Suspicious Alignment of SGD: A Fine-Grained Step Size Condition Analysis

Shenyang Deng SHENYANG.DENG.GR@DARTMOUTH.EDU
Dartmouth College

Boyao Liao BXL307@STUDENT.BHAM.AC.UK
University of Birmingham

Zhuoli Ouyang ZHUOLI.OUYANG@DARTMOUTH.EDU
Dartmouth College

Tianyu Pang TIANYU.PANG.GR@DARTMOUTH.EDU
Dartmouth College

Minhak Song MINHAKSONG@KAIST.AC.KR
KAIST

Yaoqing Yang YAOQING.YANG@DARTMOUTH.EDU
Dartmouth College

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

This paper explores the suspicious alignment phenomenon in stochastic gradient descent (SGD) under ill-conditioned optimization, where the Hessian spectrum splits into dominant and bulk subspaces. This phenomenon describes the behavior of gradient alignment in SGD updates. Specifically, during the initial phase of SGD updates, the alignment between the gradient and the dominant subspace tends to decrease. Subsequently, it enters a rising phase and eventually stabilizes in a high-alignment phase. The alignment is considered "suspicious" because, paradoxically, the projected gradient update along this highly-aligned dominant subspace proves ineffective at reducing the loss. The focus of this work is to give a fine-grained analysis in a high-dimensional quadratic setup about how step size selection produces this phenomenon. Our main contribution can be summarized as follows: We propose a step-size condition revealing that in low-alignment regimes, an adaptive critical step size η_t^* separates alignment-decreasing ($\eta_t < \eta_t^*$) from alignment-increasing ($\eta_t > \eta_t^*$) regimes, whereas in high-alignment regimes, the alignment is self-correcting and decreases regardless of the step size. We further show that under sufficient ill-conditioning, a step size interval exists where projecting the SGD updates to the bulk space decreases the loss while projecting them to the dominant space increases the loss, which explains a recent empirical observation that projecting gradient updates to the dominant subspace is ineffective. Finally, based on this adaptive step-size theory, we prove that for a constant step size and large initialization, SGD exhibits this distinct two-phase behavior: an initial alignment-decreasing phase, followed by stabilization at high alignment.

Keywords: loss landscape, suspicious alignment, step size condition, optimization theory

1. Introduction

Deep neural networks are often optimized over ill-conditioned loss landscapes. Empirical studies by Sagun et al. (2016, 2017) have shown that the Hessian of the training loss in over-parameterized models typically exhibits a bimodal eigenvalue spectrum: a small number of large eigenvalues corresponding to directions of high curvature are sharply separated from a dense bulk of near-zero eigenvalues that span the vast majority of the parameter space. This spectral structure creates an ill-conditioned loss landscape: steep along a few dominant directions, yet nearly flat across a high-dimensional subspace. Such geometry is referred to in the literature as the river-valley landscape (Wen et al., 2024), where narrow, high-curvature "valleys" are embedded within a broad, flat "river". To formalize this structure, let $\nabla^2 L(\boldsymbol{x})$ denote the Hessian at a point $\boldsymbol{x}$, and assume it admits a spectral decomposition with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_d > 0$ and orthonormal eigenvectors $\{\boldsymbol{u}_i\}_{i=1}^d$. We partition the spectrum into a *dominant block* $\mathcal{D} = \{1, \ldots, k\}$ and a *bulk block* $\mathcal{B} = \{k+1, \ldots, d\}$, separated by a non-vanishing gap

$$\mathrm{gap}_1 := \lambda_k - \lambda_{k+1} > 0.$$

This induces orthogonal projectors onto the corresponding subspaces:

$$\boldsymbol{P}^{\mathcal{D}} = \sum_{i \in \mathcal{D}} \boldsymbol{u}_i \boldsymbol{u}_i^\top, \qquad \boldsymbol{P}^{\mathcal{B}} = \sum_{i \in \mathcal{B}} \boldsymbol{u}_i \boldsymbol{u}_i^\top.$$

Typically, the (river-valley) landscapes are ill-conditioned. In such landscapes, SGD exhibits a striking phenomenon: the gradient $\nabla L(\boldsymbol{x}_t)$ becomes increasingly aligned with the dominant subspace $\mathcal{D}$ as training progresses. This *tiny subspace* effect was documented by Gur-Ari et al. (2018), who showed that late-stage gradients lie almost entirely within the span of the top few Hessian eigenvectors. Other researchers also have similar findings (Schneider et al., 2024). Following by those results, a lot of works try to use this property to design more efficient learning algorithms (Gressmann et al., 2020; Gauch et al., 2022; Li et al., 2022a). To gain a better understanding of the intuition, let's define the alignment by

$$\theta_t := \frac{\|\boldsymbol{P}^{\mathcal{D}} \nabla L(\boldsymbol{x}_t)\|_2^2}{\|\nabla L(\boldsymbol{x}_t)\|_2^2} \in [0, 1].$$

Intuitively, $\theta_t \approx 1$ suggests that optimization is focused on the most curved directions, which may seem to be an opportunity to achieve efficiency. However, above intuition is contradicted by recent findings (Song et al., 2024): in the high-alignment regime ($\theta_t \approx 1$), the SGD updates projected onto the dominant space $\mathcal{D}$ often *fail to decrease the training loss*, while the orthogonal bulk component—despite carrying negligible gradient norm—does make the loss decrease. We call this **suspicious alignment**: a state where the gradient is mainly supported on $\mathcal{D}$, yet updates along $\mathcal{D}$ give non-decreasing or even ascending loss under typical step sizes. This leads to two questions:

> **Problem 1.** *How does the step size η_t govern the evolution of alignment θ_t in ill-conditioned landscapes?*
> **Problem 2.** *Why do Dominant Projected SGD (DSGD) fail to reduce the loss while Bulk Projected SGD (BSGD) succeed under high alignment ?*

We provide complete analytical answers to both questions in the quadratic case under the high-dimensional regime where $d \to \infty$. Our main results are boxed below:

Answer to Problem 1: Step-size condition of alignment dynamics. We identify an *adaptive critical step size* $\eta_t^*(\boldsymbol{x}_t)$ (Eq. (2)) that separates two distinct regimes:

- **Low-alignment regime** ($\theta_t \le g_{\text{gap}}$, where g_{gap} is a threshold smaller than 1): alignment is *dependent* on the step size η_t. If $\eta_t < \eta_t^*$, then $\mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] < \theta_t$; if $\eta_t > \eta_t^*$, then $\mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] > \theta_t$.
- **High-alignment regime** ($\theta_t \ge \theta_t^*$, where θ_t^* is another threshold smaller than 1): alignment is *self-correcting*. $\mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] < \theta_t$ for *any* $\eta_t > 0$.

