac{\sigma_{\xi}^2}{N_A N_B d} \frac{\langle [K_A, K_B] \rangle}{N_A N_B d} \| [K_{A, \xi}]_q \|^2 \pm O \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \left( d\tau^2 + \sqrt{\delta_{A/B}} \right) \| [K_{A, \xi}]_q \|^2 \right).
$$
Combine these together and we obtain
\[
\frac{d}{dt} \| [K_{A, \xi}]_p \|^2 = -\frac{4K}{1 + K N_A N_B d} \frac{\sigma_p^2}{N_A N_B d} \langle K_A, K_B \rangle \| [K_{A, \xi}]_p \|^2 + O \left( \frac{\sigma_p^2}{N_A N_B d} (d \tau_t^2 + \sqrt{\delta_{A/B}}) \| [K_{A, \xi}]_p \|^2 \right).
\]
Interchange the roles of \( A \) and \( B \) and we obtain the formula for \( K_{B, \xi} \).

**Proof of Corollary B.8.** We write
\[
\frac{d}{dt} [K_A]_p = \left( I - [K_A]_p ( [K_A]_p )^\top \right) \frac{[K_B Q_1]_p \| [K_A]_p \|^2}{\| [K_A]_p \|^2} \frac{\sigma_p^2}{N_A N_B d} + \left( I - [K_A]_p ( [K_A]_p )^\top \right) \frac{[K_B \xi Q_1, \xi_B]_p \| [K_A]_p \|^2}{\| [K_A]_p \|^2} \frac{\sigma_p^2}{N_A N_B d}.
\]
For the first term, by Lemma B.4, we have
\[
\frac{\| [K_B]_p \| \| Q_1 \|_p}{\| [K_A]_p \|} = \left( 1 \pm \sqrt{\delta_{A/B}} \right) \left( \frac{2K}{1 + K} \pm O (d \tau_t^2) \right) = \frac{2K}{1 + K} + O \left( \sqrt{\delta_{A/B}} \right).
\]
Also by Lemma B.4, for each submand in the second term, we have
\[
\frac{\| [K_B]_p \| \| [Q_1]_{k,p} \|}{\| [K_A]_p \|} = \pm O \left( \tau_t^2 d \kappa_0 \right).
\]
Therefore,
\[
T_1 \left( \frac{d}{dt} [K_A]_p \right) = \frac{2K}{1 + K N_A N_B d} \left( I - [K_A]_p ( [K_A]_p )^\top \right) \frac{\sigma_p^2}{N_A N_B d} \sqrt{\delta_{A/B}} \left( \tau_t^2 d \kappa_0 \right). + \frac{\sigma_p^2}{N_A N_B d} \sum_{k=1}^r O \left( \tau_t^2 d \kappa_0 + \sqrt{\delta_{A/B}} \right) [K_B]_p.
\]
For the second term, by Lemma B.5, we have
\[
T_2 \left( \frac{d}{dt} [K_A]_p \right) = \frac{\sigma_p^2}{N_A N_B d} \sum_{k=1}^{d-r} \left( I - [K_A]_p ( [K_A]_p )^\top \right) \frac{[K_B \xi]_k \| [Q_1, \xi_B]_{k,p} \|}{\| [K_A]_p \|} \frac{\sigma_p^2}{N_A N_B d} \sum_{k=1}^r O \left( \tau_t^2 d \kappa_0 + \sqrt{\delta_{A/B}} \right) [K_B]_k.
\]
Combine these together, and we obtain
\[
\frac{d}{dt} [K_A]_p = \left( I - [K_A]_p ( [K_A]_p )^\top \right) \frac{\sigma_p^2}{N_A N_B d} \left( \frac{2K}{1 + K N_A N_B d} \left( I - [K_A]_p ( [K_A]_p )^\top \right) \frac{\sigma_p^2}{N_A N_B d} \right) \frac{\sigma_p^2}{N_A N_B d} \left( \tau_t^2 d \kappa_0 + \sqrt{\delta_{A/B}} \right) [K_B]_p
\]
\[
= \frac{2K}{1 + K N_A N_B d} \left( I - [K_A]_p ( [K_A]_p )^\top \right) \frac{\sigma_p^2}{N_A N_B d} \left( \tau_t^2 d \kappa_0 + \sqrt{\delta_{A/B}} \right) [K_B]_p.
\]
25
Now, we consider $K_{A,\xi}$. Again, we write
\[
\frac{d}{dt} |K_{A,\xi}|_q = \left( I - |K_{A,\xi}|_q \left( |K_{A,\xi}|_q \right)^T \right) \|K_B Q_1_\xi\|_q \frac{\sigma_\xi^2}{\|K_{A,\xi}\|_q} N_A N_B d \\
+ \left( I - |K_{A,\xi}|_q \left( |K_{A,\xi}|_q \right)^T \right) \|K_B Q_2\|_q \frac{\sigma_\xi^2}{\|K_{A,\xi}\|_q} N_A N_B d \\
=: T_1 \left( \frac{d}{dt} |K_{A,\xi}|_q \right) + T_2 \left( \frac{d}{dt} |K_{A,\xi}|_q \right)
\]
For the first term, by Lemma B.5, we have
\[
T_1 \left( \frac{d}{dt} |K_{A,\xi}|_q \right) = \sum_{k=1}^{r} \left( I - |K_{A,\xi}|_q \left( |K_{A,\xi}|_q \right)^T \right) \|K_B\|_k \frac{\|Q_{1,\xi}\|_q}{\|K_{A,\xi}\|_q} \frac{\sigma_\xi^2}{N_A N_B d} \\
= \pm \sum_{k=1}^{r} O \left( \sigma_\xi^2 \frac{\kappa_0}{N_A N_B d \rho_{N/S}} \tau_i^2 d^2 \rho_{N/S} \delta_{\xi,\perp} \right) \\
= \pm O \left( \frac{\sigma_\xi^2}{N_A N_B d} \tau_i^2 d^2 \kappa_0 \delta_{\xi,\perp} \right).
\]
The same bound also hold for $T_2$. In fact, we can have a slightly sharper bound for it because we no longer have $\|K_B\|_k / \|K_{A,\xi}\|_q$. Combine these and we obtain
\[
\frac{d}{dt} |K_{A,\xi}|_q = \pm O \left( \frac{\sigma_\xi^2}{N_A N_B d} \tau_i^2 d^2 \kappa_0 \delta_{\xi,\perp} \right).
\]

**Proof of Lemma B.9.** The proof of Corollary B.8 and Corollary B.7, mutatis mutandis, yields
\[
\frac{d}{dt} K_A = \frac{2K}{1 + K} \frac{K_B}{N_A N_B d} \text{diag}(\sigma^2) - \frac{2K}{1 + K} \frac{K_A}{N_A N_B d} \|K_B\|_F \left( \frac{d^2 \xi}{N_A^2 d^2} + \sqrt{\delta_{A/B}} \|K_A\|_F \right) \\
\frac{d}{dt} K_{A,\xi} = \frac{2K}{1 + K} \frac{K_A}{N_A N_B d} \frac{\|K_B\|_F}{N_A N_B d} \sigma_\xi^2 \pm O \left( \frac{\sigma_\xi^2}{N_A N_B d} \tau_i^2 d^2 \|K_{A,\xi}\|_F \right).
\]
Recall from Lemma A.6 that, in the non-contrastive case, we have
\[
\frac{d}{dt} K_A = E \left( \frac{f_B^+(z^+)^T}{N_A} - \frac{f_B^+}{N_A^2 d} \frac{K_A}{N_A} \right) \text{diag}(\sigma^2) \\
= \frac{K_B}{N_A N_B d} \text{diag}(\sigma^2) - \frac{\|K_B\|_F}{N_A N_B d} \frac{K_A}{N_A^2 d} \text{diag}(\sigma^2), \\
\frac{d}{dt} K_{A,\xi} = - \frac{\|K_B\|_F}{N_A N_B d} \frac{K_{A,\xi}}{N_A^2 d} \sigma_\xi^2.
\]
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 $1 - \rho_{\alpha}$ and $\rho_{N/S}$ converge to 0 and the growth rate of the condition number. The basic idea is to use the estimations we have derived in Corollary B.7 and Corollary B.8 to approximate the finite-width dynamics with the infinite-width ones.
Lemma B.10 (Convergence rate of $\rho_-$). In Stage 1, we have, for any $p \in [r]$,
\[
\frac{d}{dt} \left( \langle [K_A]_p, [K_B]_p \rangle \right) = \frac{2K}{1 + K N_A N_B d} \sigma_p^2 \left( 1 + \langle [K_A]_p, [K_B]_p \rangle \right)^2 + O \left( \frac{\sigma_p^2}{N_A N_B d} \left( \tau^2 d^2 \kappa_0 + \sqrt{\delta_{A/B}} \right) \right).
\]

Note that $1 + \langle [K_A]_p, [K_B]_p \rangle = \Theta(1)$. Hence, $\langle [K_A]_p, [K_B]_p \rangle$ will converge to 1 at a linear rate.

Lemma B.11 (Convergence rate of $\rho_{N/S}$). In Stage 1, we have
\[
\frac{d}{dt} \rho_{N/S} \leq -\frac{4K}{1 + K N_A N_B d} \sigma_{\min}^2 \rho_{N/S} + O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \tau^2 + \sqrt{\delta_{A/B}} \right) \rho_{N/S} \right).
\]

This lemma says the noise-signal ratio decreases exponentially fast.

Lemma B.12 (Growth rate of the condition number). Define $\rho_{p/q} = \| [K_A]_p \|^2 / \| [K_A]_q \|^2$. In Stage 1, we have
\[
\dot{\rho}_{p/q} \leq \frac{16K}{1 + K N_A N_B d} \left( \rho_- + \min \{ \dot{\rho}_{N/S}, 1 \} \right) \rho_{p/q} + O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \sqrt{\delta_{A/B}} + \tau^2 \right) \rho_{p/q} \right).
\]

Note that error growth slows down as $\rho_-$ and $\dot{\rho}_{N/S}$ decrease. This allows us to use Lemma B.2 (cf. Section B.4).

Omitted proofs of this subsection

Proof of Lemma B.10. By Corollary B.8, we have
\[
\frac{d}{dt} \left( \langle [K_A]_p, [K_B]_p \rangle \right) = \frac{2K}{1 + K N_A N_B d} \sigma_p^2 \left( 1 + \langle [K_A]_p, [K_B]_p \rangle \right)^2 + O \left( \frac{\sigma_p^2}{N_A N_B d} \left( \tau^2 d^2 \kappa_0 + \sqrt{\delta_{A/B}} \right) \right).
\]

Hence, by symmetry, we have
\[
\frac{d}{dt} \left( \langle [K_A]_p, [K_B]_p \rangle \right) = \frac{2K}{1 + K N_A N_B d} \sigma_p^2 \left( 1 + \langle [K_A]_p, [K_B]_p \rangle \right)^2 + O \left( \frac{\sigma_p^2}{N_A N_B d} \left( \tau^2 d^2 \kappa_0 + \sqrt{\delta_{A/B}} \right) \right).
\]

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

\[
\frac{d}{dt} K = \frac{K_B}{N_A N_B d} Q_1 \text{diag}(\sigma^2) + \frac{2K_B}{N_A N_B d} Q_{1,\xi} \text{diag}(\sigma^2) + \frac{K_A}{N_A^2 d} Q_0 \text{diag}(\sigma^2)
\]

Similarly, we also have

\[
\kappa(t) = \frac{2K_B}{N_A N_B d} \text{diag}(\sigma^2) + \frac{2K_B}{N_A N_B d} \frac{\langle K_A, K_B \rangle}{N_A N_B d} \text{diag}(\sigma^2)
\]

Define \( K = K_A + K_B \). Then, by symmetry, we have

\[
\frac{d}{dt} K = \frac{2K}{1 + K} \left( \frac{K_A}{N_A N_B d} - 1 \right) \text{diag}(\sigma^2) \pm O \left( \frac{\sigma_{max}^2}{N_A N_B d} \right) (\sqrt{\frac{\delta_A}{\delta_B}}) ,
\]

Hence,

\[
\frac{d}{dt} K = \frac{4K}{1 + K} \left( \frac{K_A}{N_A N_B d} - 1 \right) \text{diag}(\sigma^2) \pm O \left( \frac{\sigma_{max}^2}{N_A N_B d} \right) (\sqrt{\frac{\delta_A}{\delta_B}}) ,
\]

Similarly, we also have

\[
\frac{d}{dt} K_A = \frac{K_B}{N_A N_B d} Q_{1,\xi} \sigma_{\xi}^2 + \frac{K_B \xi}{N_A N_B d} Q_{2,\xi} \sigma_{\xi}^2 + \frac{K_A \xi}{N_A^2 d} Q_0 \sigma_{\xi}^2
\]

Define \( K = K_A + K_B \). Then, by symmetry, we have

\[
\frac{d}{dt} K = \frac{2K}{1 + K} \left( \frac{K_A}{N_A N_B d} - 1 \right) \text{diag}(\sigma^2) \pm O \left( \frac{\sigma_{max}^2}{N_A N_B d} \right) (\sqrt{\frac{\delta_A}{\delta_B}}) ,
\]

Hence,

\[
\frac{d}{dt} K = \frac{4K}{1 + K} \left( \frac{K_A}{N_A N_B d} - 1 \right) \text{diag}(\sigma^2) \pm O \left( \frac{\sigma_{max}^2}{N_A N_B d} \right) (\sqrt{\frac{\delta_A}{\delta_B}}) ,
\]

Therefore, we compute

\[
\frac{d}{dt} \rho_{N/S}^2 = \frac{d}{dt} \|K_A\|_F^2 - \rho_{N/S}^2 \frac{d}{dt} \|K_A\|_F^2
\]

Then, we compute

\[
\frac{d}{dt} \rho_{N/S}^2 \leq - \frac{4K}{1 + K} \frac{\sigma_{\xi}^2}{N_A^2 d} \rho_{N/S}^2 \|K_A\|_F^2 \pm O \left( \frac{\sigma_{\xi}^2}{N_A^2 d} \right) (\sqrt{\frac{\delta_A}{\delta_B}}) ,
\]

\[
\rho_{N/S}^2 \left( 4K \sigma_{\xi}^2 \left( \frac{K_A}{N_A N_B d} - 1 \right) \rho_{N/S}^2 \|K_A\|_F^2 \pm O \left( \frac{\sigma_{\xi}^2}{N_A^2 d} \right) (\sqrt{\frac{\delta_A}{\delta_B}}) \right)
\]

\[
\leq - \frac{4K}{1 + K} \frac{\sigma_{\xi}^2}{N_A N_B d} \rho_{N/S}^2 + O \left( \frac{\sigma_{max}^2}{N_A N_B d} \right) (\sqrt{\frac{\delta_A}{\delta_B}}) .
\]
Proof of Lemma B.12. For notational simplicity, define \( \rho_{p/q} := \frac{||[K_A]_p||^2}{||[K_A]_q||^2} \). By Corollary B.7, we have

\[
\hat{\rho}_{p/q} = \frac{d}{dt} \frac{||[K_A]_p||^2}{||[K_A]_q||^2} - \rho_{p/q} \frac{d}{dt} \frac{||[K_A]_p||^2}{||[K_A]_q||^2} = \frac{4K}{1 + K N_A N_B d} \frac{\sigma_p^2}{||[K_A]_p||^2} \left( \frac{\langle [K_A]_p, [B]_p \rangle}{||[K_A]_p||^2} \right) \rho_{p/q} + O \left( \frac{\sigma_p^2}{N_A N_B d} \sqrt{\frac{\delta_{A/B}}{N_A N_B d^2} + \tau^2 d} \rho_{p/q} \right) - \rho_{p/q} \frac{4K}{1 + K N_A N_B d} \frac{\sigma_q^2}{||[K_A]_q||^2} \left( \frac{\langle [K_A]_q, [B]_q \rangle}{||[K_A]_q||^2} \right) \rho_{p/q} \nonumber
\]

Now we bound \( \langle [K_A]_p, [B]_p \rangle - \langle [K_A, K_B] \rangle / (N_A N_B d) \). Clear that this term is bounded by 2. Meanwhile, by definition, we have \( \langle [K_A]_p, [B]_p \rangle = 1 + \rho_- \). For the second term, we have

\[
\frac{\langle [K_A, K_B] \rangle}{N_A N_B d} = \sum_{k=1}^r \frac{\langle [K_A]_k, [B]_k \rangle}{N_A N_B d} \frac{||K_A||_k ||K_B||_k}{||K_A||_F ||K_B||_F} = \sum_{k=1}^r (1 + \rho_-) \left( 1 - \min \{\hat{\rho}_{N/S, 1} \} \right) \frac{\sigma_k^2}{\|\kappa\|^2} (1 + \sqrt{\delta_{A/B}}) = (1 + \rho_- + \min \{\hat{\rho}_{N/S, 1} \}) (1 + \sqrt{\delta_{A/B}}).
\]

Combine these together and we obtain

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

The same is also true for \( q \). Thus,

\[
\hat{\rho}_{p/q} \leq \frac{16K}{1 + K N_A N_B d} + \min \{\hat{\rho}_{N/S, 1} \} \rho_{p/q} + O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \sqrt{\frac{\delta_{A/B}}{N_A N_B d^2} + \tau^2 d} \rho_{p/q} \right).
\]

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 \( ||[K_A]_p||^2 \) and \( ||[K_B]_p||^2 \). Instead of directly control the difference, we define

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

Note that \( \rho_{A/B, p} + \rho_{B/A, p} \geq 2, \) with equality obtained iff \( ||[K_A]_p||^2 = ||[K_B]_p||^2 \). 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 \( \rho_{A/B, p} + \rho_{B/A, p} \) is to leverage the symmetry. Similarly, we also define

\[
\rho_{A/B, \xi, q} := \frac{||[K_A]_{\xi, q}||^2}{||[K_B]_{\xi, q}||^2} \quad \text{and} \quad \rho_{B/A, \xi, q} := \frac{||[K_B]_{\xi, q}||^2}{||[K_A]_{\xi, q}||^2},
\]

and analyze \( \rho_{A/A, \xi} + \rho_{B/B, \xi} \).
Lemma B.13 (Difference of diagonal terms). In Stage 1, we have
\[
\frac{d}{dt} (\rho_{A/B,p} + \rho_{B/A,p}) \leq O\left(\frac{\sigma_p^2}{N^2N_B d} \left(\delta_{A/B}^2 + \tau_t^2d^2\right)\right),
\]
\[
\frac{d}{dt} (\rho_{A/B,\xi,q} + \rho_{B/A,\xi,q}) \leq O\left(\frac{\sigma_{\max}^2}{N^2N_B d} \left(\delta_{A/B}^2 + \tau_t^2d^2\right)\right).
\]

Note that the RHS are higher-order terms, which implies the relative difference of the norms barely grows.

Now we consider the condition number of \(K_{A,\xi}\). Unlike \(K_A\), for the noises, the \(\sigma\)'s for different coordinates are the same. Hence, we have the following simple bound on the growth rate of \(\delta_{\xi,\kappa_0}\).

**Lemma B.14** (Condition number of \(K_{A,\xi}\)). Define \(\rho_{\xi,p/q} := \|K_{A,\xi,p}\|/\|K_{A,\xi,q}\|^2\). In Stage 1, we have
\[
\dot{\rho}_{\xi,p/q} \leq O\left(\frac{\sigma_{\xi}^2}{N^2N_B d} \left(d\tau_t^2 + \sqrt{\delta_{A/B}}\right)\right).
\]

Again, the RHS is a higher-order term that can be made arbitrarily small by choosing a small \(\tau_t^2\) and maintaining a small \(\delta_{A/B}\).

Then, we consider the orthogonality conditions.