Here $\mathbb{E}[\cdot|\boldsymbol{x}_t]$ means taking the expectation with respect to the gradient noise at t step. These regimes are separated by a stable interval $(g_{\text{gap}}, \theta_t^*)$, into which θ_t tends to oscillate. Furthermore, $g_{\text{gap}} \to 1, \theta_t^* \to 1$ as $\lambda_k/\lambda_{k+1} \to \infty$. This is an informal summary of Theorems 2, 3, 4, and 6.

Answer to Problem 2: Stability disparity between subspaces. For projected updates, expected loss decreases iff $\eta_t < \eta_{\mathcal{S}}^{\text{loss}}(\boldsymbol{x}_t)$, where for $\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}$,

$$\eta_{\mathcal{S}}^{\text{loss}}(\boldsymbol{x}_t) = \frac{2\sum_{i\in\mathcal{S}} \lambda_i^2 c_{i,t}^2}{\sum_{i\in\mathcal{S}} \lambda_i^4 c_{i,t}^2 + \sum_{i\in\mathcal{S}} \lambda_i \kappa_i^2},$$

with $c_{i,t} = \langle \boldsymbol{x}_t, \boldsymbol{u}_i \rangle$ and $\kappa_i^2 = \boldsymbol{u}_i^\top \boldsymbol{\Sigma} \boldsymbol{u}_i$. There exists a critical alignment $\theta_t^{\text{crit}} \in (0,1)$ such that

$$\theta_t < \theta_t^{\text{crit}} \iff \eta_{\mathcal{D}}^{\text{loss}} < \eta_{\mathcal{B}}^{\text{loss}}.$$

As $\lambda_k/\lambda_{k+1} \to \infty$, $\theta_t^{\text{crit}} \to 1$, so for nearly all $\theta_t < 1$, there exists an unstabele regime $(\eta_{\mathcal{D}}^{\text{loss}}, \eta_{\mathcal{B}}^{\text{loss}})$ for the dominant update. When the step size is inside $(\eta_{\mathcal{D}}^{\text{loss}}, \eta_{\mathcal{B}}^{\text{loss}})$, DSGD increases loss while BSGD decreases it, resolving the paradox of Song et al. (2024). This is an informal summary of Theorems 10, 11, and 12.

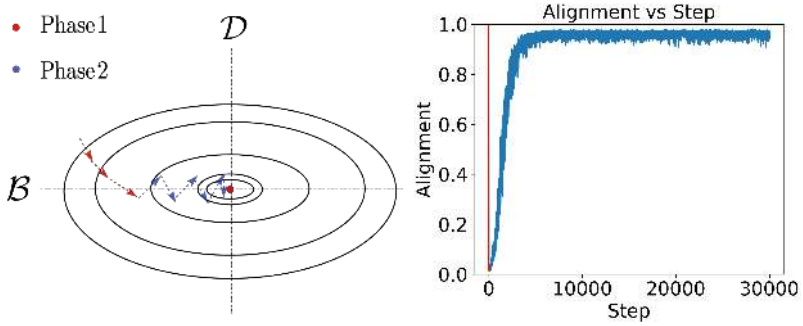


Figure 1: Two-phase SGD Dynamic: $\mathcal{D}, \mathcal{B}$ represent dominant and bulk direction respectively. This is a simulation experiment with constant step size for SGD with spectrum gap $\frac{\lambda_k}{\lambda_{k+1}} = 100$, the details can be referred to Section 7.

Building on this theoretical foundation, we further analyze the long-term alignment dynamics under *constant step size* SGD (CSGD). In Section 6, we characterize a distinct *two-phase behavior* of θ_t as show in Figure 1: an initial transient phase where alignment monotonically decreases

(Theorem 15), followed by a late-time equilibrium where θ_t converges to a noise- and spectrum-dependent limit θ_∞ (Theorem 16). This provides a complete picture of how ill-conditioning, noise geometry, and step size jointly shape the alignment trajectory of SGD.

2. Related Work

Ill-conditioned Loss Landscape The performance of deep neural networks is strongly affected by the geometric properties of their loss landscapes, as illustrated in the visualization in Li et al. (2018). Initial spectral analysis by Sagun et al. (2016, 2017) identified a bulk-dominant structure in the Hessian, characterized by a vast majority of near-zero eigenvalues alongside a small number of dominant outlier eigenvalues. Subsequent research has further characterized the properties of this structure across various tasks and architectures (Ghorbani et al., 2019; Papyan, 2019, 2020). Specifically, Ghorbani et al. (2019) observed that gradients predominantly align with the few outlier eigenvectors, while Wu et al. (2020) demonstrated that the Hessian remains consistently low-rank, with its significant curvature confined to a tiny fraction of the parameter space. These empirical observations are supported by theoretical proofs: Liao and Mahoney (2021) explained how outlier eigenvalues and bulk eigenvalues emerge from the coupling between the model and the data structure, while Singh et al. (2021) derived exact analytic formulas for the low-rank nature of Hessian maps. More recently, Hodgkinson et al. (2025) proposed that the Hessian eigenvalues of optimal models approximately follow a Heavy-Tailed Marchenko-Pastur (HTMP) distribution. A key characteristic of this distribution is its heavy-tailed nature, which is more likely to produce outlier eigenvalues than light-tail distribution. Furthermore, certain dynamic properties of this structure can exert unexpected impacts on optimization. In PINNs, Rathore et al. (2024) showed that the presence of multiple conflicting loss terms leads to a highly unstable outlier structure where outlier eigenvalues frequently shift and overlap during training, triggering severe instabilities. In addition to these primary studies, other works have explored broader connections between this structure and various optimization phenomena (Zhang et al., 2019; Yang et al., 2021; Garipov et al., 2018; Pan et al., 2021).

Gradient Alignment Theory In the context of such ill-conditioned loss landscapes, Gur-Ari et al. (2018) found that SGD consistently aligns with the large eigenvalue directions in the later stages. Since the large eigenvalue space is a very low-dimensional subspace, they posited that SGD operates within a *tiny subspace*. However, subsequent observations by Song et al. (2024) overturned this conclusion. By projecting SGD updates onto the corresponding subspaces during the middle stage of training, they discovered that updates projected onto the large eigenvalue space do not reduce the loss. Instead, updates projected onto the small eigenvalue space lead to further loss reduction. Recent theoretical work on loss landscapes has also pertained to the alignment phenomenon of gradient under ill-conditioned structures. For example, Li et al. (2022b) proved in their Theorem 3.1 that, for edge of stability (EOS) phenomenon (Cohen et al., 2021) and for quadratic programming problems, if gradient descent uses a normalized step-size schedule, the gradient $\boldsymbol{Ax}$ aligns with the largest eigenvector of the Hessian as time approaches infinity. Later, in Theorem 4.4, they established gradient alignment results for normalized gradient flow on manifolds that satisfy certain assumptions. In the context of studying LLM step size schedules, Wen et al. (2024) provided in Lemma A.13 that, for a specific loss landscape structure, after a certain time t, the ratio of the projections of the gradient flow onto the large and small eigenspaces (which is the alignment) has an asymptotic upper bound with t. But, we note that the SGD results in Wen et al. (2024) con-

cerns track the trajectory near the river only, and it does not have a result on trajectory alignment. In conclusion, these theoretical alignment results are based on gradient descent rather than SGD. Moreover, the works above are not specifically designed to study the loss landscape or the impact of step-size schedules on the alignment phenomenon. Furthermore, we cannot directly use their lemma to explain the observations reported by Song et al. (2024).

3. Preliminary

3.1. Common Notation

Before moving on to the main content, we will first introduce some common notation. We denote $\mathbb{R}^d$ as the d-dimensional Euclidean space. Bold lowercase letters (such as $\boldsymbol{u}$, $\boldsymbol{v}$) denote column vectors; Bold uppercase letters (such as $\boldsymbol{M}$) denote matrices; lowercase letters (such as a, b) denote scalars. Sets are denoted by calligraphic letters (such as $\mathcal{D}$, $\mathcal{B}$). The norm $\|\cdot\|_p$ represents the p-norm for vectors or the p-operator norm for matrices; $\|\cdot\|_F$ represents the Frobenius norm for matrices. $\mathrm{Tr}(\boldsymbol{M})$ denotes the trace of matrix $\boldsymbol{M}$.

3.2. Setup and Assumptions

We study a quadratic optimization problem under SGD with arbitrary noise of finite second moment, focusing on how the Hessian eigenspectrum and the noise covariance influence gradient alignment with the dominant eigenspace of the Hessian.

Quadratic Loss and Spectral Decomposition We consider the quadratic loss function

$$L(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \qquad \nabla L(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is symmetric positive definite. The matrix $\boldsymbol{A}$ has the spectral decomposition

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top, \quad \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d), \quad \lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_d > 0,$$

with orthonormal eigenvectors $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d]$. We express the state $\boldsymbol{x}_t$ in the eigenbasis as $\boldsymbol{x}_t = \sum_{i=1}^d c_{i,t}\boldsymbol{u}_i$, where $c_{i,t} := \langle \boldsymbol{x}_t, \boldsymbol{u}_i\rangle$.

Dominant and Bulk Spectral Recall that we partition the spectrum into a dominant block and a bulk block:

$$\mathcal{D} = \{1, \ldots, k(d)\}, \qquad \mathcal{B} = \{k(d)+1, \ldots, d\}.$$

Here, we assume $k(d)$ is a function of d. For simplicity, in the following, we denote $k(d)$ by k, where $k \in \{1, \ldots, d-1\}$. We define projection matrices for the dominant and bulk subspaces:

$$\boldsymbol{P}^{\mathcal{D}} := \sum_{i\in\mathcal{D}} \boldsymbol{u}_i\boldsymbol{u}_i^\top, \quad \boldsymbol{P}^{\mathcal{B}} := \sum_{i\in\mathcal{B}} \boldsymbol{u}_i\boldsymbol{u}_i^\top,$$

and the squared alignment function $\theta(\boldsymbol{x}_t)$ of the gradient $\boldsymbol{A}\boldsymbol{x}_t$ with the dominant subspace at time t define as

$$\theta(\boldsymbol{x}_t) := \begin{cases} \dfrac{\|\boldsymbol{P}^{\mathcal{D}}\nabla L(\boldsymbol{x}_t)\|_2^2}{\|\nabla L(\boldsymbol{x}_t)\|_2^2} = \dfrac{\|\boldsymbol{P}^{\mathcal{D}}\boldsymbol{A}\boldsymbol{x}_t\|_2^2}{\|\boldsymbol{A}\boldsymbol{x}_t\|_2^2} = \dfrac{\sum_{i\in\mathcal{D}}\lambda_i^2 c_{i,t}^2}{\sum_{i=1}^d \lambda_i^2 c_{i,t}^2} \in [0,1] & \text{if } \boldsymbol{x}_t \neq \boldsymbol{0}, \\ 0 & \text{if } \boldsymbol{x}_t = \boldsymbol{0}, \end{cases}$$

which captures the proportion of the gradient's norm in the top-k dominant eigenspace (with the convention $\theta(\mathbf{0}) = 0$ ensuring well-definedness in stochastic analysis). We will use θ_t for short in the later section. To analyze the alignment, we need to introduce the following notation for quantities defined in certain subspaces:

$$\psi_{\mathcal{D}} := \sum_{i\in\mathcal{D}} \lambda_i^2, \quad \psi_{\mathcal{B}} := \sum_{i\in\mathcal{B}} \lambda_i^2,$$

$$s_t^{\mathcal{D}} := \sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t}^2, \quad s_t^{\mathcal{B}} := \sum_{i\in\mathcal{B}} \lambda_i^2 c_{i,t}^2, \quad s_t := s_t^{\mathcal{D}} + s_t^{\mathcal{B}} = \sum_{i=1}^{d} \lambda_i^2 c_{i,t}^2. \tag{1}$$

We assume the following eigenvalue gaps to ensure a clear separation between the bulk and dominant eigenvalues. The gap is defined at the boundary between the k-th and $(k+1)$-th eigenvalues.

$$\mathrm{gap}_1 := \lambda_k - \lambda_{k+1} > 0, \quad \mathrm{gap}_2 := \lambda_k^2 - \lambda_{k+1}^2 > 0.$$

SGD Update with Noise The SGD update with step size $\eta_t > 0$ is given by

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t), \quad \mathbb{E}[\boldsymbol{\xi}_t] = \mathbf{0}, \mathrm{Cov}[\boldsymbol{\xi}_t] = \boldsymbol{\Sigma}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top \succeq 0,$$

where $\boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a Gaussian random vector. Define $\boldsymbol{C} := \boldsymbol{U}^\top \boldsymbol{\Sigma} \boldsymbol{U}$ as the noise covariance in the $\boldsymbol{A}$-basis, with per-eigendirection variances $\kappa_i^2 := (\boldsymbol{C})_{ii} = \boldsymbol{u}_i^\top \boldsymbol{\Sigma} \boldsymbol{u}_i$. We introduce the following quantities to represent block-wise noise energy:

$$e_{\mathcal{D}} := \sum_{i\in\mathcal{D}} \lambda_i^2 \kappa_i^2, \quad e_{\mathcal{B}} := \sum_{i\in\mathcal{B}} \lambda_i^2 \kappa_i^2.$$

The noise covariance eigenvalues $s_{\min} := \lambda_{\min}(\boldsymbol{\Sigma})$ and $s_{\max} := \lambda_{\max}(\boldsymbol{\Sigma})$ yield bounds:

$$s_{\min}\psi_{\mathcal{D}} \le e_{\mathcal{D}} \le s_{\max}\psi_{\mathcal{D}}, \quad s_{\min}\psi_{\mathcal{B}} \le e_{\mathcal{B}} \le s_{\max}\psi_{\mathcal{B}}.$$

Why this Model? This quadratic framework, by separately modeling the components of $\boldsymbol{x}_t$ in high-curvature (dominant) and low-curvature (bulk) directions as well as the corresponding noise($\boldsymbol{\xi}_t$) components, facilitates obtaining more fine-grained analytical conclusions and understanding the impact of alignment on optimization. It provides a tractable model for studying how step sizes and noise geometry drive gradient alignment in SGD.