**Lemma B.15** (Orthogonality between signals). For any \(p \neq q\), define
\[
\hat{\delta}_{\perp,p,q} := \left\langle [K_A]_{p,r} [K_B]_{q,r} \right\rangle^2 + \left\langle [K_B]_{p,r} [K_A]_{q,r} \right\rangle^2 + \left\langle [K_A]_{p,r} [K_A]_{q,r} \right\rangle^2 + \left\langle [K_B]_{p,r} [K_B]_{q,r} \right\rangle^2.
\]

In Stage 1, we have
\[
\frac{d}{dt} \hat{\delta}_{\perp,p,q} \leq O\left(\frac{\sigma_{\max}^2}{N^2N_B d} \left(\tau_t^2 d^2 \kappa_0 + d\sqrt{\delta_{A/B}}\right)\right).
\]

Recall that \(\rho_{\perp}\) converges to 0 at a sufficiently fast rate. As a result, 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
\[
\hat{\delta}_{\perp,\xi,p,q} := \left\langle [K_A]_{p,r} [K_A,\xi]_{q,r} \right\rangle^2 + \left\langle [K_B]_{p,r} [K_A,\xi]_{q,r} \right\rangle^2.
\]

and define \(\hat{\delta}_{\perp,\xi,B,p,q}\) similarly. In Stage 1, we have
\[
\frac{d}{dt} \hat{\delta}_{\perp,\xi,p,q} \leq O\left(\frac{\sigma_p^2}{N^2N_B d} \left(\rho_{\perp}\hat{\delta}_{\perp,\xi,p,q} \pm \delta_{\perp,\xi,p,q} \right)\right) + O\left(\frac{\sigma_{\max}^2}{N^2N_B d} \left(\tau_t^2 d^2 \kappa_0 + d\sqrt{\delta_{A/B}}\right)\right).
\]

For any \(q, s \in [d-r]\), define
\[
\hat{\delta}_{\perp,\xi,p,q} := \left\langle [K_A,\xi]_{q,r} [K_B,\xi]_{s,r} \right\rangle^2 + \left\langle [K_B,\xi]_{q,r} [K_A,\xi]_{s,r} \right\rangle^2.
\]

In Stage 1, we have
\[
\frac{d}{dt} \hat{\delta}_{\perp,\xi,p,q} \leq O\left(\frac{\sigma_{\xi}^2}{N^2N_B d} \left(\tau_t^2 \kappa_0 \delta_{\perp,\xi} \right) \sqrt{\hat{\delta}_{\perp,\xi,p,q}}\right).
\]
OMITTED PROOFS OF THIS SUBSECTION

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

\[
\frac{d}{dt} \frac{||[K_A]||^2}{||[K_B]||^2} = \frac{4K}{1+K} \frac{\sigma_p^2}{N_A N_B d^2} \langle [K_A], [K_B] \rangle - \frac{4K}{1+K} \frac{\langle [K_A], [K_B] \rangle \sigma_p^2}{N_A N_B d} \frac{||[K_A]||^2}{N_A^2 d^2}
\]

\[
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \tau^2 d \right).
\]

Interchange the roles of \( A \) and \( B \) and we obtain

\[
\frac{d}{dt} \frac{||[K_B]||^2}{||[K_A]||^2} = \frac{4K}{1+K} \frac{\sigma_p^2}{N_A N_B d^2} \langle [K_A], [K_B] \rangle - \frac{4K}{1+K} \frac{\langle [K_A], [K_B] \rangle \sigma_p^2}{N_A N_B d} \frac{||[K_B]||^2}{N_B^2 d^2}
\]

\[
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \tau^2 d \right).
\]

For notational simplicity, define \( \rho_{A/B,p} = \frac{||[K_A]||^2}{||[K_B]||^2} \). Then, we compute

\[
\dot{\rho}_{A/B,p} = \frac{d}{dt} \frac{||[K_A]||^2}{||[K_B]||^2} - \rho_{A/B,p} \frac{d}{dt} \frac{||[K_B]||^2}{||[K_A]||^2}
\]

\[
= \frac{4K}{1+K} \frac{\sigma_p^2}{N_A N_B d^2} \langle [K_A], [K_B] \rangle \rho_{A/B,p} - \frac{4K}{1+K} \frac{\langle [K_A], [K_B] \rangle \sigma_p^2 \rho_{A/B,p}}{N_A N_B d} \frac{||[K_A]||^2}{N_A^2 d^2}
\]

\[
- \frac{4K}{1+K} \frac{\sigma_p^2}{N_A N_B d^2} \langle [K_A], [K_B] \rangle \rho_{A/B,p} + \frac{4K}{1+K} \frac{\langle [K_A], [K_B] \rangle \sigma_p^2 \rho_{A/B,p}}{N_A N_B d} \frac{||[K_B]||^2}{N_B^2 d^2}
\]

\[
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \tau^2 d \right).
\]

Then, by symmetry, we have

\[
\frac{d}{dt} (\rho_{A/B,p} + \rho_{B/A,p})
\]

\[
= \frac{4K}{1+K} \frac{\sigma_p^2}{N_A N_B d^2} \langle [K_A], [K_B] \rangle \left( \frac{1}{||[K_A]||^2} - \frac{1}{||[K_B]||^2} \right) (\rho_{A/B,p} - \rho_{B/A,p})
\]

\[
- \frac{4K}{1+K} \frac{\langle [K_A], [K_B] \rangle \sigma_p^2}{N_A N_B d} \left( \frac{1}{N_A^2 d} - \frac{1}{N_B^2 d} \right) (\rho_{A/B,p} - \rho_{B/A,p})
\]

\[
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \tau^2 d \right)
\]

\[
\leq O \left( \frac{\sigma_p^2}{N_A N_B d} (\sigma_{A/B}^2 + \tau^2 d) \right).
\]

The above proof, \textit{mutatis mutandis}, yields the result for \( \rho_{A/B,\xi,q} + \rho_{B/A,\xi,q} \). \qed
Proof of Lemma B.14. By Corollary B.7, we have
\[
\dot{\rho}_{\xi,p/q} = \frac{4}{N} \left( \frac{\|K_{A,\xi}\|}{\|K_{A,\xi}\|} \right)^2 - \rho_{\xi,p/q} \frac{4}{N} \left( \frac{\|K_{A,\xi}\|}{\|K_{A,\xi}\|} \right)^2
\]
\[
= -\frac{4K}{1 + K N_{A}N_{B}} \|K_{A,\xi}\| \|K_{B}\| \rho_{\xi,p/q} \pm O \left( \frac{\sigma_{\xi}^2}{N_{A}N_{B}} \left( d\tau_{t}^2 + \sqrt{\delta_{A/B}} \right) \right)
\]
\[
- \rho_{\xi,p/q} \left( -\frac{4K}{1 + K N_{A}N_{B}} \|K_{A,\xi}\| \|K_{B}\| \pm O \left( \frac{\sigma_{\xi}^2}{N_{A}N_{B}} \left( d\tau_{t}^2 + \sqrt{\delta_{A/B}} \right) \right) \right)
\]
\[
= \pm O \left( \frac{\sigma_{\xi}^2}{N_{A}N_{B}} \left( d\tau_{t}^2 + \sqrt{\delta_{A/B}} \right) \right).
\]

Proof of Lemma B.15. By Corollary B.8, we have
\[
\left\langle \frac{d}{dt} [K_{A,p}, [K_{B,q}]_{q}] \right\rangle = \frac{2K}{1 + K N_{A}N_{B}} \left( \left\langle [K_{B,1}, [K_{B}, q]]_{q} \right\rangle - \left\langle [K_{A,1}, [K_{B}, q]]_{q} \right\rangle \right)
\]
\[
\pm O \left( \frac{\sigma_{p}^2}{N_{A}N_{B}} \left( \tau_{t}^2 d^2 \kappa_0 + d\sqrt{\delta_{A/B}} \right) \right)
\]
\[
= \frac{2K}{1 + K N_{A}N_{B}} \left( \left\langle [K_{B,1}, [K_{B}, q]]_{q} \right\rangle - \left\langle [K_{A,1}, [K_{B}, q]]_{q} \right\rangle \right)
\]
\[
+ \frac{2K}{1 + K N_{A}N_{B}} \left\langle [K_{A,1}, [K_{B}, q]]_{q} \right\rangle \left( 1 - \left\langle [K_{A,1}, [K_{B}, q]]_{q} \right\rangle \right)
\]
\[
\pm O \left( \frac{\sigma_{p}^2}{N_{A}N_{B}} \left( \tau_{t}^2 d^2 \kappa_0 + d\sqrt{\delta_{A/B}} \right) \right).
\]
Interchange the roles of \(p, q\) and \(A, B\) and we obtain
\[
\left\langle [K_{A,p}, \frac{d}{dt} [K_{B,q}]_{q}] \right\rangle = \frac{2K}{1 + K N_{A}N_{B}} \left( \left\langle [K_{A,p}, [K_{B,q}]_{q}]_{q} \right\rangle - \left\langle [K_{A,p}, [K_{B,q}]_{q}]_{q} \right\rangle \right)
\]
\[
+ \frac{2K}{1 + K N_{A}N_{B}} \left\langle [K_{A,p}, [K_{B,q}]_{q}]_{q} \right\rangle \left( 1 - \left\langle [K_{A,p}, [K_{B,q}]_{q}]_{q} \right\rangle \right)
\]
\[
\pm O \left( \frac{\sigma_{q}^2}{N_{A}N_{B}} \left( \tau_{t}^2 d^2 \kappa_0 + d\sqrt{\delta_{A/B}} \right) \right).
\]
Therefore,
\[
\frac{d}{dt} \left\langle [K_{A,p}, [K_{B,q}]_{q}] \right\rangle = \frac{2K}{1 + K N_{A}N_{B}} \left( \left\langle [K_{B,1}, [K_{B}, q]]_{q} \right\rangle - \left\langle [K_{A,1}, [K_{B}, q]]_{q} \right\rangle \right)
\]
\[
+ \frac{2K}{1 + K N_{A}N_{B}} \left\langle [K_{A,1}, [K_{B}, q]]_{q} \right\rangle \left( 1 - \left\langle [K_{A,1}, [K_{B}, q]]_{q} \right\rangle \right)
\]
\[
\pm O \left( \frac{\sigma_{\max}^2}{N_{A}N_{B}} \left( \tau_{t}^2 d^2 \kappa_0 + d\sqrt{\delta_{A/B}} \right) \right).
\]
Interchange the roles of $p$ and $q$ and we obtain
\[
\frac{d}{dt} \left( [K_B]_p, [K_A]_q \right) = 2K \frac{\sigma_q^2}{1 + K N_A N_B d} \left( \left( [K_B]_p, [K_B]_q \right) - \left( [K_B]_p, [K_A]_q \right) \right)
+ 2K \frac{\sigma_p^2}{1 + K N_A N_B d} \left( \left( [K_A]_p, [K_A]_q \right) - \left( [K_B]_p, [K_A]_q \right) \right)
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \right) \rho \left( [K_B]_p, [K_A]_q \right)
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \tau_1^2 d^2 \kappa_0 + d \sqrt{\delta A/B} \right) \right).
\]

Similarly, we compute
\[
\left\langle \frac{d}{dt} [K_A]_p, [K_A]_q \right\rangle = 2K \frac{\sigma_p^2}{1 + K N_A N_B d} \left( \left( [K_B]_p, [K_A]_q \right) - \left( [K_A]_p, [K_A]_q \right) \right)
+ 2K \frac{\sigma_p^2}{1 + K N_A N_B d} \left( \left( [K_A]_p, [K_A]_q \right) - \left( [K_B]_p, [K_B]_q \right) \right)
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \tau_1^2 d^2 \kappa_0 + d \sqrt{\delta A/B} \right) \right)
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \tau_1^2 d^2 \kappa_0 + d \sqrt{\delta A/B} \right) \right).
\]

Then, by interchanging the roles of $p$ and $q$, we obtain
\[
\left\langle [K_A]_p, \frac{d}{dt} [K_A]_q \right\rangle = 2K \frac{\sigma_q^2}{1 + K N_A N_B d} \left( \left( [K_A]_p, [K_B]_q \right) - \left( [K_A]_p, [K_A]_q \right) \right)
+ 2K \frac{\sigma_q^2}{1 + K N_A N_B d} \left( \left( [K_A]_p, [K_A]_q \right) - \left( [K_B]_p, [K_B]_q \right) \right)
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \tau_1^2 d^2 \kappa_0 + d \sqrt{\delta A/B} \right) \right).
\]

Add them together and we get
\[
\frac{d}{dt} \left( [K_A]_p, [K_A]_q \right) = 2K \frac{\sigma_p^2}{1 + K N_A N_B d} \left( \left( [K_A]_p, [K_B]_q \right) - \left( [K_A]_p, [K_A]_q \right) \right)
+ 2K \frac{\sigma_p^2}{1 + K N_A N_B d} \left( \left( [K_A]_p, [K_A]_q \right) - \left( [K_B]_p, [K_B]_q \right) \right)
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \tau_1^2 d^2 \kappa_0 + d \sqrt{\delta A/B} \right) \right).
\]

Interchange the roles of $p$ and $q$ and we obtain
\[
\frac{d}{dt} \left( [K_B]_p, [K_B]_q \right) = 2K \frac{\sigma_q^2}{1 + K N_A N_B d} \left( \left( [K_B]_p, [K_B]_q \right) - \left( [K_B]_p, [K_B]_q \right) \right)
+ 2K \frac{\sigma_q^2}{1 + K N_A N_B d} \left( \left( [K_B]_p, [K_B]_q \right) - \left( [K_B]_p, [K_B]_q \right) \right)
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} \left( \tau_1^2 d^2 \kappa_0 + d \sqrt{\delta A/B} \right) \right).
\]
For notational simplicity, define $Z_1 = \langle [K_A]_p, [K_B]_q \rangle$, $Z_2 = \langle [K_B]_p, [K_A]_q \rangle$, $Z_3 = \langle [K_A]_p, [K_A]_q \rangle$, and $Z_4 = \langle [K_B]_p, [K_B]_q \rangle$. Also define $G_p = \frac{2K}{1 + K} \sigma_p^2 A_{	ext{NBA}}$. Then, we can summarize the above equations as

$$
\frac{d}{dt} [Z_1, Z_2, Z_3, Z_4] = \begin{bmatrix}
-G_p - G_q & 0 & G_q & G_p \\
0 & -G_p - G_q & G_q & 0 \\
G_q & G_p & -G_p - G_q & 0 \\
G_p & 0 & 0 & -G_p - G_q \\
\end{bmatrix} \begin{bmatrix}
Z_1 \\
Z_2 \\
Z_3 \\
Z_4 \\
\end{bmatrix} \\
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \rho \left( \begin{array}{c}
\tau_t^2 d^2 \kappa_0 \\
\tau_t \sqrt{\delta_{A/B}} \\
\end{array} \right) \\
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \rho \left( \begin{array}{c}
\tau_t^2 d^2 \kappa_0 \\
\tau_t \sqrt{\delta_{A/B}} \\
\end{array} \right).
$$

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,

$$
\frac{d}{dt} ||Z||^2 \leq O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \rho \left( \tau_t^2 d^2 \kappa_0 + d \sqrt{\delta_{A/B}} \right).
$$

\hfill \Box

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

$$
\langle \frac{d}{dt} [K_A]_p, [K_A]_q \rangle = \frac{2K}{1 + K N_A N_B d} \left( \langle [K_B]_p, [K_A]_q \rangle - \langle [K_A]_p, [K_A]_q \rangle \langle [K_A]_p, [K_B]_p \rangle \right) \\
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \tau_t^2 d^2 \kappa_0 \sqrt{\delta_{A/B}}.
$$

and

$$
\langle [K_A]_p, \frac{d}{dt} [K_A]_q \rangle = \pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \tau_t^2 d^2 \kappa_0 \delta_{A,q}.
$$

Therefore,

$$
\frac{d}{dt} \langle [K_A]_p, [K_A]_q \rangle = \frac{2K}{1 + K N_A N_B d} \left( \langle [K_B]_p, [K_A]_q \rangle - \langle [K_A]_p, [K_A]_q \rangle \langle [K_A]_p, [K_B]_p \rangle \right) \\
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \rho \left( \tau_t^2 d^2 \kappa_0 \right) \pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \tau_t^2 d^2 \kappa_0 \sqrt{\delta_{A/B}}.
$$

Similarly, we also have

$$
\frac{d}{dt} \langle [K_B]_p, [K_A]_q \rangle = \frac{2K}{1 + K N_A N_B d} \left( \langle [K_A]_p, [K_A]_q \rangle - \langle [K_B]_p, [K_A]_q \rangle \langle [K_A]_p, [K_B]_p \rangle \right) \\
\pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \rho \left( \tau_t^2 d^2 \kappa_0 \right) \pm O \left( \frac{\sigma_p^2}{N_A N_B d} \right) \tau_t^2 d^2 \kappa_0 \sqrt{\delta_{A/B}}.
$$
Thus,
\[
\frac{d}{dt} \left( \left\langle [K_A]_p, [K_A]_q \right\rangle + \left\langle [K_B]_p, [K_A]_q \right\rangle \right) = -\frac{4K}{1 + K N A N B d} \left( \left\langle [K_A]_p, [K_A]_q \right\rangle - \left\langle [K_B]_p, [K_A]_q \right\rangle \right)^2 \\
\leq O \left( \frac{\sigma_{\max}^2}{N A N B d} \rho_{\min} \left( \left\langle [K_A]_p, [K_A]_q \right\rangle + \left\langle [K_B]_p, [K_A]_q \right\rangle \right)^2 \right) \\
\leq O \left( \frac{\sigma_{\max}^2}{N A N B d} \left( \tau_i^2 d^2 \kappa_0 + d \sqrt{\delta_{A/B}} \right) \right).
\]

For the orthogonality between noises, by Corollary B.8, we have
\[
\left\langle [K_A]_p, \frac{d}{dt} [K_A]_q \right\rangle = \pm O \left( \frac{\sigma^2}{N A N B d} \tau_i^2 d^3 \kappa_0 \delta_{\perp} \right).
\]

Clear that this bound holds for all other combinations. Thus,
\[
\frac{d}{dt} \delta_{\perp, p, q} \leq O \left( \frac{\sigma^2}{N A N B d} \tau_i^2 d^3 \kappa_0 \delta_{\perp, p, q} \right).
\]

\begin{proof}[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
\[
\frac{d}{dt} \rho_{\min} \leq -\Omega(1) \frac{\sigma_{\min}^2}{N A N B d} \rho_{\min} + O \left( \frac{\sigma_{\max}^2}{N A N B d} \left( \tau_i^2 d^2 \kappa_0 + \sqrt{\delta_{A/B}} \right) \right),
\]
\[
\frac{d}{dt} \hat{\rho}_{N/S} \leq -\Omega(1) \frac{\sigma_{\min}^2}{N A N B d} \hat{\rho}_{N/S} + O \left( \frac{\tau_i^2 d^2 \kappa_0 + \sqrt{\delta_{A/B}}}{N A N B d} \right).
\]

Define \( \rho := \max \{ \rho_{\min}, \hat{\rho}_{N/S} \} \) to be the indicator of the progress we have made. We have
\[
\frac{d}{dt} \rho \leq -\Omega(1) \frac{\sigma_{\min}^2}{N A N B d} \rho + O \left( \frac{\sigma_{\max}^2}{N A N B d} \left( \tau_i^2 d^2 \kappa_0 + \sqrt{\delta_{A/B}} \right) \right). \tag{11}
\]

Let \( \hat{\kappa}_0 := \max_{p, q} ||[K_A]_p||^2 / ||[K_A]_q||^2 \) be the condition number at time \( t \). By Lemma B.12, we have
\[
\frac{d}{dt} \hat{\kappa}_0 \leq O(1) \frac{\sigma_{\max}^2}{N A N B d} \rho \hat{\kappa}_0. \tag{12}
\]

Now we consider the discretization errors, define
\[
\hat{\delta}_{A/B} = \max_{p \in [r], q \in [d-r]} \{|\rho_{A/B, p} + \rho_{B/A, p} - 2, \rho_{A/B, \xi, q} + \rho_{B/A, \xi, q} - 2\}.
\]

Note that at time \( t \), the first condition of (10) holds with \( \hat{\delta}_{A/B} \) replaced by \( O(\hat{\delta}_{A/B}) \). Meanwhile, by Lemma B.13, we have
\[
\frac{d}{dt} \hat{\delta}_{A/B} \leq O \left( \frac{\sigma_{\max}^2}{N A N B d} \left( \hat{\delta}_{A/B}^2 + \tau_i^2 d \right) \right). \tag{13}
\]
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

\[
\frac{d}{dt} \hat{\delta}_{\xi,\kappa_0} \leq O \left( \frac{\sigma_1^2}{N d} \left( ds^2 + \sqrt{\hat{\delta}_1} \right) \right).
\]

Then, define

\[
\hat{\delta}_\perp := \max \left\{ \hat{\delta}_{1,q,p,k}, \hat{\delta}_{1,q,k,l} : q \neq k, k, l \in [d-r] \right\}
\]

Clear that the last two conditions hold at time \( t \) when \( \hat{\delta}_{AB,\perp} \) and \( \hat{\delta}_{\xi,\perp} \) are replaced by \( \sqrt{\hat{\delta}_1(t)} \). By Lemma B.15 and Lemma B.16, we have

\[
\frac{d}{dt} \hat{\delta}_\perp \leq O \left( \frac{\sigma_{\max}^2}{N d} \right) \hat{\rho}_{\perp} + O \left( \frac{\sigma_{\max}^2}{N d} \left( \tau_1^2 d^2 + d \sqrt{\hat{\delta}_1} \right) \right).
\]