Assumption 1 (Asymptotic Spectral Assumptions) *When we consider the high-dimensional regime, where both d and $k(d) \to \infty$, we assume the following conditions, summarized in Table 1.*

4. Step Size Condition Theory

For the SGD update $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t)$ with $\boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, we define the adaptive critical step size at time t as

$$\eta_t^*(\boldsymbol{x}_t) := \frac{2\Big(\big(\sum_{i\in\mathcal{B}} \lambda_i^2 c_{i,t}^2\big)\big(\sum_{j\in\mathcal{D}} \lambda_j^3 c_{j,t}^2\big) - \big(\sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t}^2\big)\big(\sum_{j\in\mathcal{B}} \lambda_j^3 c_{j,t}^2\big)\Big)}{\big(\sum_{i\in\mathcal{B}} \lambda_i^2 c_{i,t}^2\big)\Big(\sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 + e_{\mathcal{D}}\Big) - \big(\sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t}^2\big)\Big(\sum_{j\in\mathcal{B}} \lambda_j^4 c_{j,t}^2 + e_{\mathcal{B}}\Big)}. \tag{2}$$

Recall $e_{\mathcal{D}} = \sum_{i\in\mathcal{D}} \lambda_i^2 \kappa_i^2$, $e_{\mathcal{B}} = \sum_{i\in\mathcal{B}} \lambda_i^2 \kappa_i^2$ with $\kappa_i^2 = \boldsymbol{u}_i^\top \boldsymbol{\Sigma} \boldsymbol{u}_i$.

Table 1: Asymptotic Spectral Assumptions

Assumption	Description
Trajectory boundedness	The state remains bounded: $\sup_t \limsup_{d\to\infty} \frac{1}{d}\sum_{i=1}^{d} c_{i,t}^2 < \infty$.
Block proportion	The subspace dimension ratio is a fixed constant: $\rho := \frac{k}{d-k} \in (0,\infty)$. Which implies $k = \frac{\rho}{1+\rho}d$.
Block spectral moments	For $p \in \{2,3,4,6,8\}$, the block-wise spectral moments converge: $\frac{1}{k}\sum_{i\in\mathcal{D}} \lambda_i^p \to \lambda_{\mathcal{D},p} \in (0,\infty)$, $\frac{1}{d-k}\sum_{i\in\mathcal{B}} \lambda_i^p \to \lambda_{\mathcal{B},p} \in [0,\infty)$.
Noise spectral bounds	The noise covariance has a bounded trace: $Tr(\mathbf{\Sigma}) \in (0,+\infty)$

Theorem 2 (*Decrease Condition*) *Under Assumption 1, if* $0 < \eta_t < \eta_t^*(\boldsymbol{x}_t)$*, then*

$$\lim_{d\to\infty} \mathbb{E}\big[\theta_{t+1} \,\big|\, \boldsymbol{x}_t\big] \;<\; \theta_t.$$

The detailed proof of Theorem 2 is in Appendix A.6.

Interpretation When the step size is smaller than the adaptive threshold (2), the expected alignment function decreases in one step. This is the "small-step" phase. The adaptive threshold depends on $(\boldsymbol{A}, \mathbf{\Sigma})$ and the current state $\boldsymbol{x}_t$ through the function (2).

Theorem 3 (*Increase Condition*) *Under Assumption 1, let*

$$g_{\text{gap}} := \frac{1}{1 + \frac{s_{\max}}{s_{\min}} \cdot \frac{1}{\rho}\left(\frac{\lambda_{k+1}}{\lambda_k}\right)^2} \in (0,1).$$

For t *and* $\boldsymbol{x}_t$*. If* $\theta_t \le g_{\text{gap}}$ *and* $\eta_t > \eta_t^*(\boldsymbol{x}_t)$*, then*

$$\lim_{d\to\infty} \mathbb{E}\big[\theta_{t+1} \,\big|\, \boldsymbol{x}_t\big] \;>\; \theta_t.$$

The detailed proof of Theorem 3 is in Appendix A.6.

Interpretation If the current alignment is in the "small regime" (meaning less than g_{gap}), and the current step size η_t exceeds the adaptive threshold, the dominant alignment increases. Notice that g_{gap} is an upper bound which does not depend on $\boldsymbol{x}_t$. As the dominant-bulk gap grows ($\lambda_k/\lambda_{k+1} \to \infty$), $g_{\text{gap}} \to 1$, so the "increase region" $[0, g_{\text{gap}})$ expands to nearly all $\theta_t \in [0,1)$.

Theorem 4 (*Large Alignment Regime Condition*) *Under Assumption 1, there exists a critical alignment threshold* $\theta_t^* \in (0,1)$ *such that if the current alignment* $\theta_t \ge \theta_t^*$*, then for any positive step size* $\eta_t > 0$*, the expected alignment decreases:*

$$\lim_{d\to\infty} \mathbb{E}\big[\theta_{t+1} \,\big|\, \boldsymbol{x}_t\big] \;<\; \theta_t.$$

The threshold is given by $\theta_t^* := r_{0,t}/(1 + r_{0,t})$*, where* $r_{0,t}$ *is the positive root of the quadratic equation:*

$$a_t^{\text{aux}} r^2 + (a_t^{\text{aux}} - m_t^{\text{aux}} - h_t^{\text{aux}})r - h_t^{\text{aux}} = 0.$$

Recall that $\psi_{\mathcal{S}} := \sum_{i\in\mathcal{S}} \lambda_i^2$ *(*$\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}$*) and* $s_t := \sum_{i=1}^{d} \lambda_i^2 c_{i,t}^2$*, the coefficients in the quadratic equation are defined as:*

$$\begin{aligned} a_t^{\text{aux}} &:= s_{\min}\psi_{\mathcal{B}}, \\ h_t^{\text{aux}} &:= s_{\max}\psi_{\mathcal{D}}, \\ m_t^{\text{aux}} &:= s_t(\lambda_1^2 - \lambda_d^2). \end{aligned}$$

Furthermore, as $\psi_{\mathcal{D}}/\psi_{\mathcal{B}} \to \infty$*, we have* $\theta_t^* \to 1$*.*

The detailed proof of Theorem 4 is in Appendix A.6.

Theorem 5 (*Asymptotic rate of* θ_t^*) *Under Assumption 1, let* $m := \lambda_k/\lambda_{k+1} > 1$*. There exist constants* $\alpha, \beta > 0$ *such that*

$$1 - \frac{\beta}{m^2} \;\le\; \theta_t^* \;\le\; 1 - \frac{\alpha}{m^2}.$$

In particular,

$$\theta_t^* = 1 - \Theta\left(\frac{1}{m^2}\right).$$

Interpretation *Theorem 4 and 5 shows that a very large alignment value can be* self-correct*: regardless of the step size, the expected alignment decreases for* θ_t *beyond a computable threshold* θ_t^**. As the dominant-bulk gap grows (*$\lambda_k/\lambda_{k+1} \to \infty$*),* $\theta_t^* \to 1$*, so the "decrease" region* $(\theta_t^*, 1)$ *shrinks toward the maximum alignment value of 1.*

Theorem 6 (*Separation of Alignment Regimes*) *Under Assumption 1, for any nontrivial problem with a non-zero spectral gap* ($\text{gap}_2 > 0$) *and bounded noise* ($s_{\max} < \infty$)*, the low-alignment threshold* g_{gap} *is strictly less than the high-alignment threshold* θ_t^**:*

$$g_{\text{gap}} < \theta_t^*.$$

The detailed proof of Theorem 6 is in Appendix A.6.

Summary The preceding theorems delineate two primary "alignment regimes" based on the current alignment value θ_t. These regimes, which govern the core behavior of the SGD dynamics, are presented below.

$\boxed{\theta_t \in (0, g_{\text{gap}}]}$	$\boxed{\theta_t \in [\theta_t^*, 1)}$
Alignment is dependent on η_t: $\mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] > \theta_t$ if $\eta_t > \eta_t^*$, $\mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] < \theta_t$ if $\eta_t < \eta_t^*$	Alignment is self-correcting: $\mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] < \theta_t$ for any $\eta_t > 0$
Low-Alignment Regime	High-Alignment Regime

These two primary regimes are separated by an intermediate interval, $(g_{\text{gap}}, \theta_t^*)$, which can be viewed as a *stable regime*. For a sufficiently large step size (i.e., $\eta_t > \eta_t^*$), the dynamics create an oscillating behavior: if alignment drops below g_{gap}, it is pushed back up, and if it exceeds θ_t^*, it is forced back down. Consequently, the alignment θ_t will tend to *oscillate within this stable interval* $(g_{\text{gap}}, \theta_t^*)$. Crucially, this stable interval is dependent on the spectral structure of the Hessian. As the

gap between the dominant and bulk eigenvalues grows (i.e., as $\lambda_k/\lambda_{k+1} \to \infty$), both boundaries of this interval converge towards 1. This means the interval itself asymptotically approaches the single point $\{1\}$, driving the alignment to stay close to the maximum value. To provide an intuitive understanding of the connection between η_t^* and the current state of $\boldsymbol{x}_t$, we present the following theorem:

Theorem 7 (*State- and gap-aware lower bounds on* η_t^* *(with* $\|\boldsymbol{x}_t\|_2$*)) *Under Assumption 1, for any* $\boldsymbol{x}_t$*, we have*

$$\eta_t^* \;\geq\; \frac{2\,\mathrm{gap}_1}{(\lambda_1^2-\lambda_d^2) \;+\; \dfrac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2\,\|\boldsymbol{x}_t\|_2^2\,\theta_t}}.$$

The detailed proof of Theorem 7 is in Appendix A.6.

Remark The lower bounds *increase* with the parameter norm $\|\boldsymbol{x}_t\|$ and with alignment θ_t (or decrease with $1-\theta_t$ by symmetry), and with the first-order gap gap_1. Or, more essentially, based on the expression for θ_t, we can conclude that $\theta_t > \frac{\lambda_k^2\|\boldsymbol{P}^{\mathcal{D}}\boldsymbol{x}_t\|_2^2}{\lambda_1^2\|\boldsymbol{x}_t\|_2^2}$. Finally, we obtain:

$$\eta_t^* \;\geq\; \frac{2\,\mathrm{gap}_1}{(\lambda_1^2-\lambda_d^2) \;+\; \dfrac{\lambda_1^2\,s_{\max}\psi_{\mathcal{D}}}{\lambda_k^2\lambda_d^2\,\|\boldsymbol{P}^{\mathcal{D}}\boldsymbol{x}_t\|_2^2}}.$$

This means that when the dominant space has a sufficiently large projection norm (for example, at initialization, a large random initialization can satisfy the condition with high probability), η_t^* 's lower bound $\approx \frac{2\mathrm{gap}_1}{\lambda_1^2-\lambda_d^2}$. If the eigenvalues in different subspaces can be approximated (i.e., $\forall i \in \mathcal{D}, \lambda_1 \approx \lambda_i$, and $\forall i \in \mathcal{B}, \lambda_d \approx \lambda_i$), we further obtain η_t^*'s lower bound $\approx \frac{2}{\lambda_1+\lambda_d}$.

Theorem 8 (*State- and gap-aware upper bounds on* η^*) *Under Assumption 1, providing that the alignment satisfies* $\theta_t \geq \frac{e_{\mathcal{B}}}{e_{\mathcal{B}}+e_{\mathcal{D}}}$*, we have:*

$$\eta_t^* \;\leq\; \frac{(\lambda_1-\lambda_d)}{\lambda_k\lambda_1-\lambda_{k+1}\lambda_d}.$$

Corollary 9 (*The comparison between* η_t^* *and* $\frac{2}{\lambda_1}$)** *Suppose the conditions of Theorem 8 hold. We have the following relation between* η_t^* *and the convergence step size for Quadratic Programming:*

$$\eta_t^* \;\leq\; \frac{2}{\lambda_1}.$$

Remark Here, Theorem 8 offers the intuition that when the alignment function θ_t exceeds a critical threshold $\frac{e_{\mathcal{B}}}{e_{\mathcal{B}}+e_{\mathcal{D}}}$ (Note that this threshold depends solely on the noise covariance matrix), we can obtain Corollary 9 that η^* is smaller than the critical step size required for gradient descent convergence in quadratic programming($\frac{2}{\lambda_1}$).

5. Behavioral Analysis of Projected SGD under Alignment Conditions

Recalling the finding by Gur-Ari et al. (2018) that late-stage gradients are primarily confined to the dominant subspace (high alignment, $\theta_t \to 1$), the subsequent work of Song et al. (2024) revealed a seemingly paradoxical phenomenon. They demonstrated that in this high-alignment regime, updates projected onto this supposedly critical dominant direction do not necessarily decrease the loss; instead, it is often the projection onto the orthogonal bulk subspace that guarantees a loss reduction. We resolve this apparent contradiction by developing a precise theory for the step-size conditions under which each projected update is guaranteed to decrease the expected loss.