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

First, we consider \( \hat{\delta}_{AB} \). Note that the dependence of the RHS of (13) on \( \hat{\delta}_{AB} \) is quadratic. Hence, by making the initial value of \( \hat{\delta}_{AB} \), the RHS can be made to be dominated the \( \tau_1^2 \)-related terms. Hence, \( \hat{\delta}_{AB} \leq O \left( \frac{\sigma_{\max}^2}{N d} \tau_1^2 d T_1 \right) \). As we will see later, \( T_1 = \text{poly}(d) \). Therefore, by choosing \( \hat{\delta}_{AB} \) sufficiently small \( \tau_1^2 \), we can make \( \hat{\delta}_{AB} \) 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 \( \tau_1^2 \). The same is also true for (11). Hence, for these quantities, we have

\[
\frac{d}{dt} \hat{\kappa}_0 \leq -\Omega(1) \frac{\sigma_{\min}^2}{N d} \rho \left( O \left( \frac{\sigma_{\max}^2}{\sigma_{\min}^2} \right) \rho(0) \right), \quad \frac{d}{dt} \hat{\delta}_\perp \leq \hat{\delta}_\perp(0) \exp \left( O \left( \frac{\sigma_{\max}^2}{\sigma_{\min}^2} \right) \rho(0) \right).
\]

Note that \( \rho(0) = O(1) \) and \( \hat{\rho}_{N/S} \leq \left\lfloor \frac{d-r}{r} \sigma_1^2 \right\rfloor \). Therefore,

\[
\exp \left( O \left( \frac{\sigma_{\max}^2}{\sigma_{\min}^2} \right) \rho(0) \right) \leq \exp \left( O(1) \frac{\sigma_{\max}^2}{\sigma_{\min}^2} \max \left\{ 1, \frac{(d-r)\sigma_1^2}{r} \right\} \right) \leq \exp \left( \frac{1}{2} \log d \right) = \sqrt{d}.
\]

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 \( ||[K_A]_p||^2 \) can at most grow \( \sqrt{d} \) times in Stage 1. As a result, \( 1/(N d) \) is lower bounded by some \( 1/\text{poly}(d) \). Thus, by (11), \( T_1 \leq \text{poly}(d) \).

\[\square\]

### B.5 Negative Results

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

**Proof.** We choose \( d = r \) and \( \sigma_1^2 = c \log d, \sigma_2^2 = \ldots = \sigma_{d}^2 = 1 \). 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

\[
\frac{d}{dt} \langle K_A, K_B \rangle \approx \frac{4K}{1 + K} \frac{\sigma_p^2}{N d} \left( 1 + \langle K_A, K_B \rangle \right) \left( 1 - \langle K_A, K_B \rangle \right).
\]
By the proof of Lemma B.12, we have
\[ \dot{\rho}_{p/q} \simeq \frac{4K}{1 + K N_A N_B d} \frac{\sigma_p^2}{\left\langle \| K_A \|_p, \| K_B \|_p \right\rangle - \frac{\left\langle K_A, K_B \right\rangle}{N_A N_B d}} \rho_{p/q} \]
\[ - \frac{4K}{1 + K N_A N_B d} \frac{\sigma_q^2}{\left\langle \| K_A \|_q, \| K_B \|_q \right\rangle - \frac{\left\langle K_A, K_B \right\rangle}{N_A N_B d}} \rho_{p/q}. \]

Note that, in the infinite-width case, we have
\[ (16) \]
Note that, in the infinite-width case, we have
\[ (16) \]
Lemma C.1 (Stage 2)
the following.
Therefore, \( \left\langle K_A \|_p, K_B \|_p \right\rangle - \frac{\left\langle K_A, K_B \right\rangle}{N_A N_B d} \geq 0 \) and \( \left\langle K_A \|_q, K_B \|_q \right\rangle - \frac{\left\langle K_A, K_B \right\rangle}{N_A N_B d} \leq 0 \) for any \( q \geq 2 \). Hence,
\[ \dot{\rho}_{1/2} \geq \frac{4K}{1 + K N_A N_B d} \frac{\sigma_1^2}{\left\langle \| K_A \|_1, \| K_B \|_1 \right\rangle - \frac{\left\langle K_A, K_B \right\rangle}{N_A N_B d}} \rho_{1/2} \]
\[ \geq \frac{4K}{1 + K N_A N_B d} \frac{\sigma_1^2}{\left\langle \| K_A \|_1, \| K_B \|_1 \right\rangle - \frac{\left\langle K_A, K_B \right\rangle}{N_A N_B d}} \left( 1 - \frac{\kappa_1^2}{||\kappa||^2} \right) \left( \left\langle \| K_A \|_1, \| K_B \|_1 \right\rangle - \frac{\left\langle K_A, K_B \right\rangle}{N_A N_B d} \frac{d - 1}{d} \right) \rho_{1/2}. \]

For notational simplicity, define \( X_1 = 1 - \left\langle K_A \|_1, \| K_B \|_1 \right\rangle \), \( X_2 = 1 - \left\langle K_A \|_2, \| K_B \|_2 \right\rangle \), \( Y = \rho_{1/2}, A = \frac{4K}{1 + K N_A N_B d} \). Then we have
\[ X_1 \leq -\sigma_1^2 A X_1, \quad X_2 \geq -2 A X_2. \]

First, by Gronwall’s lemma, we have \( X_1(T) \leq \exp \left( -\sigma_1^2 \int_0^T A \right) \) and
\[ X_2(T) \geq \exp \left( -2 \int_0^T A \right) \geq X_2^{2/\sigma_1^2}(T). \]

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,
\[ \rho_{1/2}(T_2) \geq \rho_{1/2}(T_1) \exp \left( \Omega(1) \sigma_1^2 \int_{T_1}^{T_2} A \right) \geq 2^{O(1)} \Omega(1). \]

\[ \square \]

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\(^3\). Let \( c_{\text{target}} > 1 \) be a constant. Let \( T_1 \) be the earliest time that
\[ \frac{\| [K_A]_p \|_p^2}{\| [K_A]_q \|_q^2} \leq c_{\text{target}}, \quad \forall p, q \in [r]. \]

\(^3\)By Lemma B.1, this condition indeed holds.
We have $T_1 \leq \text{poly}(d)$. Moreover, throughout Stage 2, we have

\[
\max \left\{ 1 - \left( \frac{K}{\|K\|} \right)^p, 1 - \frac{\|K\|^p}{\|K\|^q} \right\} \leq \delta_-, \quad \forall p \in [r],
\]

\[
\max \left\{ 1 - \frac{\|K\|^p}{\|K\|^q}, 1 - \frac{\|K\|^p}{\|K\|^q} \right\} \leq \delta_-, \quad \forall p, q \in [d - r],
\]

\[
\max \left\{ \frac{\|K\|^p}{\|K\|^q} : C, D \in \{A, B\} \right\} \leq \delta_{N/S}, \quad \forall p \in [r], q \in [d - r],
\]

\[
\max \left\{ \frac{\|K\|^p}{\|K\|^q} : C, D \in \{A, B\} \right\} \leq \delta_{AB,\perp}, \quad \forall p \neq q \in [r],
\]

\[
\max \left\{ \frac{\|K\|^p}{\|K\|^q}, \frac{\|K\|^p}{\|K\|^q} : C, D \in \{A, B\} \right\} \leq \delta_{\xi,\perp}, \quad \forall p \in [r], q, s \in [d - r],
\]

(16)

where the $\delta$’s are some small $1/\text{poly}(d)$ values.

The rest of this section is organized as follows. We derive estimations for the $Q$-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 $Q$-matrices in this subsection. The analysis here will be more complicated than the one in Section B.1 since now $	au_i^2$ is no longer close to 0, and we cannot simply approximate $S_A$ and $S_B$ with $(1 + K)^{-1}$. However, the idea is still fairly straightforward. We split all terms into the infinite-width part and the discretization error part. Then, we Taylor expand the corresponding function around the infinite-width part to factor out the first-order error terms. Then, we evaluate and simplify these first-order terms with $\mathbb{E}z^-$ and $\mathbb{E}z^+$.

First, we need the following lemma which gives closed-form formulas for some expectations we will encounter later.

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

\[
\mathbb{E}_{z^-} \left\{ \exp \left( \frac{\langle z^+, z^- \rangle_{K}^2}{N_A N_B} \right) : z_0 = 0, \ldots, z_r = 0, z_{r+1} = T_p \right\} = Z_{p+},
\]

\[
\mathbb{E}_{z^+} \left\{ \exp \left( \frac{\langle z^+, z^- \rangle_{K}^2}{N_A N_B} \right) : z_0 = 0, \ldots, z_r = 0, z_{r+1} = T_p \right\} = Z_{p+},
\]

\[
\mathbb{E}_{z^+} \left\{ \exp \left( \frac{\langle z^+, z^- \rangle_{K}^2}{N_A N_B} \right) : z_0 = 0, \ldots, z_r = 0, z_{r+1} = T_p \right\} = Z_{p+},
\]

\[
\mathbb{E}_{z^+} \left\{ \exp \left( \frac{\langle z^+, z^- \rangle_{K}^2}{N_A N_B} \right) : z_0 = 0, \ldots, z_r = 0, z_{r+1} = T_p \right\} = Z_{p+},
\]

Then, we derive estimations for $S_A$ and $S_B$. There are two types of errors we need to consider. The first one comes from the noises and the second one from the non-diagonals of $K_A K_B$. 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

\[
\delta_{+, z^-} := \frac{\langle K A z^+, B z^- \rangle}{N_A N_B}, \quad \delta_{+, z^+} := \frac{\langle K A z^+, B z^+ \rangle}{N_A N_B},
\]

\[
\delta_{+, \xi, +} := \frac{\langle K A z^+, B \xi_B \rangle}{N_A N_B}, \quad \delta_{+, \xi, -} := \frac{\langle K A_\xi A \xi_B \rangle \text{diag}([T_k]_{k \in [r]} z^+) \rangle}{N_A N_B},
\]

\[
Z_{A,c} := \mathbb{E}_{z^-} E_{+, -}, \quad Z_{B,c} := \mathbb{E}_{z^+} E_{+, +},
\]

\[
\hat{S}_A := E_{+, +}, \quad \hat{S}_B := E_{+, +} + K Z_{B,c}, \quad \hat{S} := \frac{\hat{S}_A}{E_{+, +} + K Z_{A,c}}.
\]
In Stage 2, we have
\[ S_A = \tilde{S}_A + \tilde{S}(1 - \tilde{S})(\delta_{+} + \delta_{+} - \delta_{+}) \pm O \left( d^3 \left( \delta_{\perp} + \delta_{N/S} \right) \right), \]
\[ S_B = \tilde{S}_B + \tilde{S}(1 - \tilde{S})(\delta_{+} + \delta_{+} - \delta_{+}) \pm O \left( d^3 \left( \delta_{\perp} + \delta_{N/S} \right) \right), \]
and
\[ \frac{S_A(x_A^+, x_B^+)}{\exp(f_A^+ + f_B^+)} = \frac{\tilde{S}_A E_{+,+}}{E_{+,+}} + \frac{\tilde{S}_B E_{+,+}}{E_{+,+}} \left( (1 - \tilde{S}) \delta_{+} + \frac{\tilde{S}}{E_{+,+}} \right) \pm O \left( d^3 \left( \delta_{\perp} + \delta_{N/S} \right) \right). \]

Then, we consider the error comes from the non-diagonality of \( K_A^T K_B \).

**Lemma C.4 (Further estimations for \( S \)).** Define
\[ \tilde{E}_0 := \tilde{E}_{+,+} = \tilde{E}_{+,+} = \exp \left( \frac{\| \tilde{K} \|^2}{N_A N_B} \right). \]

In Stage 2, we have
\[ \tilde{S}_A(x_A^+) = \tilde{S}_A \left( 1 + \tilde{S}(1 - \tilde{S}) (\delta_{+} - \delta_{+}) \right) \pm O \left( d^2 \delta_{\perp} \right), \]
\[ \tilde{S}_B(x_B^+) = \tilde{S}_B \left( 1 + \tilde{S}(1 - \tilde{S}) (\delta_{+} - \delta_{+}) \right) \pm O \left( d^2 \delta_{\perp} \right), \]
\[ \tilde{S}_A E_{+,+}/E_{+,+} = \tilde{S}_A E_{+,+}/E_{+,+} (1 - \tilde{S}) \delta_{+,+} - \delta_{+,+} \right) \pm O \left( d^2 \delta_{\perp} \right), \]
\[ \tilde{S}_B E_{+,+}/E_{+,+} = \tilde{S}_B E_{+,+}/E_{+,+} (1 - \tilde{S}) \delta_{+,+} - \delta_{+,+} \right) \pm O \left( d^2 \delta_{\perp} \right). \]

With the above two lemmas in hand, we can now derive estimations for the \( Q \)-matrices.

**Lemma C.5 (Estimations for \( Q_1 \)).** Define \( K_{AB} = K_A^T K_B \) and \( K_{BA} = K_B^T K_A \). In Stage 2, for any \( p \neq q \in [r] \), we have
\[ |Q_1|_{p, p} = 2(1 - \tilde{S}) (T_p - T_p) \pm O \left( d^2 \delta_{\perp} \right), \]
\[ |Q_1|_{p, q} = -\tilde{S} (1 - \tilde{S}) \left( \frac{K_{AB}}{N_A N_B} \right) \frac{\exp(2 - T_p - T_q) - (2 - \tilde{S} + T_p - T_q)}{N_A N_B} \]
\[ - (1 - \tilde{S}) \left( \frac{\exp(T_p + T_q - 1 - \tilde{S})}{N_A N_B} \right) \pm O \left( d^2 \delta_{\perp} \right). \]

In particular, we have
\[ |Q_1|_{p, q} \leq O \left( \frac{\delta_{\perp}}{\sqrt{d}} \right). \]

Note that the diagonal term has a zero-order term, i.e., it is not proportional to some error. That is the signal term. On the other hand, all off-diagonal terms depend on \( |K_{AB}|_{p, q} \) \((p \neq q)\). Recall that the dynamics of \( K_A \) and \( K_B \) can be described using these \( Q \)-matrices. Therefore, for any \( p \neq q \), we can have equations of form \( \frac{\delta_{\perp}}{\sqrt{d}} |K_{AB}|_{p, q} \approx G(|K_{AB}|_{p, q}) \left( |K_{AB}|_{p, q} \right), \) where \( G \) is some complicated matrix. By carefully analyzing \( G \), we can then derive bounds for the off-diagonal entries using Gronwall’s lemma.

Similar things also hold for \( Q_1, \xi \) and \( Q_2 \). The difference here is that for these two matrices, we do not have signal terms.
Lemma C.6 (Estimations for $Q_{1,ξ}$). In Stage 2, for any $p \in [d-r]$ and $q \in [r]$, we have
\[
|Q_{1,ξ}|_{p,q} = -(1 - \tilde{S}) \left(1 + \tilde{S} + (1 - \tilde{S})T_{q}^{1}\right) \frac{\langle [K_{B}]_{p}, [K_{A,ξ}]_{p} \rangle}{N_{A}N_{B}d} \pm O \left(d^{3} (\delta_{AB,⊥} + \delta_{N/S}) \delta_{ξ,⊥} \delta_{N/S}\right),
\]
\[
|Q_{1,ξ,2}|_{p,q} = -(1 - \tilde{S}) \left(1 + \tilde{S} + (1 - \tilde{S})T_{q}^{1}\right) \frac{\langle [K_{A,ξ}]_{p}, [K_{B,ξ}]_{p} \rangle}{N_{A}N_{B}d} \pm O \left(d^{3} (\delta_{AB,⊥} + \delta_{N/S}) \delta_{ξ,⊥} \delta_{N/S}\right).
\]
In particular, we have
\[
\max \{ |Q_{1,ξ}|_{p,q}, |Q_{1,ξ,2}|_{p,q} \} \leq O \left(\delta_{N/S} \delta_{ξ,⊥}\right).
\]
Lemma C.7 (Estimations for $Q_{2}$). In Stage 2, for any $p, q \in [d-r]$, we have
\[
|Q_{2}|_{p,q} = \pm O \left(d^{3} (\delta_{AB,⊥} + \delta_{N/S}) \delta_{ξ,⊥} \delta_{N/S}\right).
\]
Lemma C.8 (Estimations for $Q_{0}$). In Stage 2, we have
\[
Q_{0} = -\sum_{k=1}^{r} \frac{\kappa_{k}^{2}}{||κ||} [Q_{1}]_{k,k} \pm O \left(d^{2} \delta_{AB,⊥}^{2} + \delta_{-} + \kappa_{0} d \delta_{N/S} \delta_{ξ,⊥}\right).
\]
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
\[
\frac{d}{dt} \| [K_{A}]_{p} \|^{2} = 2\sigma_{p}^{2} \frac{\| [K_{A}]_{p} \| \| [K_{B}]_{p} \|}{N_{A}N_{B}d} \left(\langle [K_{A}]_{p}, [K_{B}]_{p} \rangle Q_{1,p} + 2 \frac{\| [K_{A}]_{p} \|^{2}}{N_{A}d} Q_{0} \delta_{ξ}^{2}\right)
\]
\[
\pm O \left(\frac{\sigma_{p}^{2} \kappa_{k}^{2}}{N_{A}N_{B}d} \delta_{AB,⊥}\right)
\]
\[
\frac{d}{dt} \| [K_{A,ξ}]_{p} \|^{2} = 2 \frac{\| [K_{A,ξ}]_{p} \|^{2}}{N_{A}d} Q_{0} \delta_{ξ}^{2} \pm O \left(\frac{\sigma_{p}^{2} \| [K_{A,ξ}]_{p} \|^{2}}{N_{A}N_{B}d} d^{2} \delta_{ξ,⊥}^{2}\right).
\]
The formulas for $\| [K_{B}]_{p} \|$ and $\| [K_{B,ξ}]_{q} \|$ can be obtained by interchanging the roles of $A$ and $B$.