Projected SGD To rigorously analyze this phenomenon, we define two idealized algorithms where the stochastic gradient update is explicitly projected onto the dominant and bulk subspaces, respectively. The update rules for the Projected SGD algorithms are given as follows:

$$\textit{Dominant Projected SGD (DSGD):} \qquad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{P}^{\mathcal{D}}(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t). \tag{3}$$

$$\textit{Bulk Projected SGD (BSGD):} \qquad \boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{P}^{\mathcal{B}}(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t). \tag{4}$$

To analyze their decay condition, we first define the necessary state- and noise-dependent quantities for a subspace $\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}$:

$$\tau_t^{\mathcal{S}} := \sum_{i\in\mathcal{S}} \lambda_i^3 c_{i,t}^2, \qquad n_{\mathcal{S}}^{\text{loss}} := \sum_{i\in\mathcal{S}} \lambda_i \kappa_i^2, \qquad \mu_t^{\mathcal{S}}(p,q) := \frac{\sum_{i\in\mathcal{S}} \lambda_i^p c_{i,t}^2}{\sum_{i\in\mathcal{S}} \lambda_i^q c_{i,t}^2}. \tag{5}$$

Theorem 10 (*Step Size Condition for Projected Updates*) *For a given state $\boldsymbol{x}_t$, the one-step expected loss decreases if and only if the step size η_t satisfies the following conditions for the dominant and bulk projected updates, respectively:*

$$\textit{For } \mathcal{D}: \quad \mathbb{E}[L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) \mid \boldsymbol{x}_t] < 0 \iff 0 < \eta_t < \eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) := \frac{2\, s_t^{\mathcal{D}}}{\tau_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}}}$$

$$\textit{For } \mathcal{B}: \quad \mathbb{E}[L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) \mid \boldsymbol{x}_t] < 0 \iff 0 < \eta_t < \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t) := \frac{2\, s_t^{\mathcal{B}}}{\tau_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}}$$

The detailed proof of Theorem 10 is in Appendix A.7. The relationship between these two thresholds is not fixed; it is dynamically determined by the alignment of the current state. The following theorems characterize this crossover phenomenon.

Theorem 11 (*Condition Differences on Different Alignment Regime*) *Under Assumption 1, the relative ordering of the two loss thresholds $\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t)$ and $\eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t)$ is determined by the alignment θ_t of a given $\boldsymbol{x}_t$ relative to the unique critical threshold, $\theta_t^{crit} \in (0,1)$. This threshold is the root of the quadratic equation $h(\theta) = \alpha_t\theta^2 + \beta_t\theta + \gamma_t = 0$, with coefficients defined by:*

$$\alpha_t = s_t \left(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\right) \tag{6}$$

$$\beta_t = -\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}} \tag{7}$$

$$\gamma_t = -n_{\mathcal{D}}^{\text{loss}} \tag{8}$$

Where $n_{\mathcal{S}}^{\text{loss}}, \mu_t^{\mathcal{S}}(p,q)$ is defined in formula (5). The ordering of the thresholds in the two resulting regimes is as follows:

$$\begin{aligned} \theta_t < \theta_t^{crit} \quad &\Longleftrightarrow \quad \eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) < \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t), \\ \theta_t > \theta_t^{crit} \quad &\Longleftrightarrow \quad \eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) > \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t). \end{aligned}$$

The detailed proof of Theorem 11 is in Appendix A.7. Furthermore, as stated in the following theorem, this critical alignment threshold exhibits a predictable behavior in ill-conditioned settings, which converges to a state of full alignment 1.

Theorem 12 (*Asymptotic Limit of the Alignment Threshold*) *Under Assumption 1, let the spectral gap be denoted by $m := \lambda_k/\lambda_{k+1}$. In the limit as the gap grows infinitely large, the critical threshold in Theorem 11 converges to 1:*

$$\lim_{m\to\infty} \theta_{\text{crit}}^{\eta}(\boldsymbol{x}_t) = 1.$$

Summary The central finding in this section is the existence of a critical alignment threshold, θ_t^{crit}, that dictates the relative stability between DSGD and BSGD. This threshold partitions the alignment space into two regimes with inverted stability orderings, as summarized below:

$\boxed{\theta_t < \theta_t^{\text{crit}}}$	$\boxed{\theta_t > \theta_t^{\text{crit}}}$
$\underbrace{\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) < \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t)}$	$\underbrace{\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) > \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t)}$
Low-Alignment Regime	High-Alignment Regime

Our asymptotic analysis shows that as the spectral gap grows ($m := \lambda_k/\lambda_{k+1} \to \infty$), the critical threshold converges to one: $\lim_{m\to\infty} \theta_t^{\text{crit}} = 1$. This has a direct consequence on the operational regimes: the high-alignment regime $(\theta_t^{\text{crit}}, 1)$ asymptotically vanishes to $\{1\}$. Conversely, the low-alignment regime $[0, \theta_t^{\text{crit}})$ expands to occupy nearly the entire alignment space $[0, 1)$. The dynamics within this now-predominant low-alignment regime are therefore of primary importance. Here, the stability thresholds are ordered such that $\eta_{\mathcal{D}}^{\text{loss}} < \eta_{\mathcal{B}}^{\text{loss}}$, which leads to the following different descent behaviors:

$$\forall\, \eta_t \in (\eta_{\mathcal{D}}^{\text{loss}}, \eta_{\mathcal{B}}^{\text{loss}}), \quad \begin{cases} \text{For } \mathcal{D}: & \mathbb{E}[L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) \mid \boldsymbol{x}_t] > 0 \quad \text{(Loss Increases)} \\ \text{For } \mathcal{B}: & \mathbb{E}[L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) \mid \boldsymbol{x}_t] < 0 \quad \text{(Loss Decreases)} \end{cases}$$

This provides a theoretical explanation for the empirical findings of Song et al. (2024), which observed that updates along dominant directions can increase while those in the bulk remain decreases, although the gradient is aligned with dominant space. However, it should be noted that the two alignment regimes here are not directly related to those in the previous Section 4, although they exhibit the same behavior in the asymptotic case.

Theorem 13 (*Rate of the critical alignment threshold*) *Under Assumption 1, let $m := \frac{\lambda_k}{\lambda_{k+1}} > 1$. Then the critical alignment threshold $\theta_t^{\text{crit}} \in (0,1)$ satisfies*

$$\frac{n_{\mathcal{B}}^{\text{loss}}}{s_t\,\lambda_{k+1}(m-1) + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}} \;\le\; 1 - \theta_t^{\text{crit}} \;\le\; \frac{n_{\mathcal{B}}^{\text{loss}}}{s_t\,\lambda_{k+1}(m-1)}.$$

If $\lambda_{k+1} = \Theta(1)$*, consequently,*

$$\lambda_k = \Theta(m), 1 - \theta_t^{\text{crit}} \in \Theta\left(\frac{1}{s_t\,(m-1)}\right).$$

Remark Note that the asymptotic order here is simultaneously related to both s_t and m. s_t is positively correlated with the loss function $L(\boldsymbol{x}_t)$. Although we discussed in Theorem 5 that θ_t^* is asymptotically 1 with order $\Theta\left(\frac{1}{m^2}\right)$. Notice that $\theta_t^* > \theta_t^{crit}$ with large enough m, which implies that the alignment stable regime may drop in Bulk projection unstable regime. This seems to contradict the empirical phenomena, but actually it doesn't. We recall the phenomenon observed by Song et al. (2024): when the loss has not yet converged (i.e., $L(\boldsymbol{x}_t)$ has not reached a very small regime), even the alignment is already very high at this point, if we start performing subspace projection on the updates, we will observe the suspicious alignment phenomenon. The conditions here, in addition to the inherently ill-conditioned structure of the loss Hessian itself, also require that the loss $L(\boldsymbol{x}_t)$ is still in the pre-convergence stage.