OMITTED PROOFS OF THIS SUBSECTION

Proof of Lemma C.2. First, we compute
\[
E_{z^{-}} \left\{ \exp \left(\frac{z^{+}, z^{-}}{N_{A}N_{B}} \right) \right\} = \prod_{k=1}^{r} E_{z_{k}} \left\{ \exp \left(\frac{\tilde{κ}_{k,ξ}^{2} z_{k}^{+} z_{k}^{-}}{N_{A}N_{B}} \right) \right\} = \prod_{k=1}^{r} \frac{1}{2} \left( \exp \left(\frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right) + \exp \left(- \frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right) \right)
\]
\[
= \prod_{k=1}^{r} \cosh \left(\frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right).
\]
Similarly, we also have
\[
E_{z^{-}} \left\{ \exp \left(\frac{(z^{+}, z^{-})}{N_{A}N_{B}} \right) z_{p}^{-} \right\} = E_{z_{p}} \left\{ \exp \left(\frac{\hat{κ}_{k,ξ}^{2} z_{p}^{+} z_{p}^{-}}{N_{A}N_{B}} \right) \right\} \prod_{k\neq p} E_{z_{k}} \left\{ \exp \left(\frac{\hat{κ}_{k,ξ}^{2} z_{k}^{+} z_{k}^{-}}{N_{A}N_{B}} \right) \right\}.
\]
Each factor in $\prod_{k\neq p}$ is still $\cosh \left(\frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right)$. For the first term, we have
\[
E_{z_{p}} \left\{ \exp \left(\frac{\hat{κ}_{k,ξ}^{2} z_{p}^{+} z_{p}^{-}}{N_{A}N_{B}} \right) z_{p}^{-} \right\} = \frac{1}{2\sqrt{d}} \frac{1}{z_{p}} \left\{ \exp \left(\frac{\hat{κ}_{k,ξ}^{2} z_{p}^{+} / \sqrt{d}}{N_{A}N_{B}} \right) \right\} - \frac{1}{2\sqrt{d}} \frac{1}{z_{p}} \left\{ \exp \left(-\frac{\hat{κ}_{k,ξ}^{2} z_{p}^{+} / \sqrt{d}}{N_{A}N_{B}} \right) z_{p}^{-} \right\}
\]
\[
= \frac{1}{\sqrt{d}} \sinh \left(\frac{\hat{κ}_{k,ξ}^{2} z_{p}^{+} / \sqrt{d}}{N_{A}N_{B}} \right) = \frac{\text{sgn} z_{p}^{+} \sqrt{d}}{\sqrt{d}} \sinh \left(\frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right) = \sinh \left(\frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right) z_{p}^{+}.
\]
Therefore,
\[
E_{z^{-}} \left\{ \exp \left(\frac{(z^{+}, z^{-})}{N_{A}N_{B}} \right) z_{p}^{-} \right\} = z_{p}^{+} \sinh \left(\frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right) \prod_{k\neq p} \cosh \left(\frac{\tilde{κ}_{k}^{2} / d}{N_{A}N_{B}} \right) = Z_{c}T_{p} z_{p}^{+}.
\]
The above calculation, mutatis mutandis, also yields the last identity. □
Proof of Lemma C.3. First, we write

\[
\langle f_A^+, f_B^- \rangle = \frac{\langle K_A z^+ + K_A \xi_A^+, K_B z^- + K_B \xi_B^+ \rangle}{N_A N_B} = \frac{\langle K_A z^+, K_B z^- \rangle}{N_A N_B} + \frac{\langle K_A z^+, K_B \xi_B^+ \rangle}{N_A N_B} + \frac{\langle K_A \xi_A^+, K_B z^- \rangle}{N_A N_B} \pm O \left( d^2 \delta_{\xi,\perp} \delta_{N/S} \right).
\]

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

\[
\exp \left( f_A^+ \cdot f_B^- \right) = \exp \left( 1 + \frac{\langle K_A z^+, K_B \xi_B^+ \rangle}{N_A N_B} + \frac{\langle K_A \xi_A^+, K_B z^- \rangle}{N_A N_B} \right) \pm O \left( d^2 \delta_{\xi,\perp} \delta_{N/S}^2 \right).
\]

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

\[
\exp \left( f_A^+ \cdot f_B^- \right) = \mathbb{E}_{x_B^-} \left( E_{+,-} (1 + \delta_{\xi,+} + \delta_{\xi,-}) \right) \pm O \left( d^2 \delta_{\xi,\perp} \delta_{N/S}^2 \right),
\]

\[
\exp \left( f_A^+ \cdot f_B^+ \right) = \mathbb{E}_{x_B^+} \left( E_{+,-} (1 + \delta_{\xi,+} + \delta_{\xi,-}) \right) \pm O \left( d^2 \delta_{\xi,\perp} \delta_{N/S}^2 \right),
\]

\[
\exp \left( f_A^- \cdot f_B^+ \right) = \mathbb{E}_{x_B^-} \left( E_{+,-} (1 + \delta_{\xi,+} + \delta_{\xi,-}) \right) \pm O \left( d^2 \delta_{\xi,\perp} \delta_{N/S}^2 \right).
\]

To compute \( S_A \) and \( S_B \), we then need to take expectations over the negative examples. Note that by taking expectation over \( \tilde{E}_{\xi,-} \), the term \( E_{+,-}(1+\delta_{\xi,-}) \) becomes 0. Therefore, we have

\[
\mathbb{E}_{x_B^-} \left( E_{+,-}(1+\delta_{\xi,-}) \right) = \mathbb{E}_{x_B^-} \left( \hat{E}_{+,+} (1 \pm O(d \delta_{AB,\perp}) \right) \delta_{\xi,-} + (1 \pm O(d \delta_{AB,\perp}) \delta_{\xi,-} \pm O(d^2 \delta_{\xi,\perp} \delta_{N/S}).
\]

Recall Lemma C.2. Then, we compute

\[
\mathbb{E}_{x_B^-} \left( \hat{E}_{+,+}(1+\delta_{\xi,-}) \right) = \mathbb{E}_{x_B^-} \left( \hat{E}_{+,+}(1+\delta_{\xi,-}) \right) + \mathbb{E}_{x_B^-} \left( \hat{E}_{+,+}(1+\delta_{\xi,-}) \right) \pm O(d^2 \delta_{AB,\perp} \delta_{\xi,\perp} \delta_{N/S}).
\]

Hence,

\[
\mathbb{E}_{x_B^-} \left( \hat{E}_{+,+}(1+\delta_{\xi,-}) \right) = Z_{A,c} + Z_{c} \delta_{\xi,+} + O(d^3 \delta_{\xi,\perp} \delta_{N/S}).
\]

Similarly, we also have

\[
\mathbb{E}_{x_B^-} \left( \hat{E}_{+,+}(1+\delta_{\xi,+}) \right) = Z_{B,c} + Z_{c} \delta_{\xi,+} + O(d^3 \delta_{\xi,\perp} \delta_{N/S}).
\]

Recall that \( \exp \left( f_A^+ \cdot f_B^- \right) = E_{+,-} (1 + \delta_{\xi,+} + \delta_{\xi,-}) \), hence, we have

\[
S_A(x_A^+, x_B^-) = \frac{E_{+,-}(1 + \delta_{\xi,+} + \delta_{\xi,-})}{E_{+,-}(1 + \delta_{\xi,+} + \delta_{\xi,-}) + KZ_{A,c} + KZ_{c}\delta_{\xi,+}} \pm O \left( d^3 \left( \delta_{AB,\perp} + \delta_{N/S} \right) \delta_{\xi,\perp} \delta_{N/S} \right)
\]

\[
= \hat{S}_A \left( 1 - \hat{S}_A \delta_{\xi,+} + \delta_{\xi,+} - (1 - \hat{S}_A) \delta_{\xi,+} + \delta_{\xi,-} \right) + \hat{S}_A(\delta_{\xi,+} + \delta_{\xi,-} \delta_{\xi,+} + \delta_{\xi,-} \delta_{\xi,+}) \pm O \left( d^3 \left( \delta_{AB,\perp} + \delta_{N/S} \right) \delta_{\xi,\perp} \delta_{N/S} \right).
\]
Similarly, we also have
\[ S_B(x_A^+, x_B^+) = \tilde{S}_B + \tilde{S}(1 - \tilde{S}) (\delta_{+,-} + \delta_{+,-} - \delta_{+,-}) \leq O \left( d^3 (\delta_{AB,\perp} + \delta_{NS}) \delta_{\perp} \delta_{NS} \right). \]
Then, we compute
\[
\begin{align*}
S_A(x_A^+, x_B^+) \exp(f_A^+ \cdot f_B) = & \left( \tilde{S}_A + \tilde{S}(1 - \tilde{S}) (\delta_{+,-} + \delta_{+,-} - \delta_{+,-}) \right) E_{+,-} (1 + \delta_{+,-} + \delta_{+,-}) \\
& \pm O \left( d^3 (\delta_{AB,\perp} + \delta_{NS}) \delta_{\perp} \delta_{NS} \right) \\
& = \tilde{S}_A E_{+,-} + \tilde{S}(1 + \delta_{+,-} + \delta_{+,-} - \tilde{S} (\delta_{+,-} + \delta_{+,-} - (1 - \tilde{S}) \delta_{+,-})) \\
& \pm O \left( d^3 (\delta_{AB,\perp} + \delta_{NS}) \delta_{\perp} \delta_{NS} \right) \\
& = \tilde{S}_A E_{+,-} + \tilde{S}
\end{align*}
\]
and
\[
\begin{align*}
S_B(x_A^+, x_B^+) \exp(f_A^+ \cdot f_B) = & \left( \tilde{S}_B + \tilde{S}(1 - \tilde{S}) (\delta_{+,-} + \delta_{+,-} - \delta_{+,-}) \right) E_{+,-} (1 + \delta_{+,-} + \delta_{+,-}) \\
& \pm O \left( d^3 (\delta_{AB,\perp} + \delta_{NS}) \delta_{\perp} \delta_{NS} \right) \\
& = \tilde{S}_B E_{+,-} + \tilde{S}(1 + \delta_{+,-} + \delta_{+,-} - \tilde{S} (\delta_{+,-} + \delta_{+,-} - (1 - \tilde{S}) \delta_{+,-})) \\
& \pm O \left( d^3 (\delta_{AB,\perp} + \delta_{NS}) \delta_{\perp} \delta_{NS} \right).
\end{align*}
\]

\[ \text{Proof of Lemma C.4.} \] We write
\[
\langle K_A z^+, K_B z^- \rangle = \sum_{k=1}^{t} \frac{\langle [K_A]_i, [K_B]_j \rangle}{N_A N_B} z^+_i z^-_j + \sum_{i \neq j} \frac{\langle [K_A]_i, [K_B]_j \rangle}{N_A N_B} z^+_i z^-_j =: I_{+,-} + \hat{\delta}_{+,-}.
\]
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 \( z^\pm \). Also note that \( I_{+,-} \) is bounded by \( O(d\delta_{AB,\perp}) \). Then, we compute
\[ E_{+,-} = \exp(I_{+,-}) \left( 1 + \hat{\delta}_{+,-} + O \left( d^2 \delta_{AB,\perp} \delta_{NS} \right) \right) \quad \text{and} \quad E_{+,+} = \exp(0) \left( 1 + \hat{\delta}_{+,-} \pm O \left( d^2 \delta_{AB,\perp} \delta_{NS} \right) \right). \]
Take expectation over \( z^- \) and we obtain
\[ \mathbb{E} z^- \left\lbrace \exp(I_{+,-}) \delta_{+,-} \right\rbrace = \sum_{i \neq j} \frac{\langle [K_A]_i, [K_B]_j \rangle}{N_A N_B} \mathbb{E} z^- \left\lbrace \exp(I_{+,-}) z^{-}_j \right\rbrace z^+_i, \]
where the second line again comes from Lemma C.2. Hence, we have
\[ \mathbb{E} z^- E_{+,-} = Z_c + Z_c \hat{\delta}_{+,-} \pm O \left( d^2 \delta_{AB,\perp} \delta_{NS} \right). \]
Then, for \( \tilde{S}_A \), we have
\[
\tilde{S}_A(z^+) = \frac{E_{+,+}}{E_{+,+} + K \mathbb{E} z^- E_{+,-}} = \frac{\exp(I_0) \left( 1 + \hat{\delta}_{+,-} \right)}{\exp(I_0) \left( 1 + \hat{\delta}_{+,-} \right) + K Z_c + K Z_c \hat{\delta}_{+,-}} \pm O \left( d^2 \delta_{AB,\perp} \delta_{NS} \right) \\
= \tilde{S}_A \left( 1 - \tilde{S} \hat{\delta}_{+,-} - (1 - \tilde{S}) \hat{\delta}_{+,-} \right) \pm O \left( d^2 \delta_{AB,\perp} \delta_{NS} \right) \\
= \tilde{S}_A \left( 1 + (1 - \tilde{S}) (\hat{\delta}_{+,-} - \hat{\delta}_{+,-}) \right) \pm O \left( d^2 \delta_{AB,\perp} \delta_{NS} \right).
Similarly, we also have
\[
\tilde{S}_B(z^+) = \tilde{S} \left(1 + (1 - \tilde{S})(\delta_{+,+} - \delta_{T+,+})\right) \pm O \left(d^2 \delta_{AB,\perp}^2\right).
\]

Then, we compute
\[
\frac{\tilde{S}_A E_{+-}}{E_{++,+}} = \frac{\tilde{S} (1 + (1 - \tilde{S})(\delta_{++,+} - \delta_{+,T+}))}{\exp(I_{+,+})} \left(1 + \frac{\tilde{S}_A}{\exp(I_0)} (1 - \tilde{S}\delta_{++,+} + (1 - \tilde{S})\delta_{+,T+} + \delta_{+,+})\right) \pm O \left(d^2 \delta_{AB,\perp}^2\right).
\]

Similarly, we also have
\[
\frac{\tilde{S}_B E_{-+,+}}{E_{++,+}} = \frac{\tilde{S} \exp(I_{-+,+})}{\exp(I_0)} (1 - \tilde{S}\delta_{++,+} - (1 - \tilde{S})\delta_{T+,+} + \delta_{+,+}) \pm O \left(d^2 \delta_{AB,\perp}^2\right).
\]

**Proof of Lemma C.5.** Recall that
\[
Q_1 := \mathbb{E} \left\{ (2 - S_A(x_A^+,x_B^+)) - S_B(x_A^+,x_B^+) \right\} z^+(z^+)^T d
\]
\[
- K \mathbb{E} \left\{ \frac{S_A(x_A^+,x_B^+)}{\exp(f_A^+ \cdot f_B^+)} z^-(z^+)^T d \right\} - K \mathbb{E} \left\{ \frac{S_B(x_A^+,x_B^+)}{\exp(f_A^+ \cdot f_B^+)} z^-(z^-)^T d \right\}.
\]

Note that there is no \( \xi \) here other than the ones in the coefficient. As a result, all terms contain \( \delta_+, \xi_+ \)
and alike are 0. Hence, we have
\[
Q_1 := \mathbb{E} \left\{ \left(2 - \tilde{S}_A(x_A^+,x_B^+) - \tilde{S}_B(x_A^+,x_B^+) \right) z^+(z^+)^T d \right\}
\]
\[
- K \mathbb{E} \left\{ \frac{\tilde{S}_A E_{+-}}{E_{++,+}} z^+(z^-)^T d \right\} - K \mathbb{E} \left\{ \frac{\tilde{S}_B E_{-+,+}}{E_{++,+}} z^+(z^-)^T d \right\}
\]
\[
=: T_1(Q_1) + T_2(Q_1) + T_3(Q_1).
\]

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
\[
T_1[(Q_1)_{p,p}] = \mathbb{E} \left\{ 2 - \tilde{S}_A(x_A^+,x_B^+) - \tilde{S}_B(x_A^+,x_B^+) \right\}
\]
\[
= \mathbb{E} \left\{ 2 - \tilde{S}(1 + (1 - \tilde{S})(\delta_{++,+} - \delta_{+,T+})) - \tilde{S}(1 + (1 - \tilde{S})(\delta_{++,+} - \delta_{+,T+})) \right\} \pm O \left(d^2 \delta_{AB,\perp}^2\right)
\]
\[
= 2(1 - \tilde{S}) \pm O \left(d^2 \delta_{AB,\perp}^2\right).
\]

Note that we use the fact that all these \( \delta \)'s have mean 0. Also by Lemma C.4, we have
\[
T_2[(Q_1)_{p,p}] = -K \mathbb{E} \left\{ \left(\frac{\tilde{S} E_{+-}}{E_0} (1 - \tilde{S}\delta_{++,+} - (1 - \tilde{S})\delta_{+,T+} + \delta_{+,+})\right) z_p^+ z_p^+ d \right\} \pm O \left(d^2 \delta_{AB,\perp}^2\right)
\]
\[
= - \frac{\tilde{S} K}{E_0} \mathbb{E} \left\{ E_{+-} z_p^+ z_p^+ d \right\}
\]
\[
- \frac{\tilde{S} K}{E_0} \mathbb{E} \left\{ (E_{+-} - (1 - \tilde{S})\delta_{++,+} + \delta_{+,+}) z_p^+ z_p^+ d \right\} \pm O \left(d^2 \delta_{AB,\perp}^2\right).
\]

Note that \( \delta_{++,+}, \delta_{+,+}, \delta_{+,T+}\) only contain cross terms of form \( z_i^+ z_j^+ \) with \( i \neq j \). Hence, the second term is 0. Meanwhile, by Lemma C.2, the first term is
\[
- \frac{\tilde{S} K}{E_0} \mathbb{E} \left\{ E_{+-} z_p^+ z_p^+ d \right\} = - \frac{\tilde{S} K z_p T_p}{E_0} = -(1 - \tilde{S})T_p.
\]

As a result,
\[
T_2[(Q_1)_{p,p}] = -(1 - \tilde{S})T_p \pm O \left(d^2 \delta_{AB,\perp}^2\right).
\]
Similarly, one can show that \( T_3([Q_1]_{p,p}) = -(1 - \tilde{S})T_p \pm O(d^2 \delta_{AB,\perp}) \) also holds. Thus,

\[
[Q_1]_{p,p} = 2(1 - \tilde{S})(1 - T_p) \pm O(d^2 \delta_{AB,\perp}).
\]

Now, we consider the off-diagonal terms. For notational simplicity, we define \( K_{AB} = K_A^T K_B \).

For any \( p \neq q \), we compute

\[
T_1([Q_1]_{p,q}) = - \mathbb{E} \left\{ \left( \tilde{S}_A(x_A^+ \cdot x_B^+) + \tilde{S}_B(x_A^+ \cdot x_B^+) \right) z_p^+ z_q^+ d \right\}
= - \tilde{S}(1 - \tilde{S}) \mathbb{E} \left\{ (2 \tilde{\delta}_{+,+} - \tilde{\delta}_{+,T+} - \tilde{\delta}_{T+,T}) z_p^+ z_q^+ d \right\} \pm O(d^2 \delta_{AB,\perp})
= - \tilde{S}(1 - \tilde{S}) \sum_{i \neq j} \mathbb{E} \left\{ \frac{K_{AB}}{N_AN_B} (2 - T_i - T_j) \right\} \mathbb{E} \{ z_i^+ z_j^+ z_p^+ z_q^+ d \} \pm O(d^2 \delta_{AB,\perp}).
\]

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

\[
T_1([Q_1]_{p,q}) = - \tilde{S}(1 - \tilde{S}) \frac{K_{AB}}{N_AN_B} + \frac{K_{AB}}{N_AN_B} (2 - T_p - T_q) \pm O(d^2 \delta_{AB,\perp}).
\]

Then, for \( T_2 \), we compute

\[
T_2([Q_1]_{p,q}) = -K \mathbb{E} \left\{ \frac{S_A(x_A^+ \cdot x_B^+) \exp(f_A^+ \cdot f_B^-)}{\exp(f_A^+ \cdot f_B^-)} z_p^- z_q^- d \right\}
= -K \mathbb{E} \left\{ \left( \frac{\tilde{S}E_{+,+}}{E_0} - (1 - \tilde{S}) \tilde{\delta}_{+,T} + \tilde{\delta}_{+,.-} \right) z_p^- z_q^- d \right\} \pm O(d^2 \delta_{AB,\perp})
= -K \mathbb{E} \left\{ \left( \frac{\tilde{S}E_{+,+}}{E_0} - (1 - \tilde{S}) \tilde{\delta}_{+,T} + \tilde{\delta}_{+,.-} \right) z_p^- z_q^- d \right\} \pm O(d^2 \delta_{AB,\perp})
= -K \sum_{i \neq j} \mathbb{E} \left\{ \left( \frac{\tilde{S}E_{+,+}}{E_0} - (1 - \tilde{S}) \tilde{\delta}_{+,T} + \tilde{\delta}_{+,.-} \right) z_i^+ z_j^+ z_p^- z_q^- d \right\} \pm O(d^2 \delta_{AB,\perp}).
\]

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

\[
T_2([Q_1]_{p,q}) = - \frac{\tilde{S}K}{} \frac{[K_{AB}]_{p,q}}{E_0} \left( - \tilde{S} + (1 - \tilde{S}) T_q \right) \mathbb{E} \{ \exp(I_{+,.-}) z_p^- z_q^- \} + \mathbb{E} \{ \exp(I_{+,.-}) z_p^+ z_q^+ z_p^- z_q^- d \}
- \frac{\tilde{S}K}{} \frac{[K_{AB}]_{p,q}}{E_0} \left( - \tilde{S} + (1 - \tilde{S}) T_p \right) \mathbb{E} \{ \exp(I_{+,.-}) z_p^+ z_q^- \} + \frac{d^{-1}}{} \mathbb{E} \{ \exp(I_{+,.-}) \} \pm O(d^2 \delta_{AB,\perp})
- \frac{\tilde{S}K}{} \frac{[K_{AB}]_{p,q}}{E_0} \left( - \tilde{S} + (1 - \tilde{S}) T_q \right) T_p + T_p T_q
- \frac{\tilde{S}K}{} \frac{[K_{AB}]_{q,p}}{E_0} \left( - \tilde{S} + (1 - \tilde{S}) T_p \right) T_p + 1 \pm O(d^2 \delta_{AB,\perp})
= -(1 - \tilde{S}) \frac{[K_{AB}]_{p,q}}{N_AN_B} \tilde{S} T_p (T_q - 1) - (1 - \tilde{S}) \frac{[K_{AB}]_{q,p}}{N_AN_B} \left( - \tilde{S} T_p - (1 - \tilde{S}) T_p^2 + 1 \right) \pm O(d^2 \delta_{AB,\perp}).
\]
Similarly, for $T_3$, we have
\[
T_3((Q_1)_{p,q}) = -K \mathbb{E} \left\{ \frac{S_B(x_A^\top, x_B^\top) \exp(f_A^\top \cdot f_B^\top) z_p^\top z_q^{-} d}{\exp(f_A^\top \cdot f_B^\top)} \right\} \\
= -K \mathbb{E} \left\{ \frac{\tilde{S} E_{-+} (1 - \tilde{S} \delta_{+,+} - (1 - \tilde{S}) \delta_{T+,+} + \tilde{S} \delta_{-,+}) z_p^\top z_q^{-} d}{E_0} \right\} \pm O \left( d^2 \delta_{AB,\perp}^2 \right) \\
= -\frac{\tilde{S} K}{E_0} \sum_{i \neq j} [K_{AB}]_{i,j} \mathbb{E} \left\{ \tilde{E}_{-+} \left( (1 - \tilde{S} - (1 - \tilde{S})T_i) z_p^\top z_j^{-} + z_i^\top z_j^{+} \right) z_p^\top z_q^{-} d \right\} \pm O \left( d^2 \delta_{AB,\perp}^2 \right) \\
= -\frac{\tilde{S} K}{E_0} \sum_{i \neq j} [K_{AB}]_{i,j} \mathbb{E} \left\{ \tilde{E}_{-+} \left( (1 - \tilde{S} - (1 - \tilde{S})T_i) z_p^\top z_j^{-} + z_i^\top z_j^{+} \right) z_p^\top z_q^{-} d \right\} \pm O \left( d^2 \delta_{AB,\perp}^2 \right) \\
= -\frac{\tilde{S} K}{E_0} \sum_{i \neq j} \left[ K_{AB} \right]_{i,j} \mathbb{E} \left\{ \tilde{E}_{-+} \left( (1 - \tilde{S} - (1 - \tilde{S})T_i) z_p^\top z_j^{-} + z_i^\top z_j^{+} \right) z_p^\top z_q^{-} d \right\} \pm O \left( d^2 \delta_{AB,\perp}^2 \right) \\
= -(1 - \tilde{S}) \left[ K_{AB} \right]_{p,q} ST_q(T_p - 1) - (1 - \tilde{S}) \left[ K_{AB} \right]_{q,p} ST_q(T_p - 1) \pm O \left( d^2 \delta_{AB,\perp}^2 \right).
\]
Combining these together we obtain
\[
(Q_1)_{p,q} = -\tilde{S}(1 - \tilde{S}) \left\{ [K_{AB}]_{p,q} + [K_{AB}]_{q,p} (2 - T_p - T_q) \right\} \\
- (1 - \tilde{S}) \left[ K_{AB} \right]_{p,q} \tilde{S} (2T_pT_q - T_p - T_q) \\
- (1 - \tilde{S}) \left[ K_{AB} \right]_{q,p} \left( 2 - \tilde{S}(T_p + T_q) - (1 - \tilde{S})(T_q^2 + T_q^2) \right) \pm O \left( d^2 \delta_{AB,\perp}^2 \right).
\]

\[\square\]

**Proof of Lemma C.6.** We write
\[
Q_{1,\xi_B} := \mathbb{E} \left\{ (2 - S_K(x_A^\top, x_B^\top) - S_B(x_A^\top, x_B^\top)) \xi_B^\top(z^\top)^{-} d \right\} \\
- K \mathbb{E} \left\{ S_A(x_A^\top, x_B^\top) \exp(f_A^\top \cdot f_B^\top) \xi_B^\top(z^\top)^{-} d \right\} - K \mathbb{E} \left\{ S_B(x_A^\top, x_B^\top) \exp(f_A^\top \cdot f_B^\top) \xi_B^\top(z^\top)^{-} d \right\} \\
=: T_1(Q_{1,\xi_B}) + T_2(Q_{1,\xi_B}) + T_3(Q_{1,\xi_B}).
\]
We will use the fact that if some quantity $X$ does not depend on $\xi$, then $\mathbb{E} \{ X \xi \} = 0$ to simplify these terms. For $T_1$, by Lemma C.3, we have
\[
T_1((Q_{1,\xi_B})_{p,q}) = -\mathbb{E} \left\{ (S_K(x_A^\top, x_B^\top) + S_B(x_A^\top, x_B^\top)) \xi_B^\top z_q^+ d \right\} \\
= -\tilde{S} (1 - \tilde{S}) \mathbb{E} \left\{ (2\delta_{\xi+,+} + 2\delta_{\xi+,+} - \delta_{\xi+,+} - \delta_{T+,+}) \xi_B^\top z_q^+ d \right\} \\
\pm O \left( d^3 (\delta_{AB,\perp} + \delta_{N/S}) \delta_{\xi,\perp} \delta_{N/S} \right) \\
= -\tilde{S} (1 - \tilde{S}) \mathbb{E} \left\{ (2\delta_{\xi+,+} + 2\delta_{\xi+,+} - \delta_{T+,+}) \xi_B^\top z_q^+ d \right\} \pm O \left( d^3 (\delta_{AB,\perp} + \delta_{N/S}) \delta_{\xi,\perp} \delta_{N/S} \right).
\]
Recall that
\[
\delta_{\xi,+} = \sum_{i,j} \left[ K_{A}^\top K_{B,\xi} \right]_{i,j} \frac{z_i^+}{N_B} \xi_B^\top \xi_B^j, \quad \delta_{T+,\xi,+} = \sum_{i,j} \left[ K_{A}^\top K_{B,\xi} \right]_{i,j} T_i z_i^+ \xi_B^\top \xi_B^j.
\]
Hence, we can further rewrite $T_1((Q_{1,\xi_B})_{p,q})$ as
\[
T_1((Q_{1,\xi_B})_{p,q}) = -\tilde{S} (1 - \tilde{S}) \sum_{i,j} \left[ K_{A}^\top K_{B,\xi} \right]_{i,j} (2 - T_i) \mathbb{E} \{ z_i^+ \xi_B^\top \xi_B^j \} \\
\pm O \left( d^3 (\delta_{AB,\perp} + \delta_{N/S}) \delta_{\xi,\perp} \delta_{N/S} \right).
\]
Note that the summand is nonzero only if \( i = q \) and \( j = p \). Thus,

\[
T_1 ([Q_{1,ξB}]_{p,q}) = -\bar{S}(1 - \bar{S})(2 - T_q) \frac{[K_A]_{q,p} [K_{B,ξ}]_{j,p}}{N_A N_B d} \pm O \left( d^3 \left( \delta_{AB,⊥} + \delta_{N/S} \right) \delta_{ξ,⊥} \delta_{N/S} \right).
\]

Similarly, we also have

\[
T_2 ([Q_{1,ξB}]_{p,q}) = -\frac{K\bar{S}}{E} E \left\{ \tilde{E}_{p,-} \left( z_q^+ - \bar{S}z_q^+ \right) \delta_{ξ,⊥} \delta_{N/S} \right\},
\]

\[
T_3 ([Q_{1,ξB}]_{p,q}) = -\frac{K\bar{S}}{E} E \left\{ \tilde{E}_{p,+} \left( \tilde{E}_{p,-} \right) \delta_{ξ,⊥} \delta_{N/S} \right\},
\]

Then, we write

\[
E \left\{ \tilde{E}_{p,+} \left( z_q^+ \right) \delta_{ξ,⊥} \delta_{N/S} \right\} = \sum_{i,j} \frac{[K_A^T K_{B,ξ}]_{i,p}}{N_A N_B} \pm O \left( d^3 \left( \delta_{AB,⊥} + \delta_{N/S} \right) \delta_{ξ,⊥} \delta_{N/S} \right).
\]

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

\[
T_2 ([Q_{1,ξB}]_{p,q}) = -\frac{K\bar{S}}{E} E \left\{ \tilde{E}_{p,+} \left( z_q^+ \right) \delta_{ξ,⊥} \delta_{N/S} \right\},
\]

Similarly, for \( T_3 \), we compute

\[
E \left\{ \tilde{E}_{p,+} \left( z_q^+ \right) \delta_{ξ,⊥} \delta_{N/S} \right\} = \sum_{i,j} \frac{[K_A^T K_{B,ξ}]_{j,p}}{N_A N_B} \pm O \left( d^3 \left( \delta_{AB,⊥} + \delta_{N/S} \right) \delta_{ξ,⊥} \delta_{N/S} \right).
\]

Thus,

\[
T_3 ([Q_{1,ξB}]_{p,q}) = -\left( 1 - \bar{S} \right) \left( 1 + \bar{S} \right) \frac{[K_A]_{q,p} [K_{B,ξ}]_{j,p}}{N_A N_B d} \pm O \left( d^3 \left( \delta_{AB,⊥} + \delta_{N/S} \right) \delta_{ξ,⊥} \delta_{N/S} \right).
\]

Combine these together, rearrange terms and we obtain

\[
[Q_{1,ξB}]_{p,q} = -\left( 1 - \bar{S} \right) \left( 1 + \bar{S} \right) \frac{[K_A]_{q,p} [K_{B,ξ}]_{j,p}}{N_A N_B d} \pm O \left( d^3 \left( \delta_{AB,⊥} + \delta_{N/S} \right) \delta_{ξ,⊥} \delta_{N/S} \right).
\]

To obtain the formula for \( Q_{1,ξA} \), it suffices to interchange the roles of \( A \) and \( B \). \( \square \)

**Proof of Lemma C.7.** Recall that

\[
Q_2 := \mathbb{E} \left\{ \left( 2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+) \right) \xi_A^+ (\xi_A^+)^T d \right\} - K \mathbb{E} \left\{ \frac{S_A(x_A^+, x_B^+)}{\exp(f_A^+ \cdot f_B^+)} \xi_A^+ (\xi_A^+)^T d \right\} - K \mathbb{E} \left\{ \frac{S_B(x_A^+, x_B^+)}{\exp(f_A^+ \cdot f_B^+)} \xi_B^+ (\xi_A^+)^T d \right\} =: T_1(Q_2) + T_2(Q_2) + T_3(Q_2).
\]

46
We now estimate each of these three terms. Again, the strategy is to leverage the symmetry of the distribution of $\xi$ to argue that some part of the expectation is 0. For $T_1$, we have

\[
T_1([Q_2]_{p,q}) = -\tilde{S}(1 - \tilde{S}) \mathbb{E}\left\{ (2\delta_{+} + 2\delta_{+,-} - \delta_{+,-} + \delta_{+,+} - \delta_{+,-} + \delta_{+,-}) |\xi_B^+| p |\xi_A^+| q d \right\} + O\left( d^3 \left( \delta_{AB,-1} + \delta_{N/S} \right) \delta_{\xi,1} \delta_{N/S} \right).
\]

Note that none of these $\delta$'s depends on both $\xi_A^+$ and $\xi_B^+$. Therefore, the first term is 0. Similarly, for $T_2$, we have

\[
T_2([Q_2]_{p,q}) = -\frac{K \tilde{S}}{E} \mathbb{E}\left\{ \left( \tilde{E}_{+,-} + \tilde{E}_{+,-} + \tilde{E}_{+,-} + \tilde{E}_{+,-} + \tilde{E}_{+,-} + \tilde{E}_{+,-} - \tilde{E}_{+,-} \right) \right\} |\xi_B^+| p |\xi_A^+| q d \right\} + O\left( d^3 \left( \delta_{AB,-1} + \delta_{N/S} \right) \delta_{\xi,1} \delta_{N/S} \right).
\]

The same is also true for $T_3$. Thus,

\[
[Q_2]_{p,q} = +\frac{O\left( d^3 \left( \delta_{AB,-1} + \delta_{N/S} \right) \delta_{\xi,1} \delta_{N/S} \right)}{}
\]

\[\Box\]

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

\[
Q_0 := - \mathbb{E}_{x_A^+, x_B^+} \left\{ \left( 2 - S_A(x_A^+, x_B^+) - S_B(x_A^+, x_B^+) \right) \langle f_A^+, f_B^- \rangle \right\} + K \mathbb{E}_{x_A^+, x_B^+} \left\{ \frac{S_A(x_A^+, x_B^+)}{E} \exp(\tau_1^t f_A^+ \cdot f_B^-) \langle f_A^+, f_B^- \rangle \right\} + K \mathbb{E}_{x_A^+, x_B^+} \left\{ \frac{S_B(x_A^+, x_B^+)}{E} \exp(\tau_1^t f_A^+ \cdot f_B^-) \langle f_A^+, f_B^- \rangle \right\}.
\]

We write

\[
\langle f_A^+, f_B^- \rangle = \langle K_A z^+, K_B z^- \rangle N_A N_B \pm O \left( \kappa_0 \delta_{N/S} \delta \right)
\]

\[
= \sum_{i,j} \frac{|K_{AB}|_{i,j} z_i^+ z_j^-} {N_A N_B} \pm O \left( \kappa_0 \delta_{N/S} \delta \right).
\]

Therefore, we have

\[
Q_0 = - \sum_{i,j} \frac{|K_{AB}|_{i,j}} {N_A N_B} \langle Q_1 \rangle_{i,j} \pm O \left( \kappa_0 \delta_{N/S} \delta \right)
\]

\[
= - \sum_{k=1}^r \frac{\kappa_0^2} {\kappa_2^2} \langle Q_1 \rangle_{k,k} \pm O \left( d^2 \delta_{AB,-1} \delta \right).
\]

\[\Box\]

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

\[
\frac{d}{dt} \|K_A\|^2 = 2 \frac{\langle K_A \rangle_p \cdot \langle K_B \rangle_p \cdot \langle Q_1 \rangle_{p,q} \cdot \sigma_p^2 + 2 \frac{\langle K_A \rangle_p \cdot \langle K_B \rangle \cdot \langle Q_1 \rangle_{p,q} \cdot \sigma_p^2 + 2 \frac{\langle K_A \rangle_{p,q} \cdot \langle Q_1 \rangle_{p,q} \cdot \sigma_p^2} {N_A N_B d^2} Q_0 \sigma_p^2}{N_A N_B d^2} Q_0 \sigma_p^2 =: \sum_{i=1}^3 T_i \left( \frac{d}{dt} \|K_A\|^2 \right).
\]

By Lemma C.5, we have

\[
T_1 \left( \frac{d}{dt} \|K_A\|^2 \right) = 2 \sum_{k=1}^r \frac{\| K_A \| p \cdot \| K_B \| k}{N_A N_B d^2} \langle K_A \| p \cdot \| K_B \| k \rangle \langle Q_1 \rangle_{k,p} \sigma_p^2 + 2 \frac{\| K_A \|_{p,q} \cdot \| K_B \|_{p,q} \cdot \langle Q_1 \rangle_{p,q} \cdot \sigma_p^2}{N_A N_B d^2} + O \left( \frac{\sigma_p^2 \kappa_0^2}{N_A N_B d^2} \kappa_0 \delta_{AB,-1} \right).
\]

47
By Lemma C.6, we have
\[
T_2 \left( \frac{d}{dt} \|K_A\|_p^2 \right) = 2 \sum_{k=1}^{d-r} \frac{\|K_A\|_p \|K_B\|_k}{N_A N_B d} \langle [K_A]_{p}, [K_B]_{k} \rangle \|Q_{1,1} \|_{k,p} \sigma_p^2
\]
\[
= O \left( \frac{\sigma_p^2 \kappa_p^2}{N_A N_B d} d \delta_{N/S}^2 \delta_{\xi, \perp}^2 \right).
\]

Combine these together and we obtain
\[
\frac{d}{dt} \|K_A\|_p^2 = 2 \sum_{k=1}^{d-r} \frac{\|K_A\|_p \|K_B\|_k}{N_A N_B d} \langle [K_A]_{p}, [K_B]_{k} \rangle \|Q_{1,1} \|_{k,p} \sigma_p^2 + \frac{2 \|K_A\|_p^2}{N_A d} Q_0 \sigma_p^2 
\]
\[
\pm O \left( \frac{\sigma_p^2 \kappa_p^2}{N_A N_B d} \kappa_0 d \delta_{AB, \perp}^2 \right).
\]

To get the formula for \( \frac{d}{dt} \|K_B\|_p^2 \), it suffices to interchange the roles of \( A \) and \( B \). Similarly, for the noises, we write
\[
\frac{d}{dt} \|K_A\|_q^2 = 2 \sum_{i=1}^{3} T_i \left( \frac{d}{dt} \|K_A\|_q^2 \right).
\]
By Lemma C.6, we have
\[
T_1 \left( \frac{d}{dt} \|K_A\|_q^2 \right) = 2 \sum_{k=1}^{d-r} \frac{\|K_A\|_q \|K_B\|_k}{N_A N_B d} \langle [K_A]_{q}, [K_B]_{k} \rangle \|Q_{1,1} \|_{q,k} \sigma_q^2
\]
\[
= O \left( \frac{\sigma_q^2 \|K_A\|_q^2}{N_A N_B d} d \delta_{\xi, \perp}^2 \right).
\]

By Lemma C.7, we have
\[
T_1 \left( \frac{d}{dt} \|K_A\|_q^2 \right) = 2 \sum_{k=1}^{d-r} \frac{\|K_A\|_q \|K_B\|_k}{N_A N_B d} \langle [K_A]_{q}, [K_B]_{k} \rangle \|Q_{2} \|_{k,q} \sigma_q^2
\]
\[
= O \left( \frac{\sigma_q^2 \|K_A\|_q^2}{N_A N_B d} d \delta_{\xi, \perp}^2 \delta_{N/S}^2 \delta_{\xi, \perp} \right).
\]

Combine these together and we get
\[
\frac{d}{dt} \|K_A\|_q^2 = 2 \frac{\|K_A\|_q^2}{N_A d} Q_0 \sigma_q^2 \pm O \left( \frac{\sigma_q^2 \|K_A\|_q^2}{N_A N_B d} d \delta_{\xi, \perp}^2 \right).
\]

\[
\square
\]

C.2 Maintaining the Orthogonality

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

48
Lemma C.10. In Stage 2, for any \( p \neq q \in [r] \), we have

\[
\frac{d}{dt} \left< K_{A|p}, K_{B|q} \right> = [Q_1]_{|p,q} \frac{\kappa_p/\kappa_p \sigma_p^2 + \kappa_q/\kappa_q \sigma_q^2}{N_A N_B d} \\
- \left< K_{A|p}, K_{B|q} \right> \left( [Q_1]_{|p,p} \sigma_p^2 + [Q_1]_{|q,q} \sigma_q^2 \right) \frac{N_A N_B d}{N_A N_B d} \\
+ \left< K_{A|p}, K_{A|q} \right> [Q_1]_{|q,q} \sigma_q^2 \frac{N_A N_B d}{N_A N_B d} + \left< K_{B|p}, K_{B|q} \right> [Q_1]_{|p,p} \sigma_p^2 \frac{N_A N_B d}{N_A N_B d} \\
\pm O \left( \frac{\sigma_{\max}^2}{N_A N_B d} d_{k_0 \delta_{AB,1}} \left( \sqrt{\delta} + \delta_{AB,1} \right) \right).
\]

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

\[
\frac{d}{dt} \left< K_{A|p}, K_{A|q} \right> = [Q_1]_{|p,q} \frac{\kappa_p/\kappa_p \sigma_p^2 + \kappa_q/\kappa_q \sigma_q^2}{N_A N_B d} \\
- \left< K_{A|p}, K_{A|q} \right> \left( [Q_1]_{|p,p} \sigma_p^2 + [Q_1]_{|q,q} \sigma_q^2 \right) \frac{N_A N_B d}{N_A N_B d} \\
+ \left< K_{A|p}, K_{B|q} \right> [Q_1]_{|q,q} \sigma_q^2 \frac{N_A N_B d}{N_A N_B d} + \left< K_{B|p}, K_{B|q} \right> [Q_1]_{|p,p} \sigma_p^2 \frac{N_A N_B d}{N_A N_B d} \\
\pm \left( \frac{\sigma_{\max}^2}{N_A N_B d} d_{k_0 \delta_{AB,1}} \left( \sqrt{\delta} + \delta_{AB,1} \right) \right),
\]

and

\[
\frac{d}{dt} \left< K_{B|p}, K_{B|q} \right> = [Q_1]_{|p,q} \frac{\kappa_p/\kappa_p \sigma_p^2 + \kappa_q/\kappa_q \sigma_q^2}{N_A N_B d} \\
- \left< K_{B|p}, K_{B|q} \right> \left( [Q_1]_{|p,p} \sigma_p^2 + [Q_1]_{|q,q} \sigma_q^2 \right) \frac{N_A N_B d}{N_A N_B d} \\
+ \left< K_{B|p}, K_{A|q} \right> [Q_1]_{|q,q} \sigma_q^2 \frac{N_A N_B d}{N_A N_B d} + \left< K_{A|p}, K_{B|q} \right> [Q_1]_{|p,p} \sigma_p^2 \frac{N_A N_B d}{N_A N_B d} \\
\pm \left( \frac{\sigma_{\max}^2}{N_A N_B d} d_{k_0 \delta_{AB,1}} \left( \sqrt{\delta} + \delta_{AB,1} \right) \right).
\]

Now, we are ready to control the off-diagonality. The proof is similar to the one of Lemma B.15. To provide intuitions, we first consider an idealized case. That is, we assume here the noise-signal ratio is 0 and \( K_A = K_B \), and explain how to control \( \left< K_{A|p}, K_{B|q} \right> \) for \( p \neq q \in [r] \) under these assumptions. In this case, we have

\[
\frac{d}{dt} \left< K_{A|p}, K_{B|q} \right> \approx \left< Q_1 \right>_{|p,q} \frac{\kappa_p/\kappa_p \sigma_p^2 + \kappa_q/\kappa_q \sigma_q^2}{N_A N_B d},
\]

and

\[
\left< Q_1 \right>_{|p,q} \approx - (1 - \tilde{S}) \left< 2 - T_p^2 - T_q^2 + \tilde{S} (2 - T_p - T_q) \right> \frac{K_{AB|p,q}}{N_A N_B d}.
\]

Note that the coefficient is negative. As a result, \( \left< K_{A|p}, K_{B|q} \right> \) will move towards 0.