6. Constant Step Size SGD: Two-Phase Dynamics of θ_t

Building upon the step size condition theory from the preceding sections, we now investigate the long-term dynamics under the *constant step size* SGD, which we will refer to as CSGD in the later content. This specialization allows us to characterize a distinct two-phase learning dynamic: an initial, transient phase driven by the initial state $\boldsymbol{x}_0$, and a late-time, steady-state equilibrium driven by noise. In this section, we make the following assumptions about the step size and initialization for CSGD. Recall that $c_{i,t} := \langle \boldsymbol{x}_t, \boldsymbol{u}_i \rangle$ is the projection of the state $\boldsymbol{x}_t$ onto the i-th eigenvector of matrix $\boldsymbol{A}$. To give a further analysis, we define several key quantities:

$$\beta_i := \frac{\eta\,\kappa_i^2}{2\lambda_i - \eta\lambda_i^2} > 0, \varrho_{\mathcal{D}} := \sum_{i\in\mathcal{D}} (c_{i,0}^2 - \beta_i).$$

$$\delta := \frac{s_{\max}\psi_{\mathcal{D}}\lambda_1^2}{\lambda_d^2\lambda_k^2\left(\frac{2(\lambda_k-\lambda_{k+1})}{\eta} - (\lambda_1^2 - \lambda_d^2)\right)}.$$

Assumption 14 *Our analysis is based on the following assumptions:*

Table 2: Assumptions for CSGD

Assumption	**Description**
Constant Step Size	$\eta_t = \eta, \quad 0 < \eta < \min\left\{\frac{2}{\lambda_1}, \frac{2(\lambda_k-\lambda_{k+1})}{\lambda_1^2-\lambda_d^2}\right\}$
Initialization for $\boldsymbol{x}_0$	$\forall i \in \mathcal{D}, \quad c_{i,0}^2 > \beta_i,$ $\varrho_{\mathcal{D}} > \delta - \sum_{i\in\mathcal{D}} \beta_i$

Theorem 15 *(Initial Decrease Phase) Under Assumption 1 and Assumption 14, let the t^* be defined as*

$$t^* := \left\lfloor \frac{\log\left(\dfrac{\varrho_{\mathcal{D}}}{\delta - \sum_{i\in\mathcal{D}} \beta_i}\right)}{-2\log(1-\eta\lambda_1)} \right\rfloor.$$

Where $\lfloor\cdot\rfloor$ is the floor function. Then for all time steps $t \in \{0, 1, \ldots, t^* - 1\}$, the expected alignment is strictly decreasing:

$$\lim_{d\to\infty} \mathbb{E}[\theta_{t+1}] < \lim_{d\to\infty} \mathbb{E}[\theta_t].$$

The proof of Theorem 15 is in Appendix A.9.

Interpretation Theorem 15 establishes the existence of a predictable initial phase in SGD dynamics, guaranteeing that for a sufficiently large initialization, the expected alignment $\mathbb{E}[\theta_t]$ will monotonically decrease for a calculable duration of t^* steps. The formula for t^* explicitly links this phase duration to the initial state: a positive-length phase is guaranteed when the initial $\varrho_{\mathcal{D}}$, exceeds a threshold. Furthermore, the theorem shows that the length of this phase, t^*, grows logarithmically with this initial $\varrho_{\mathcal{D}}$.

Theorem 16 *(Late Phase) Under Assumption 1 and Assumption 14, the late-time asymptotic expected alignment is given by*

$$\theta_\infty := \lim_{t\to\infty}\lim_{d\to\infty} \mathbb{E}_t[\theta_t] = \frac{\lim_{d\to\infty}\sum_{i\in\mathcal{D}} \lambda_i^2\beta_i}{\lim_{d\to\infty}\sum_{i=1}^{d} \lambda_i^2\beta_i},$$

where $\beta_i = \dfrac{\eta\kappa_i^2}{2\lambda_i - \eta\lambda_i^2} > 0$. Equivalently,

$$\theta_\infty = \frac{\lim_{d\to\infty}\sum_{i\in\mathcal{D}} \dfrac{\eta\lambda_i^2\kappa_i^2}{2\lambda_i - \eta\lambda_i^2}}{\lim_{d\to\infty}\sum_{i=1}^{d} \dfrac{\eta\lambda_i^2\kappa_i^2}{2\lambda_i - \eta\lambda_i^2}}.$$

The proof of Theorem 16 is in Appendix A.9.

Interpretation After the transient decay, the alignment settles to a stable level θ_∞ determined solely by the Hessian spectrum $\{\lambda_i\}$, the noise covariance $\{\kappa_i^2\}$, and the step size η. In the case of isotropic noise ($\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, so $\kappa_i^2 = \sigma^2$ for all i) and small step size $\eta \ll 1$, we have the approximation

$$\theta_\infty \approx \frac{\lim_{d\to\infty}\sum_{i\in\mathcal{D}} \lambda_i}{\lim_{d\to\infty}\sum_{i=1}^{d} \lambda_i}.$$

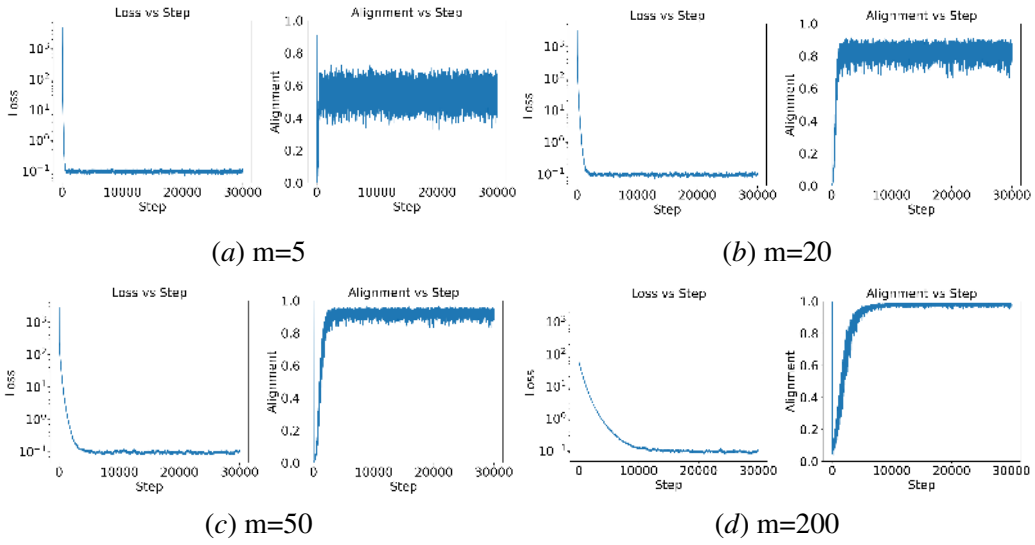
Furthermore, if the dominant–bulk eigenvalue gap grows such that $\lambda_k/\lambda_{k+1} \to \infty$ while the bulk spectrum remains bounded away from zero, then $\theta_\infty \to 1$. This provides a theoretical explanation for the long-run alignment of SGD with the dominant eigenspace.