When \( K_A = K_B \) is not exactly true, the situation is trickier because we need to deal with the middle four terms of the RHS of \( \frac{d}{dt} \left< K_{A|p}, K_{B|q} \right> \). The idea is to view all these off-diagonal errors as a whole and show that their sum is non-increasing, up to some higher-order terms.

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

\[
\delta_{\perp,p,q} := \left< K_{A|p}, K_{B|q} \right>^2 + \left< K_{B|p}, K_{A|q} \right>^2 + \left< K_{A|p}, K_{A|q} \right>^2 + \left< K_{B|p}, K_{B|q} \right>^2.
\]
In Stage 2, we have
\[
\frac{d}{dt} \delta_{\perp, p, q} \leq O \left( \frac{\sigma_{\text{max}}^2}{N_A N_B d} \kappa_0 \left( d^2 \delta_{\perp, \perp}^2 + \delta_{\perp, \perp} \sqrt{\delta_{\perp} + \delta_{\perp}} \right) \frac{\sigma_\perp^2}{\delta_{\perp, p, q}} \right).
\]

Note that the LHS is of order \(\delta_{\perp, \perp}^2\) and for the RHS, the only term that can potentially have the same order is the \(\left( \delta_{\perp} \sqrt{\delta_{\perp, p, q}} \right)\)-related term. We will show later that \(\delta_{\perp} = o(\delta_{\perp, \perp})\), 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
\[
\hat{\delta}_{\perp, p, q} := \left( [K_A]_p, [K_A, \xi]_q \right)^2 + \left( [K_B]_p, [K_A, \xi]_q \right)^2 + \left( [K_B]_p, [K_B, \xi]_q \right)^2.
\]

In Stage 2, we have
\[
\frac{d}{dt} \hat{\delta}_{\perp, p, q} \leq O \left( \frac{\sigma_{\text{max}}^2}{N_A N_B d} d \kappa_0 \delta_{\perp, \perp} \left( \delta_{\perp, \perp} + \sqrt{\delta_{\perp}} \right) \right).
\]

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
\[
\frac{d}{dt} \left( [K_A, \xi]_p, [K_B, \xi]_q \right) = \pm O \left( \frac{\sigma_\perp^2}{N_A N_B d} \frac{d \delta_{\perp, \perp}^2}{\delta_{\perp, \perp}} \right).
\]

**Omitted proof of this subsection**

**Proof of Lemma C.10.** Recall that
\[
\frac{d}{dt} [K_A]_p = \left( I - [K_A]_p [K_A]_p^\top \right) \left( \frac{[K_B]_p [Q_{1, \xi}]_p}{\| [K_A]_p \|} + \frac{[K_B, \xi Q_{1, \xi}]_p}{\| [K_A]_p \|} \right) \frac{\sigma_\perp^2}{N_A N_B d},
\]
\[
\frac{d}{dt} [K_B]_q = \left( I - [K_B]_q [K_B]_q^\top \right) \left( \frac{[K_A]_q [Q_{1, \xi}]_q}{\| [K_B]_q \|} + \frac{[K_A, \xi Q_{1, \xi}]_q}{\| [K_B]_q \|} \right) \frac{\sigma_\perp^2}{N_A N_B d}.
\]

Then, we write
\[
\left( [K_A]_p, \frac{d}{dt} [K_B]_q \right) = \left( [K_A]_p \right)^\top \left( I - [K_B]_q [K_B]_q^\top \right) \left( \frac{[K_A]_q [Q_{1, \xi}]_q}{\| [K_B]_q \|} \right) \frac{\sigma_\perp^2}{N_A N_B d}.
\]

For the second term, by Lemma C.6, we have
\[
\frac{d}{dt} [K_A]_p \left( I - [K_B]_q [K_B]_q^\top \right) \left( \frac{[K_A]_q [Q_{1, \xi}]_q}{\| [K_B]_q \|} \right) \frac{\sigma_\perp^2}{N_A N_B d} = \sum_{k=1}^{d-r} \left( \left( [K_A]_p, [K_A, \xi]_k \right) - \left( [K_A]_p, [K_B]_q \right) \left( [K_B]_q, [K_A, \xi]_k \right) \right) \frac{|| [K_A, \xi]_k || || [Q_{1, \xi}]_k, q ||}{|| [K_B]_q ||} = \pm O \left( d \delta_{\perp, \perp}^2 \frac{\sigma_{\text{max}}^2}{N_A N_B d} \right).
\]

Hence,
\[
\left( [K_A]_p, \frac{d}{dt} [K_B]_q \right) = \left( [K_A]_p \right)^\top \left( I - [K_B]_q [K_B]_q^\top \right) \left( \frac{[K_A]_q [Q_{1, \xi}]_q}{\| [K_B]_q \|} \right) \frac{\sigma_\perp^2}{N_A N_B d} \pm O \left( \frac{\sigma_{\text{max}}^2}{N_A N_B d} d \delta_{\perp, \perp}^2 \frac{\sigma_{\text{max}}^2}{N_A N_B d} \right).
\]
For the first term, we have
\[
\begin{align*}
[K_A]_p^T \left( I - [K_B]_q \left( [K_B]_q \right)^T \right) [K_A Q_1]_q \\
= \sum_{k=1}^r \left( \left( [K_A]_p, [K_A]_k \right) - \left( [K_A]_p, [K_B]_q \right) \left( [K_B]_q, [K_A]_k \right) \right) \frac{[K_A]_k \| [Q_1]_{q,k}}{\| [K_B]_q \|}.
\end{align*}
\]
When \( k \notin \{p, q\} \), we have \( \| [K_A]_p, [K_A]_k \| \leq O(\delta_{AB, \perp}) \), \( \| [K_B]_q, [K_A]_k \| \leq O(\delta_{AB, \perp}) \), and \( \| [K_A]_k \| / \| [K_B]_q \| \leq \kappa_0 \). Meanwhile, by Lemma C.5, we also have \( \| [Q_1]_{q,k} \| \leq O(\delta_{AB, \perp}) \).

Hence,
\[
\begin{align*}
\sum_{k \notin \{p, q\}} \left( \left( [K_A]_p, [K_A]_k \right) - \left( [K_A]_p, [K_B]_q \right) \left( [K_B]_q, [K_A]_k \right) \right) \frac{[K_A]_k \| [Q_1]_{q,k}}{\| [K_B]_q \|} \\
\leq O \left( d_\kappa \delta_{AB, \perp}^2 \right).
\end{align*}
\]
Therefore,
\[
\begin{align*}
\left( 1 - \left[ [K_A]_p, [K_B]_q \right] \left( [K_B]_q, [K_A]_p \right) \right) \frac{\| [K_A]_p \| [Q_1]_{q,p}}{\| [K_B]_q \|} \\
= \left( 1 \pm \delta_{AB, \perp} \right) \frac{\kappa_p [Q_1]_{q,p}}{\kappa_q} \left( 1 \pm \sqrt{\delta_{-}} \right) \\
= \frac{\kappa_p [Q_1]_{q,p}}{\kappa_q} \pm O \left( \kappa_0 \delta_{AB, \perp} \sqrt{\delta_{-}} \right) \pm O \left( \kappa_0 \delta_{AB, \perp}^3 \right).
\end{align*}
\]
For the second term, we have
\[
\begin{align*}
\left( [K_A]_p, [K_A]_q \right) - \left( [K_A]_p, [K_B]_q \right) \left( [K_B]_q, [K_A]_q \right) \frac{\| [K_A]_q \| [Q_1]_{q,q}}{\| [K_B]_q \|} \\
= \left( [K_A]_p, [K_A]_q \right) - \left( [K_A]_p, [K_B]_q \right) \left( 1 \pm \delta_{-} \right) \| [Q_1]_{q,q} \left( 1 \pm \sqrt{\delta_{-}} \right) \\
= \left( [K_A]_p, [K_A]_q \right) - \left( [K_A]_p, [K_B]_q \right) \| [Q_1]_{q,q} \pm O \left( \delta_{-} \sqrt{\delta_{-}} \right).
\end{align*}
\]
Thus,
\[
\begin{align*}
\left( [K_A]_p, \frac{d}{dt} [K_B]_q \right) \left( \frac{\| [K_A]_q \| [Q_1]_{q,q}}{\| [K_B]_q \|} \right)^{-1} \\
= \frac{\kappa_p [Q_1]_{q,p}}{\kappa_q} \pm O \left( \kappa_0 \delta_{AB, \perp} \sqrt{\delta_{-}} \right) \pm O \left( \kappa_0 \delta_{AB, \perp}^3 \right) \\
\quad + \left( \left( [K_A]_p, [K_A]_q \right) - \left( [K_A]_p, [K_B]_q \right) \right) \| [Q_1]_{q,q} \pm O \left( \delta_{-} \sqrt{\delta_{-}} \right) \\
\quad \pm O \left( d_\kappa \delta_{AB, \perp}^2 \right) \pm O \left( d_\delta \delta_{AB, \perp}^2 \right) \\
= \frac{\kappa_p [Q_1]_{q,p}}{\kappa_q} \pm \left( \| [K_A]_p, [K_A]_q \| - \left( [K_A]_p, [K_B]_q \right) \right) \| [Q_1]_{q,q} \pm O \left( \delta_{-} \sqrt{\delta_{-}} \right) \\
\quad \pm O \left( d_\kappa \delta_{AB, \perp} \left( \sqrt{\delta_{-}} + \delta_{AB, \perp} \right) \right).
\end{align*}
\]
Hence,
\[
\left< \frac{d}{dt} [K_A]_p, [K_B]_q \right> = \frac{\kappa_p [Q_1]_{q,p}}{\kappa_q} \frac{\sigma^2_p}{N_A N_B d} + \left( \left< \frac{d}{dt} [K_A]_p, [K_B]_q \right> - \left< K_A, [K_B]_q \right> \right) \frac{[Q_1]_{q,p} \sigma^2_q}{N_A N_B d} \\
\pm O \left( \frac{\sigma^2_{\max}}{N_A N_B d} d_{\kappa_0} \delta_{AB,\perp} \left( \sqrt{\delta_-} + \delta_{AB,\perp} \right) \right).
\]

Interchange the roles of \( A, B, p, q \), replace \( Q_1 \) with \( Q_1^T \), and we obtain
\[
\left< \frac{d}{dt} [K_A]_p, [K_B]_q \right> = \frac{\kappa_p [Q_1]_{q,p}}{\kappa_q} \frac{\sigma^2_p}{N_A N_B d} + \left( \left< [K_B]_p, [K_B]_q \right> - \left< \frac{d}{dt} [K_A]_p, [K_B]_q \right> \right) \frac{[Q_1]_{p,p} \sigma^2_p}{N_A N_B d} \\
\pm O \left( \frac{\sigma^2_{\max}}{N_A N_B d} d_{\kappa_0} \delta_{AB,\perp} \left( \sqrt{\delta_-} + \delta_{AB,\perp} \right) \right).
\]

Combine these together and we complete the proof.

\( \square \)

**Proof of Lemma C.11.** Similar to the proof of the previous lemma, we compute
\[
\left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> = \left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> \left( I - [K_A]_p \left( [K_A]_p \right)^T \right) \left( \frac{[K_B]_q}{\| [K_A]_p \|} + \frac{[K_B]_q Q_1_{q,p}}{\| [K_A]_p \|} \right) \frac{\sigma^2_p}{N_A N_B d} \\
= \left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> \left( I - [K_A]_p \left( [K_A]_p \right)^T \right) \frac{[K_B]_q Q_1_{q,p}}{\| [K_A]_p \|} \| [K_B]_k \| \| Q_1 \|_{p,p} \\
+ \sum_{k \neq \{p,q\}} \left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> \left( I - [K_A]_p \left( [K_A]_p \right)^T \right) \frac{[K_B]_q}{\| [K_A]_p \|} \| [K_B]_k \| \| Q_1 \|_{k,p} \\
= \frac{\kappa_p [Q_1]_{q,p}}{\kappa_q} + \left( \left< [K_B]_p, [K_A]_q \right> - \left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> \right) [Q_1]_{p,p} \pm \left( d_{\kappa_0} \delta_{AB,\perp} \left( \sqrt{\delta_-} + \delta_{AB,\perp} \right) \right).
\]

Therefore,
\[
\left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> = \left< [Q_1]_{q,p} \right> \frac{\kappa_p}{\kappa_q} \frac{\sigma^2_p}{N_A N_B d} + \left( \left< [K_B]_p, [K_A]_q \right> - \left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> \right) \frac{[Q_1]_{q,p} \sigma^2_q}{N_A N_B d} \\
\pm \left( \frac{\sigma^2_{\max}}{N_A N_B d} d_{\kappa_0} \delta_{AB,\perp} \left( \sqrt{\delta_-} + \delta_{AB,\perp} \right) \right).
\]

Interchange the roles of \( p, q \) and we obtain
\[
\left< [K_A]_p, \frac{d}{dt} [K_A]_q \right> = \left< [Q_1]_{q,p} \right> \frac{\kappa_q}{\kappa_p} \frac{\sigma^2_p}{N_A N_B d} + \left( \left< [K_B]_q, [K_A]_p \right> - \left< \frac{d}{dt} [K_A]_p, [K_A]_q \right> \right) \frac{[Q_1]_{q,p} \sigma^2_q}{N_A N_B d} \\
\pm \left( \frac{\sigma^2_{\max}}{N_A N_B d} d_{\kappa_0} \delta_{AB,\perp} \left( \sqrt{\delta_-} + \delta_{AB,\perp} \right) \right).
\]
Combine these together and we get
\[
\frac{d}{dt} \langle [K_{A}]_{p}, [K_{A}]_{q} \rangle = |Q_1|_{q,p} \frac{k_q/\kappa_p \sigma_p^2}{N^2_{A}N^2_{B}d} + |Q_1|_{p,q} \frac{k_p/\kappa_q \sigma_q^2}{N^2_{A}N^2_{B}d} - \langle [K_{A}]_{p}, [K_{A}]_{q} \rangle \left( |Q_1|_{p,q} \sigma_p^2 \frac{1}{N^2_{A}N^2_{B}d} + |Q_1|_{q,q} \sigma_q^2 \frac{1}{N^2_{A}N^2_{B}d} \right) + \langle [K_{A}]_{p}, [K_{B}]_{q} \rangle \frac{|Q_1|_{q,p} \sigma_p^2}{N^2_{A}N^2_{B}d} + \langle [K_{B}]_{p}, [K_{A}]_{q} \rangle \frac{|Q_1|_{p,q} \sigma_q^2}{N^2_{A}N^2_{B}d} \pm \left( \frac{\sigma_{\max}}{N^2_{A}N^2_{B}d} d \kappa_0 \delta_{AB,\perp} \left( \sqrt{\delta_\perp} + \delta_{AB,\perp} \right) \right).
\]
Interchange the roles of $A, B$, replace $Q_1$ with $Q_1^\perp$, and we obtain the formula for $\frac{d}{dt} \langle [K_{B}]_{p}, [K_{B}]_{q} \rangle$.

**Proof of Lemma C.12.** First, we consider the $|Q_1|_{p,q}$-related terms, by Lemma C.5, we have
\[
|Q_1|_{p,q} = -\tilde{S}(1 - \tilde{S}) \frac{[K_{AB}]_{p,q} + [K_{BA}]_{q,p}}{N^2_{A}N^2_{B}d} (2 - T_p - T_q) - (1 - \tilde{S}) \frac{[K_{AB}]_{p,q}}{N^2_{A}N^2_{B}d} \tilde{S} (2T_p T_q - T_p - T_q) - (1 - \tilde{S}) \frac{[K_{BA}]_{q,p}}{N^2_{A}N^2_{B}d} (2 - \tilde{S}(T_p + T_q) - (1 - \tilde{S})(T_p^2 + T_q^2)) \pm O \left( \delta^2 \right) = O \left( \frac{k_p k_q \delta_{AB,\perp}}{N^2_{A}N^2_{B}d} \right).
\]
Hence,
\[
|Q_1|_{p,q} \frac{k_q/\kappa_p \sigma_p^2}{N^2_{A}N^2_{B}d} = O \left( \frac{\sigma_p^2}{N^2_{A}N^2_{B}d \delta_{AB,\perp}} \right) = O \left( \frac{1}{\sqrt{d}} \frac{\sigma_{\max}}{N^2_{A}N^2_{B}d} \delta_{AB,\perp} \right).
\]
The same bound also hold for other $|Q_1|_{p,q}$-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 $Z_1 = \langle [K_{A}]_{p}, [K_{A}]_{q} \rangle$, $Z_2 = \langle [K_{B}]_{p}, [K_{B}]_{q} \rangle$, $Z_3 = \langle [K_{A}]_{p}, [K_{B}]_{q} \rangle$, and $Z_4 = \langle [K_{B}]_{p}, [K_{A}]_{q} \rangle$. Also define $G_p := |Q_1|_{p,q} \sigma_p^2/(N^2_{A}N^2_{B}d)$ and similarly for $G_q$. Then, we can write the results of Lemma C.10 and Lemma C.11 as
\[
\frac{d}{dt} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} = \begin{bmatrix} -G_p - G_q & 0 & G_q & G_p \\ 0 & -G_p - G_q & G_p & 0 \\ G_q & G_p & -G_p - G_q & 0 \\ G_p & 0 & -G_p - G_q & 0 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{bmatrix} \pm O \left( \frac{1}{\sqrt{d}} \frac{\sigma_{\max}}{N^2_{A}N^2_{B}d} \delta_{AB,\perp} \right) \pm O \left( \frac{\sigma_{\max}}{N^2_{A}N^2_{B}d} d \kappa_0 \delta_{AB,\perp} \left( \sqrt{\delta_\perp} + \delta_{AB,\perp} \right) \right).
\]
The eigenvalues of the first matrix are $-2G_p, -2G_q, -2G_p - 2G_q$ and 0. For the first three eigenvalues, note that
\[
G_p = \frac{|Q_1|_{p,q} \sigma_p^2}{N^2_{A}N^2_{B}d} \geq \frac{\sigma_{\max}^2}{\sqrt{d} N^2_{A}N^2_{B}d}.
\]
Hence, the signals will dominate the noises, in particular, the first term on the second line, and push $\|Z\|$ toward 0. Now we consider the eigen-pair $(0, (1, 1, 1, 1))$, for which we will use the actual
form of $[Q_1]_{p,q}$ we obtained in Lemma C.5. We have

$$
\frac{d}{dt} \sum_{i=1}^{4} Z_i = 2 \left( [Q_1]_{p,q} + [Q_1]_{q,p} \right) \kappa_q / \kappa_p \sigma_p^2 + \kappa_p / \kappa_q \sigma_q^2 \frac{N_A N_B d}{N_A N_B d}
+ O \left( \frac{\sigma_{\text{max}}}{N_A N_B d} \kappa_0 \delta_{AB, \perp} \left( \sqrt{\delta_-} + \delta_{AB, \perp} \right) \right).
$$

By Lemma C.5, we have

$$
[Q_1]_{p,q} + [Q_1]_{q,p} = -(1 - \hat{S}) \frac{[K_{AB}]_{p,q} + [K_{BA}]_{p,q}}{N_A N_B d} \left( 2 - T_p^2 - T_q^2 + \hat{S}(2 - T_p - T_q)^2 \right) + O \left( d^2 \delta_{AB, \perp} \right).
$$

Then, we write

$$
[K_{AB}]_{p,q} + [K_{BA}]_{p,q} = \|K_A\| [K_B]_{p,q} + \|K_B\| [K_A]_{p,q} \leq \delta_-. \quad (Z_1 + Z_2) \kappa_p \kappa_q (1 \pm \sqrt{\delta_-}).
$$

Hence,

$$
[Q_1]_{p,q} + [Q_1]_{q,p} = -(1 - \hat{S}) \frac{\kappa_p \kappa_q}{N_A N_B d} \left( Z_1 + Z_2 \right) \left( 2 - T_p^2 - T_q^2 + \hat{S}(2 - T_p - T_q)^2 \right) + O \left( d^2 \delta_{AB, \perp} \right).
$$

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

$$
Z_1 + Z_2 - Z_3 - Z_4 = \left\langle [K_A]_{p,q} - [K_B]_{p,q}, [K_A]_{q,p} - [K_B]_{q,p} \right\rangle \leq \delta_-. \quad (Z_1 + Z_2) \kappa_p \kappa_q (1 \pm \sqrt{\delta_-}).
$$

Therefore,

$$
[Q_1]_{p,q} + [Q_1]_{q,p} = -\frac{1 - \hat{S}}{2} \frac{\kappa_p \kappa_q}{N_A N_B d} \left( 2 - T_p^2 - T_q^2 + \hat{S}(2 - T_p - T_q)^2 \right) \sum_{i=1}^{4} Z_i + O \left( d^2 \delta_{AB, \perp} \right).
$$

Thus,

$$
\frac{d}{dt} \sum_{i=1}^{4} Z_i = -(1 - \hat{S}) \frac{\kappa_q / \kappa_p \sigma_p^2 + \kappa_p / \kappa_q \sigma_q^2}{N_A N_B d} \frac{\kappa_p \kappa_q}{N_A N_B d} \left( 2 - T_p^2 - T_q^2 + \hat{S}(2 - T_p - T_q)^2 \right) \sum_{i=1}^{4} Z_i + O \left( \frac{\sigma_{\text{max}}}{N_A N_B d} \kappa_0 \delta_{AB, \perp} \left( d^2 \delta_{AB, \perp} \right) \right).
$$

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

**Proof of Lemma C.13.** Recall that

$$
\frac{d}{dt} [K_A]_{p,q} = \left( I - [K_A]_{p,q} \right) \frac{[K_B Q_1]_{p} \|K_A\| + [K_B Q_1]_{q} \|K_A\| \|K_A\|}{N_A N_B d} \frac{\sigma_p^2}{N_A N_B d},
$$

$$
\frac{d}{dt} [K_A \xi]_{q} = \left( I - [K_A \xi]_{q} \right) \frac{[K_B Q_1 \xi]_{q} \|K_A\|}{N_A N_B d} \frac{\sigma_\xi^2}{N_A N_B d}.
$$

We now compute $\left\langle \frac{d}{dt} [K_A]_{p,q}, [K_A \xi]_{q} \right\rangle$ and $\left\langle [K_A]_{p,q}, \frac{d}{dt} [K_A \xi]_{q} \right\rangle$ separately. First, we write

$$
\left\langle \frac{d}{dt} [K_A]_{p,q}, [K_A \xi]_{q} \right\rangle = [K_A \xi]_{q} \left( I - [K_A]_{p,q} \right) \frac{[K_B Q_1]_{p} \|K_A\|}{N_A N_B d} \frac{\sigma_p^2}{N_A N_B d}
+ [K_A \xi]_{q} \left( I - [K_A]_{p,q} \right) \frac{[K_B Q_1 \xi]_{q} \|K_A\|}{N_A N_B d} \frac{\sigma_\xi^2}{N_A N_B d}
= : T_1 \left( \left\langle \frac{d}{dt} [K_A]_{p,q}, [K_A \xi]_{q} \right\rangle \right) + T_2 \left( \left\langle \frac{d}{dt} [K_A]_{p,q}, [K_A \xi]_{q} \right\rangle \right).
$$
Then, we compute
\[
T_1 \left( \frac{d}{dt} \langle [K_A]_p, [K_A, \xi]_q \rangle \right) = \left( \langle [K_A, \xi]_q, [K_B]_p \rangle - \langle [K_A]_p, [K_A, \xi]_q \rangle \langle [K_A]_p, [K_B]_p \rangle \right) \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d.
\]
\[
+ \sum_{k \neq p} \left( \langle [K_A, \xi]_q, [K_B]_k \rangle - \langle [K_A]_p, [K_A, \xi]_q \rangle \langle [K_A]_p, [K_B]_k \rangle \right) \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d
\]
\[
= \left( \langle [K_A, \xi]_q, [K_B]_p \rangle - \langle [K_A]_p, [K_A, \xi]_q \rangle \right) \frac{Q_1|_p \|Q_1\|}{N_A N_B} \pm O \left( \frac{\sigma_p^2 \|\xi\|}{N_A N_B}^d \right),
\]
and, by Lemma C.6, we have
\[
T_2 \left( \frac{d}{dt} \langle [K_A]_p, [K_A, \xi]_q \rangle \right) = \sum_{k=1}^{d-r} \langle [K_A]_p, [K_A, \xi]_q \rangle \left( I - [K_A]_p \langle [K_A]_p \rangle \right) \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d
\]
\[
= \pm O \left( \frac{\sigma_p^2 \|\xi\|}{N_A N_B}^d \right).
\]
Combine these together and we get
\[
\langle \frac{d}{dt} [K_A]_p, [K_A, \xi]_q \rangle = \left( \langle [K_A, \xi]_q, [K_B]_p \rangle - \langle [K_A]_p, [K_A, \xi]_q \rangle \right) \frac{Q_1|_p \|Q_1\|}{N_A N_B} \pm O \left( \frac{\sigma_p^2 \|\xi\|}{N_A N_B}^d \right).
\]
Then, we compute \( \langle [K_A]_p, \frac{d}{dt} [K_A, \xi]_q \rangle \). We write
\[
\langle [K_A]_p, \frac{d}{dt} [K_A, \xi]_q \rangle = \sum_{k=1}^{r} \langle [K_A]_p, [K_A, \xi]_q \rangle \left( I - [K_A]_p \langle [K_A]_p \rangle \right) \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d
\]
\[
+ \sum_{k=1}^{d-r} \langle [K_A]_p, [K_A, \xi]_q \rangle \left( I - [K_A]_p \langle [K_A]_p \rangle \right) \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d
\]
\[
=: T_1 \left( \langle [K_A]_p, \frac{d}{dt} [K_A, \xi]_q \rangle \right) + T_2 \left( \langle [K_A]_p, \frac{d}{dt} [K_A, \xi]_q \rangle \right).
\]
For \( T_1 \), when \( k = p \), by Lemma C.6, we have
\[
\langle [K_A]_p \langle I - [K_A, \xi]_q \rangle \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d \rangle
\]
\[
= \left( \langle [K_A]_p, [K_B]_p \rangle - \langle [K_A]_p, [K_A, \xi]_q \rangle \langle [K_A]_p, [K_B]_p \rangle \right) \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d
\]
\[
= - (1 - \tilde{S}) \left( 1 + \tilde{S} + (1 - \tilde{S})T_0^2 \right) (1 \pm O (\delta_+ + \delta_{\xi,\perp}^2)) \frac{\|K_B\|}{N_A N_B} \frac{\|Q_1\|}{\|K_A\|} \pm O \left( \|\xi\| \delta_+ + \delta_{\xi,\perp} + \delta_N \right) \frac{\|\xi\|}{N_A N_B}^d
\]
\[
= - \Theta \left( \frac{\|K_B\|}{N_A N_B} \right) \langle [K_B]_p, [K_A, \xi]_q \rangle \pm O \left( \|\xi\| \delta_+ + \delta_{\xi,\perp} + \delta_N \right) \frac{\|\xi\|}{N_A N_B}^d
\]
When \( k \neq p \), we have
\[
\langle [K_A]_p \langle I - [K_A, \xi]_q \rangle \frac{\|K_B\|}{\|K_A\|} \frac{\|Q_1\|}{\|K_A\|} \sigma_p^2 \frac{\|\xi\|}{N_A N_B}^d \rangle = O \left( \kappa_0 \|\xi\| \delta_+ + \delta_{\xi,\perp} \right).
\]
Hence,

\[ T_1 \left( \langle \tilde{K}_A[p], \frac{d}{dt} \tilde{K}_A[q] \rangle \right) = -\Theta \left( \frac{\|K_B[p]\|^2}{N_AN_B^d} \right) \frac{\sigma^2_\xi}{N_AN_B^d} \langle \tilde{K}_B[p], \tilde{K}_A[q] \rangle \\
+ O \left( \frac{\sigma^2_{\max}}{N_AN_B^d} d^3 \kappa_0 \left( \delta_{AB,\perp} + \delta_{N/S} \right) \delta_{\xi,\perp} \delta_{N/S} \right). \]

Now we consider \( T_2 \). By Lemma C.7, we have

\[ T_2 \left( \langle \tilde{K}_A[p], \frac{d}{dt} \tilde{K}_A[q] \rangle \right) = O \left( \frac{\sigma^2_\xi}{N_AN_B^d} d^4 \delta_{\xi,\perp} \delta_{N/S} \left( \delta_{AB,\perp} + \delta_{N/S} \right) \right). \]

Therefore,

\[ \langle \tilde{K}_A[p], \frac{d}{dt} \tilde{K}_A[q] \rangle = -\Theta \left( \frac{\|K_B[p]\|^2}{N_AN_B^d} \right) \frac{\sigma^2_\xi}{N_AN_B^d} \langle \tilde{K}_B[p], \tilde{K}_A[q] \rangle \\
+ O \left( \frac{\sigma^2_{\max}}{N_AN_B^d} d^3 \kappa_0 \left( \delta_{AB,\perp} + \delta_{N/S} \right) \delta_{\xi,\perp} \delta_{N/S} \right). \]

Thus,

\[ \frac{d}{dt} \langle \tilde{K}_A[p], \tilde{K}_A[q] \rangle = \left( \langle \tilde{K}_A[q], \tilde{K}_B[p] \rangle - \langle \tilde{K}_A[q], \tilde{K}_A[p] \rangle \right) \frac{Q[p,p]\sigma^2_\xi}{N_AN_B^d} \\
- \Theta \left( \frac{\|K_B[p]\|^2}{N_AN_B^d} \right) \frac{\sigma^2_\xi}{N_AN_B^d} \langle \tilde{K}_B[p], \tilde{K}_A[q] \rangle \\
+ O \left( \frac{\sigma^2_{\max}}{N_AN_B^d} d^3 \kappa_0 \left( \delta_{\xi,\perp} + \delta_{\perp} \right) \delta_{N/S} \right). \]

Similarly, one can show that

\[ \frac{d}{dt} \langle \tilde{K}_B[p], \tilde{K}_A[q] \rangle = \left( \langle \tilde{K}_A[q], \tilde{K}_A[p] \rangle - \langle \tilde{K}_A[q], \tilde{K}_B[p] \rangle \right) \frac{Q[p,p]\sigma^2_\xi}{N_AN_B^d} \\
- \Theta \left( \frac{\|K_A[p]\|^2}{N_AN_B^d} \right) \frac{\sigma^2_\xi}{N_AN_B^d} \langle \tilde{K}_A[p], \tilde{K}_A[q] \rangle \\
+ O \left( \frac{\sigma^2_{\max}}{N_AN_B^d} d^3 \kappa_0 \left( \delta_{\xi,\perp} + \delta_{\perp} \right) \delta_{N/S} \right). \]

For notational simplicity, define

\[ X = \langle \tilde{K}_A[p], \tilde{K}_A[q] \rangle, \quad Y = \langle \tilde{K}_B[p], \tilde{K}_A[q] \rangle, \quad C_1 = \frac{Q[p,p]\sigma^2_\xi}{N_AN_B^d}, \quad C_2 = \frac{n^2_p}{N_AN_B^d} \frac{\sigma^2_\xi}{N_AN_B^d}. \]

Then, we can write

\[ \frac{d}{dt} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} -C_1 & C_1 - \Theta(C_2) \\ C_1 - \Theta(C_2) & -C_1 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + O \left( \frac{\sigma^2_{\max}}{N_AN_B^d} d^3 \kappa_0 \left( \delta_{\xi,\perp} + \delta_{\perp} \right) \delta_{N/S} \right). \]

The first matrix is negative semi-definite, whence

\[ \frac{d}{dt} \left( X^2 + Y^2 \right) \leq O \left( \frac{\sigma^2_{\max}}{N_AN_B^d} d^3 \kappa_0 \left( \delta_{\xi,\perp} + \delta_{\perp} \right) \delta_{N/S} \right). \]

Since the roles of \( \tilde{K}_{A,\xi} \) and \( \tilde{K}_{B,\xi} \) are interchangeable, the same bound also holds for \( \langle \tilde{K}_A[p], \tilde{K}_B[q] \rangle \) and \( \langle \tilde{K}_A[p], \tilde{K}_B[q] \rangle \).

\[ \square \]
Proof of Lemma C.14. We write
\[
\left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle = \left[ K_{A,\xi} \right]_p^T \left( I - [K_{B,\xi}]_q \left[ [K_{B,\xi}]_q \right]^T \right) \left( \frac{[K_A Q_{1,\xi}]_{q}}{\| [K_{B,\xi}]_q \|} + \frac{[K_A Q_{2,\xi}]_{q}}{\| [K_{B,\xi}]_q \|} \right) \frac{\sigma^2_{\xi}}{N_A N_B d} \\
= \left[ K_{A,\xi} \right]_p^T \left( I - [K_{B,\xi}]_q \left[ [K_{B,\xi}]_q \right]^T \right) \left( \frac{[K_A Q_{1,\xi}]_{q}}{\| [K_{B,\xi}]_q \|} \frac{\sigma^2_{\xi}}{N_A N_B d} \right) \\
+ \left[ K_{A,\xi} \right]_p^T \left( I - [K_{B,\xi}]_q \left[ [K_{B,\xi}]_q \right]^T \right) \left( \frac{[K_A Q_{2,\xi}]_{q}}{\| [K_{B,\xi}]_q \|} \frac{\sigma^2_{\xi}}{N_A N_B d} \right) \\
= : T_1 \left( \left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle \right) + T_2 \left( \left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle \right).
\]

Then, we compute
\[
T_1 \left( \left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle \right) = \sum_{k=1}^{r} \left( \left\langle [K_{A,\xi}]_p, [K_{A}]_k \right\rangle - \left\langle [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle \left\langle [K_{B,\xi}]_q, [K_{A}]_k \right\rangle \right) \frac{\| [K_A]_k \| \| [Q_{1,\xi,k}]_{q,k} \|}{\| [K_{B,\xi}]_q \|} \frac{\sigma^2_{\xi}}{N_A N_B d}.
\]

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
\[
\frac{\| [K_A]_k \| \| [Q_{1,\xi,k}]_{q,k} \|}{\| [K_{B,\xi}]_q \|} = \pm \frac{\| [K_A]_k \|}{\| [K_{B,\xi}]_q \|} \left( O(1) \frac{\| [K_B]_p, [K_A]_k \|}{N_A N_B d} \pm O \left( d^4 (\delta_{AB,\perp} + \delta_{N,S}) \delta_{\xi,\perp} \delta_{N,S} \right) \right)
\]

Hence,
\[
T_1 \left( \left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle \right) = O \left( \frac{\sigma^2_{\xi}}{N_A N_B d} d^4 \delta_{\xi,\perp} \right).
\]

Now we consider \( T_2 \). By Lemma C.7, We have
\[
T_2 \left( \left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle \right) = \sum_{k=1}^{d-r} \frac{\| [K_{A}]_k \|}{\| [K_{B,\xi}]_q \|} \left( I - [K_{B,\xi}]_q \left[ [K_{B,\xi}]_q \right]^T \right) \left[ K_{A} \right]_k \frac{\| [K_A]_k \| \| [Q_{2,k}]_{q,k} \|}{\| [K_{B,\xi}]_q \|} \frac{\sigma^2_{\xi}}{N_A N_B d}
\]
\[
= \pm \sum_{k=1}^{d-r} O(1) \frac{\| [Q_{2,k}]_{q,k} \|}{\| [K_{B,\xi}]_q \|} \frac{\sigma^2_{\xi}}{N_A N_B d}
\]
\[
= \pm O \left( \frac{\sigma^2_{\max}}{N_A N_B d} d^4 (\delta_{AB,\perp} + \delta_{N,S}) \delta_{\xi,\perp} \delta_{N,S} \right).
\]

Combine this together and we obtain
\[
\left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle = \pm O \left( \frac{\sigma^2_{\xi}}{N_A N_B d} d^4 \delta_{\xi,\perp} \right).
\]

Similarly, one can derive the same bound for \( \left\langle \frac{d}{dt} [K_{A,\xi}]_p, [K_{B,\xi}]_q \right\rangle \) and complete the proof. \( \square \)

C.3 MAINTAINING \( K_A \approx K_B \)

In this subsection, we show that \( \left\langle K_A \right\rangle_{p} = \| K_A \| \approx \| K_B \| \approx \left\langle K_B \right\rangle_{p} \) and also \( \| K_{A,\xi} \| \approx \| K_{B,\xi} \| \) throughout Stage 2. The proofs are similar to the corresponding ones in Stage 1.
Lemma C.15. In Stage 2, we have
\[ \frac{d}{dt} \langle [K_A]_p, [K_B]_p \rangle = \Omega \left( \frac{\sigma_p^2 [Q_1]_{p,p}}{N_A N_B d} \right) \left( 1 - \langle [K_A]_p, [K_B]_p \rangle \right) + O \left( \frac{\sigma_{d_{\max}}^2 d_{\max}^2 \delta_{AA, \perp}}{N_A N_B d} \right). \]

Lemma C.16. Define \( \rho_{A/B,p} := \| [K_A]_p \|^2 / \| [K_B]_p \|^2 \) and \( \rho_{B/A,p} := \| [K_B]_p \|^2 / \| [K_A]_p \|^2 \).
In Stage 2, we have
\[ \frac{d}{dt} (\rho_{A/B,p} + \rho_{B/A,p}) = \frac{2 [Q_1]_{p,p} \sigma_p^2}{N_A N_B d} \left( [K_A]_p - \langle [K_A]_p \rangle \right)^\top \left( I - \langle [K_A]_p \rangle \right) [K_B]_p \| [K_A]_p \| \| [K_B]_p \| \frac{\sigma_p^2}{N_A N_B d} + O \left( \frac{\sigma_{d_{\max}}^2 d_{\max}^2 \delta_{AB, \perp}}{N_A N_B d} \right). \]

Lemma C.17. For any \( p, q \in [d - r] \), define \( \rho_{\xi, A/A, \xi/q} = \| [K_A, \xi]_p \|^2 / \| [K_A, \xi]_q \|^2 \) and \( \rho_{\xi, A/B, \xi/q} = \| [K_A, \xi]_p \|^2 / \| [K_B, \xi]_q \|^2 \). In Stage 2, we have
\[ \frac{d}{dt} \rho_{\xi, A/A, \xi/q} = \pm O \left( \frac{\sigma_{\xi}^2 \| [K_A, \xi]_p \| \| [K_A, \xi]_q \| \delta_{\xi, \perp}}{N_A N_B d} \right), \]
\[ \frac{d}{dt} \left( \rho_{\xi, A/B, \xi/q} + \rho_{\xi, A/A, \xi/q} \right) = \pm O \left( \frac{\sigma_{\xi}^2 \| [K_A, \xi]_p \| \| [K_A, \xi]_q \| \delta_{\xi, \perp} + \delta_{\xi, \perp}^2} {N_A N_B d} \right). \]

OMITTED PROOF OF THIS SUBSECTION

Proof of Lemma C.15. First, we write
\[ \left\langle \frac{d}{dt} [K_A]_p, [K_B]_p \right\rangle = [K_B]_p^\top \left( I - [K_A]_p \right) [K_A]_p \| [K_A]_p \| \| [K_B]_p \| \frac{\sigma_p^2}{N_A N_B d} + \sum_{k \neq p} [K_B]_p^\top \left( I - [K_A]_p \right) [K_A]_p \| [K_B]_p \| \| [K_A]_p \| \| [K_B]_k \| \| [Q_1]_{p,p} \| \sigma_p^2 \frac{\sigma_{\xi}^2 \| [K_A, \xi]_p \| \| [K_A, \xi]_q \| \delta_{\xi, \perp}} {N_A N_B d} \]
\[ = \left( 1 - \langle [K_A]_p, [K_B]_p \rangle \right)^2 \| [K_B]_p \| \| [Q_1]_{p,p} \| \sigma_p^2 \frac{\sigma_{\xi}^2 \| [K_A, \xi]_p \| \| [K_A, \xi]_q \| \delta_{\xi, \perp}} {N_A N_B d} + O \left( \frac{\sigma_{d_{\max}}^2 d_{\max}^2 \delta_{A B, \perp}} {N_A N_B d} \right). \]

For T2, we compute
\[ T_2 \left( \left\langle \frac{d}{dt} [K_A]_p, [K_B]_p \right\rangle \right) = \sum_{k=1}^{d-r} [K_B]_p^\top \left( I - [K_A]_p \right) [K_A]_p \| [K_B]_p \| \| [Q_1]_{k,p} \| \sigma_p^2 \frac{\sigma_{\xi}^2 \| [K_A, \xi]_p \| \| [K_A, \xi]_q \| \delta_{\xi, \perp}} {N_A N_B d} \]
\[ = \pm O \left( \frac{\sigma_{d_{\max}}^2 d_{\max}^2 \delta_{\xi, \perp}} {N_A N_B d} \right). \]

Therefore,
\[ \left\langle \frac{d}{dt} [K_A]_p, [K_B]_p \right\rangle = \left( 1 - \langle [K_A]_p, [K_B]_p \rangle \right)^2 \| [K_B]_p \| \| [Q_1]_{p,p} \| \sigma_p^2 \frac{\sigma_{\xi}^2 \| [K_A, \xi]_p \| \| [K_A, \xi]_q \| \delta_{\xi, \perp}} {N_A N_B d} + O \left( \frac{\sigma_{d_{\max}}^2 d_{\max}^2 \delta_{A B, \perp}} {N_A N_B d} \right). \]
Then, by symmetry, we have
\[
\frac{d}{dt} \langle \sigma (K_A_p, K_B_p) \rangle = \left( 1 - \langle K_A_p, K_B_p \rangle^2 \right) [Q1]_{p,p} \left( \frac{||K_B_p||}{||K_A_p||} + \frac{||K_A_p||}{||K_B_p||} \right) \frac{\sigma_p^2}{N_A N_B d} \\
+ O \left( \frac{\sigma_p^2}{N_A N_B d} \right) - d_{\xi}^2 \delta_{AB,\perp}^2
\]
\[
= \Omega \left( \frac{\sigma_p^2}{N_A N_B d} \right) \left( 1 - \langle K_A_p, K_B_p \rangle \right) + O \left( \frac{\sigma_p^2}{N_A N_B d} \right) - d_{\xi}^2 \delta_{AB,\perp}^2
\].

Proof of Lemma C.16. Similar to Stage 1, we define \( \rho_{A/B,p} = \|K_A_p\|^2 / \|K_B_p\|^2 \). By Lemma A.5, we have
\[
\frac{d}{dt} \rho_{A/B,p} = \frac{d}{dt} \left( \frac{\|K_A_p\|^2}{\|K_B_p\|^2} - \rho_{A/B,p} \right) \frac{d}{dt} \left( \|K_B_p\|^2 \right) \\
= 2 \frac{\langle [K_A_p, K_B Q_1]_p, [K_B Q_1]_p \rangle}{N_A N_B d \|K_B_p\|^2} \|K_B_p\|^2 + 2 \frac{\langle [K_A_p, K_B Q_1]_p, [K_B Q_1]_p \rangle}{N_A N_B d \|K_B_p\|^2} \|K_B_p\|^2 \\
- \rho_{A/B,p} \left( \frac{2 \langle [K_B_p, [K_A_p, Q_1]_p]_p \rangle}{N_A N_B d \|K_B_p\|^2} \|K_B_p\|^2 + 2 \frac{\langle [K_B_p, [K_A_p, Q_1]_p]_p \rangle}{N_A N_B d \|K_B_p\|^2} \|K_B_p\|^2 \right) \\
+ 2 \frac{Q_0 \sigma^2}{N_A d} \right) - \frac{1}{N_B^2 d} \right) \rho_{A/B,p}.
\]
Then, by Lemma C.5 and Lemma C.6, we have
\[
\langle [K_A_p, [K_B Q_1]_p] = \langle [K_A_p, [K_B_p] Q_1]_p \rangle + O \left( d_{\xi}^2 \delta_{N/S} \delta_{\perp}^2 \right), \\
\langle [K_B_p, [K_A Q_1]_p] = \langle [K_B_p, [K_A_p] Q_1]_p \rangle + O \left( d_{\xi}^2 \delta_{N/S} \delta_{\perp}^2 \right), \\
\langle [K_A_p, [K_B Q_1]_p] \rangle = \langle [K_A_p, [K_B_p] Q_1]_p \rangle + O \left( d_{\xi}^2 \delta_{N/S} \delta_{\perp}^2 \right), \\
\langle [K_B_p, [K_A Q_1]_p] \rangle = \langle [K_B_p, [K_A_p] Q_1]_p \rangle + O \left( d_{\xi}^2 \delta_{N/S} \delta_{\perp}^2 \right).
\]

Thus,
\[
\frac{d}{dt} \rho_{A/B,p} = 2 \frac{Q_1 \sigma^2}{N_A N_B d} \|K_B_p\|^2 \left( 1 - \rho_{A/B,p} \right) \\
+ 2Q_0 \sigma^2 \left( \frac{1}{N_A^2 d} - \frac{1}{N_B^2 d} \right) \rho_{A/B,p} + O \left( \frac{\sigma_p^2}{N_A N_B d} \right) - d_{\xi}^2 \delta_{AB,\perp}^2
\].

Interchange the roles of \( A, B \) and we get
\[
\frac{d}{dt} \rho_{B/A,p} = 2 \frac{Q_1 \sigma^2}{N_A N_B d} \|K_A_p\|^2 \left( 1 - \rho_{B/A,p} \right) \\
+ 2Q_0 \sigma^2 \left( \frac{1}{N_B^2 d} - \frac{1}{N_A^2 d} \right) \rho_{B/A,p} + O \left( \frac{\sigma_p^2}{N_A N_B d} \right) - d_{\xi}^2 \delta_{AB,\perp}^2
\].
Hence,
\[
\frac{d}{dt} (\rho_{A/B,p} + \rho_{B/A,p}) = 2 \frac{[Q]_{p,p} \sigma_p^2}{N_A N_B d} \left( \langle [K_A]_p, [K_B]_p \rangle \left( \frac{1 - \rho_{A/B,p}}{\| [K_B]_p \|^2} + \frac{1 - \rho_{B/A,p}}{\| [K_A]_p \|^2} \right) \right.
\]
\[
+ 2 Q_0 \sigma_p^2 \left( \frac{1}{N_A^2 d} - \frac{1}{N_B^2 d} \right) \left( \rho_{A/B,p} - \rho_{B/A,p} \right) \pm O \left( \frac{\sigma_{\text{max}}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right)
\]
\[
= 2 \frac{[Q]_{p,p} \sigma_p^2}{N_A N_B d} \left( \langle [K_A]_p, [K_B]_p \rangle (2 - \rho_{A/B,p} - \rho_{B/A,p}) \right)
\]
\[
\pm O \left( \frac{\sigma_{\text{max}}^2}{N_A N_B d} (d \delta_{\xi,\bot} + \delta^2) \right).
\]

Proof of Lemma C.17. For notational simplicity, we drop the subscript \( \xi \) in the proof. Recall from Corollary C.9 that
\[
\frac{d}{dt} \| [K_{A,\xi}]_q \|^2 = 2 \frac{\| [K_{A,\xi}]_q \|^2}{N_A^2 d} Q_0 \sigma_q^2 \pm O \left( \frac{\sigma_{\xi}^2 \| [K_{A,\xi}]_q \|^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right).
\]

Hence, for any \( p, q \in [d-r] \), we have
\[
\frac{d}{dt} \rho_{A,p/q} = \frac{d}{dt} \frac{\| [K_{A,\xi}]_q \|^2}{\| [K_{A,\xi}]_q \|^2} \rho_{A,p/q} = \frac{\| [K_{A,\xi}]_q \|^2}{\| [K_{A,\xi}]_q \|^2} \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right)
\]
\[
\pm O \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right).
\]
Similarly, we have
\[
\frac{d}{dt} \rho_{B,p/q} = \frac{d}{dt} \frac{\| [K_{B,\xi}]_q \|^2}{\| [K_{B,\xi}]_q \|^2} \rho_{B,p/q} = \frac{\| [K_{B,\xi}]_q \|^2}{\| [K_{B,\xi}]_q \|^2} \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right)
\]
\[
\pm O \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right).
\]
By symmetry, we also have
\[
\frac{d}{dt} \rho_{B,q/p} = \frac{d}{dt} \frac{\| [K_{B,\xi}]_q \|^2}{\| [K_{B,\xi}]_q \|^2} \rho_{B,q/p} = \frac{\| [K_{B,\xi}]_q \|^2}{\| [K_{B,\xi}]_q \|^2} \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right)
\]
\[
\pm O \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right).
\]
Hence,
\[
\frac{d}{dt} (\rho_{A,B,p/q} + \rho_{B,A,q/p}) = \frac{2}{N_A^2 d} - \frac{2}{N_B^2 d} \left( \rho_{A,B,p/q} - \rho_{B,A,q/p} \right) \pm O \left( \frac{\sigma_{\xi}^2}{N_A N_B d} \delta_{\xi,\bot}^2 \right)
\]
\[
= \pm O \left( \frac{\sigma_{\xi}^2}{N_A N_B d} (d \delta_{\xi,\bot} + \delta^2) \right).
\]
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 \([||K_A||, \ldots, ||K_A||_{k \in [r]}\) be the smallest among all \([||K_A||, \ldots, ||K_A||_{k \in [r]}\). For any \(q \in [r]\), in Stage 2, we have 
\[
\frac{d}{dt} \frac{||K_A||_q^2}{||K_A||_p^2} \leq O \left( \frac{\sigma_q^2}{N_A N_B d} \left( \delta_+^2 + \delta_{AB, \perp}^2 + \delta_- + d^2 \delta_{NS, \perp} \right) \right) \frac{||K_A||_q^2}{||K_A||_p^2}.
\]

**Proof.** Recall from Corollary C.9 that
\[
\frac{d}{dt} ||K_A||_p^2 = \frac{2\sigma_p^2 ||K_A|| \langle ||K_B|| \rangle}{N_A N_B d} \langle [K_A]_p, [K_B]_p \rangle [Q_1]_{p,p} + 2 \frac{||K_A||_p^2}{N_A d} Q_0 \sigma_p^2 \pm O \left( \frac{\sigma_p^2 ||K_A||_q^2}{N_A N_B d} \delta_{AB, \perp}^2 \right).
\]

Since the condition number of \(K_A\) is bounded by \(\sqrt{d}\), it suffices to consider the smallest \(||K_A||\), for which we have
\[
\frac{d}{dt} ||K_A||_p^2 = \frac{2\sigma_p^2 ||K_A|| \langle [Q_1]_{p,p} \rangle \langle [K_A]_p \rangle}{N_A N_B d} + \frac{2\sigma_p^2}{N_A d} Q_0 ||[K_A]_p||^2 \pm O \left( \frac{\sigma_p^2 ||K_A||_q^2}{N_A N_B d} \delta_{AB, \perp}^2 \right)
\]
\[
= \frac{2\sigma_p^2}{N_A N_B d} \langle [Q_1]_{p,p} \rangle \langle [K_A]_p \rangle - \frac{2\sigma_p^2}{N_A d} \left( \sum_{k=1}^{r} \kappa_k^2 \langle [Q_1]_{k,k} \rangle \right) ||[K_A]_p||^2 \pm O \left( \frac{\sigma_p^2 ||K_A||_q^2}{N_A N_B d} \left( d^2 \delta_{AB, \perp}^2 + d^2 \delta_{NS, \perp} + \delta_- \right) \right),
\]
where the last line comes from Lemma C.8. By Lemma B.4, \([Q_1]_{p,p}\) is negative correlated with \(\kappa_p^2\).

As a result, we have
\[
\frac{d}{dt} ||K_A||_p^2 \geq -O \left( \frac{\sigma_p^2 ||K_A||_q^2}{N_A N_B d} \left( d^2 \delta_{AB, \perp}^2 + d^2 \delta_{NS, \perp} + \delta_- \right) \right).
\]

For \(K_A, q\), we simply have
\[
\frac{d}{dt} ||K_A||_q^2 \leq O \left( \frac{\sigma_q^2}{N_A N_B d} \left( \delta_+^2 + \delta_{AB, \perp}^2 + \delta_- + d^2 \delta_{NS, \perp} \right) \right) ||K_A||_q^2.
\]

Thus,
\[
\frac{d}{dt} \frac{||K_A||_q^2}{||K_A||_p^2} = \frac{d}{dt} \frac{||K_A||_q^2}{||K_A||_p^2} - \frac{||K_A||_q^2}{||K_A||_p^2} \frac{d}{dt} \frac{||K_A||_p^2}{||K_A||_p^2} \leq O \left( \frac{\sigma_q^2}{N_A N_B d} \left( \delta_+^2 + \delta_{AB, \perp}^2 + \delta_- + d^2 \delta_{NS, \perp} \right) \right) ||K_A||_q^2.
\]
\[
\square
\]
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 $\|K_A\|_p$ is the largest and $\|K_A\|_q$ the smallest among all $\{\|K_A\|_{k}\}_{k \in [r]}$. In Stage 2, we have

$$\frac{d}{dt} \|K_A\|_p^2 \leq -\frac{4(1-\tilde{S})\sigma_{\min}^2}{N_A N_B d} (T_p - T_q) \|K_A\|_p^2$$

$$+ O \left( \frac{\sigma_p^2}{N_A N_B d} (d^2 \delta_{AB, \perp} + \delta_- + d^2 \delta_{N/S \delta_{\xi, \perp}}) \right).$$

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

$$\frac{d}{dt} \|K_A\|_p^2 = \frac{2\sigma_p^2}{N A N_B d} [Q_1]_{p,p} \|K_A\|_p^2 + \frac{2\sigma_p^2}{N A d} Q_0 \|K_A\|_p^2$$

$$+ O \left( \frac{\sigma_p^2}{N_A N_B d} (\kappa_0 d \delta_{AB, \perp} + \delta_-) \right)$$

$$= \frac{4(1-\tilde{S})\sigma_p^2}{N_A N_B d} (1-T_p) \|K_A\|_p^2 + \frac{4(1-\tilde{S})\sigma_p^2}{N_A d} \left( \sum_{k=1}^r \kappa_k^2 (1-T_p) \right) \|K_A\|_p^2$$

$$+ O \left( \frac{\sigma_p^2}{N_A N_B d} (d^2 \delta_{AB, \perp} + \delta_- + d^2 \delta_{N/S \delta_{\xi, \perp}}) \right)$$

$$= -\frac{4(1-\tilde{S})\sigma_p^2}{N_A N_B d} (T_p - \tilde{T}) \|K_A\|_p^2 + O \left( \frac{\sigma_p^2}{N_A N_B d} (d^2 \delta_{AB, \perp} + \delta_- + d^2 \delta_{N/S \delta_{\xi, \perp}}) \right),$$

where $\tilde{T} = \sum_{k=1}^r \frac{\kappa_k^2}{\kappa_0} T_p$. Then, we compute

$$\frac{d}{dt} \|K_A\|_q^2 = \frac{d}{dt} \|K_A\|_p^2 - \frac{\|K_A\|_p^2}{\|K_A\|_q^2} \frac{d}{dt} \|K_A\|_q^2$$

$$= -\frac{4(1-\tilde{S})}{N_A N_B d} \left( \sigma_p^2 (T_p - \tilde{T}) - \sigma_q^2 (T_q - \tilde{T}) \right) \|K_A\|_p^2$$

$$+ O \left( \frac{\sigma_p^2}{N_A N_B d} (d^2 \delta_{AB, \perp} + \delta_- + d^2 \delta_{N/S \delta_{\xi, \perp}}) \right).$$

Since $\|K_A\|_p$ is the largest, $\|K_A\|_q$ is the smallest, $T_p$ is positively correlated with $\|K_A\|_p$, and $\tilde{T}$ is a weighted average of $T_p$, we have

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

Thus,

$$\frac{d}{dt} \|K_A\|_q^2 \leq -\frac{4(1-\tilde{S})\sigma_{\min}^2}{N_A N_B d} (T_p - T_q) \|K_A\|_p^2$$

$$+ O \left( \frac{\sigma_p^2}{N_A N_B d} (d^2 \delta_{AB, \perp} + \delta_- + d^2 \delta_{N/S \delta_{\xi, \perp}}) \right).$$
Proof of Corollary C.20. Recall that
\[ T_p = \tanh \left( \frac{k_p^2}{NKN_Bd} \right) = \tanh \left( \frac{\|K_A\|_F^2}{\|K_A\|_F^2} \left( 1 \pm O(\delta_+ + \delta_{N/S}) \right) \right). \]
Since \( \kappa_0 \leq \sqrt{d} \), we have \( \|K_A\|_F^2 \leq \frac{1}{2} \). Note that \( \tanh'(z) = 1 - \tanh^2(z) = \Omega(1) \) for any \( z \leq 1.1/2 \). Therefore,
\[ T_p - T_q \geq \Omega \left( \frac{\|K_A\|_F^2 - \|K_A\|_q^2}{\|K_A\|_F^2} \right) \pm O(\delta_+ + \delta_{N/S}) \geq \Omega \left( \frac{1}{d} \right). \]
Then, by Lemma C.19, we have
\[ \frac{d}{dt} \left[ \|K_A\|_p^2 \right] \leq -\frac{4(1 - \delta)}{NKN_Bd} \frac{\|K_A\|_F^2}{\|K_A\|_p^2} (T_p - T_q) \frac{\|K_A\|_p^2}{\|K_A\|_q^2} \pm O \left( \frac{\sigma_p^2}{NKN_Bd} (d^2\delta^2_{AB,\perp} + \delta_+ + d^2\delta_{N/S}\delta_{\xi,\perp}) \kappa_0^2 \right) \]
\[ \leq -\Omega \left( \frac{\sigma_p^2}{NKN_Bd} \right) \frac{\|K_A\|_p^2}{\|K_A\|_q^2}. \]
By the proof of Lemma C.19, the largest \( \|K_A\|_p^2 \) is non-increasing. Hence, \( NKN_Bd \) is upper bounded by some \( \text{poly}(d) \). Thus, it takes at most \( \text{poly}(d) \) for \( \|K_A\|_p^2 / \|K_A\|_q^2 \) 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
\[ \frac{d}{dt} \delta_{\perp,p,q} \leq O \left( \frac{\sigma^2_{\max}}{NKN_Bd} \kappa_0 (d^2\delta^2_{AB,\perp} + \delta_+ \sqrt{\delta_+ + \delta_-}) \right), \]
\[ \frac{d}{dt} \delta_{\xi,\perp,p,q} \leq O \left( \frac{\sigma^2_{\max}}{NKN_Bd} d\kappa_0 \delta^2_{\xi,\perp} \right), \]
\[ \frac{d}{dt} \left( \|K_{A,q}\|_p, \|K_{B,q}\|_q \right) = \pm O \left( \frac{\sigma^2_{\xi}}{NKN_Bd} \right). \]
Recall from Lemma C.15, Lemma C.16, and Lemma C.17 that
\[ \frac{d}{dt} \left( \langle K_{A,p}, K_{B,p} \rangle \right) = \Omega \left( \frac{\sigma^2_p (Q)_p}{NKN_Bd} \right) \left( 1 - \langle K_{A,p}, K_{B,p} \rangle \right) \pm O \left( \frac{\sigma^2_{\max}}{NKN_Bd} d\kappa_0 \delta^2_{AB,\perp} \right), \]
\[ \frac{d}{dt} \left( \rho_{A/B,p} + \rho_{B/A,p} \right) \leq O \left( \frac{\sigma^2_{\max}}{NKN_Bd} (d\kappa_0 \delta^2_{AB,\perp} + \delta_+^2) \right), \]
\[ \frac{d}{dt} \rho_{\xi,A/p} \leq O \left( \frac{\sigma^2_{\xi}}{NKN_Bd} \right), \]
\[ \frac{d}{dt} \left( \rho_{\xi,A/B,p/q} + \rho_{\xi,B/A,q/p} \right) \leq O \left( \frac{\sigma^2_{\xi}}{NKN_Bd} (d\delta^2_{\xi,\perp} + \delta_+^2) \right). \]
(17)

By Lemma C.18, we have
\[ \frac{d}{dt} \frac{\|K_{A,q}\|_p^2}{\|K_{A,p}\|_p^2} \leq O \left( \frac{\sigma^2_{\xi}}{NKN_Bd} (d\delta^2_{\xi,\perp} + d^2\delta^2_{AB,\perp} + \delta_+ + d^2\delta_{N/S}\delta_{\xi,\perp}) \right) \frac{\|K_{A,q}\|_d^2}{\|K_{A,p}\|_p^2}. \]
Note that on the RHS of these equations, the only terms whose order may potentially be smaller than or equal to LHS are the \( \delta_+ \)-related terms. However, (17), we can make sure \( \delta_- \) is at most \( \delta^1_{AB,\perp} \). 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 \( \text{poly}(d) \) time if they are sufficiently small at the beginning of Stage 2. \( \square \)
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.