7. Numerical Experiment

Numerical Simulation with Different Spectrum Gap We performed several numerical simulations of constant step size SGD under varying spectral gaps. In these experiments, we fixed the dimension $d = 500$, the dominant space dimension $k = 50$, the step size $\eta = 0.003$, and the total number of steps $T = 30,000$. For each spectral gap

$$m = \lambda_k/\lambda_{k+1} \in \{5, 10, 20, 50, 100, 200, 300, 400, 500\},$$

We randomly initialized positive definite symmetric matrices A with the corresponding spectral gap and conducted comparative experiments. For brevity, We selected four experiments to present in the main text, with simulations for $m = 5, 20, 50, 200$. We also plotted the loss decline curves to ensure that the algorithm converges to the convergence stage, with specific results shown in Figure 2. The complete experimental results and setup are available at Appendix B. It can be observed that all experiments exhibit a two-phase phenomenon, where the alignment function initially decreases in the short term and then increases. Furthermore, as m increases, the stable region of the alignment function gradually approaches 1.



(*a*) m=5 (*b*) m=20

(*c*) m=50 (*d*) m=200

Figure 2: Numerical simulation experiments with different spectral gaps ($m = \lambda_k/\lambda_{k+1}$)

The Simulation for $\mathbb{E}[\theta_\infty]$ and Spectrum Gap Following the setup from the previous experiment (detailed in Appendix B, where a constant step size is employed for SGD), we performed a simulation study to investigate the relationship between the limiting stable value θ_∞ and the spectral gap $m = \lambda_k/\lambda_{k+1}$ in the expectation context. Here, θ_T denotes the value at time T, which marks the beginning of the second phase where the system enters stability, and T_{end} denotes the end time of the experiment. We use the statistic:

$$\mathbb{E}[\text{Alignment}] = \frac{1}{T_{\text{end}} - T} \sum_{t=T}^{T_{\text{end}}} \theta_t$$

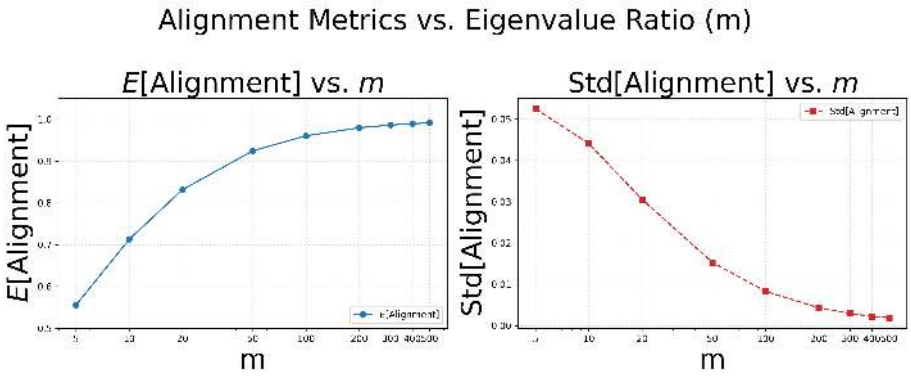


Figure 3: $\mathbb{E}[\text{Alignment}]$ and Std[Alignment] vs $m = \lambda_k/\lambda_{k+1}$

to estimate $\mathbb{E}[\theta_\infty]$. We plotted the trend of $\mathbb{E}[\text{Alignment}]$ and recorded the variance of this statistic as functions of m, the latter reflecting the trend in the volatility of θ_t in the second phase, as shown in Figure 3. We also performed tests with different random seeds, the detail results can be found in Appendix B. It can be observed that as m gradually increases, $\mathbb{E}[\text{Alignment}]$ gradually approaches 1 with fewer fluctuations, which is consistent with the result of Theorem 16. Besides the primary numerical simulation results discussed above, the AppendixB contains additional simulation analyses, including results on the decay rate in the first phase and a discussion of the order in m at which θ_∞ asymptotically approaches 1. The code for our simulation experiments is available at link.

8. Conclusion

Combining Sections 4, 5, and 6, this work points out that for ill-conditioned loss structures, when optimized using SGD, the gradient dynamics exhibit two distinct alignment trends depending on the choice of step size. In the early stages of optimization, following sufficiently large initialization, if the step size η_t is relatively large compared to the norm of the optimization variable $\boldsymbol{x}_t$, the gradient gradually aligns with the bulk space. However, as the optimization progresses and the norm of the optimization variable $\boldsymbol{x}_t$ decreases, entering a small-scale regime, if η_t does not keep pace with the decay of the norm of $\boldsymbol{x}_t$, the gradient gradually aligns with the dominant space (Theorems 2, 3, and 4). However, this alignment does not imply that the dominant direction of the gradient is beneficial to loss reduction. On the contrary, the direction that truly facilitates loss reduction remains the bulk space (Theorems 10, 11, and 12). Using the "river-valley" structure analogy (Wen et al., 2024), each step of SGD's update has a critical step size η_t^*. When $\boldsymbol{x}_t$ is sufficiently far from the "river" or the update step size is smaller than η_t^*, $\boldsymbol{x}_t$ moves closer to the "river." When $\boldsymbol{x}_t$ is sufficiently close to the "river" and the update step size is larger than η_t^*, $\boldsymbol{x}_t$ moves toward the "valley." This critical step size η_t^* gradually decreases and converges to approximately lower than $\approx \dfrac{1}{\lambda_1 + \lambda_d}$ (Theorem 6). This suggests that to track the "river" during the optimization process, the step size should be set to less than $\approx \dfrac{1}{\lambda_1 + \lambda_d}$. However, to ensure sufficiently fast movement in the direction of the "river," the step size needs to be sufficiently large. This indicates that to balance tracking the "river" and maintaining fast optimization progress, the step size should be at most

$\approx \frac{1}{\lambda_1 + \lambda_d}$. This proves that for SGD, tracking the "river" while preserving optimization efficiency is constrained by the ill-conditioned structure. This provides an explanation for SGD to maintain optimization effectiveness while ensuring efficiency. To achieve better optimization efficiency in such cases, further improvements to the descent direction are necessary, such as those found in many existing preconditioning methods, including methods that approximate the inverse Hessian or assign step sizes to individual parameters (Kingma, 2014; Gupta et al., 2018; Yao et al., 2021; Song et al., 2025). The analysis of the relationship between step size selection and alignment for these preconditioning methods can be left as a direction for future research.

Acknowledgments

We thank our colleagues and funding agencies. This work is supported by DOE under Award Number DE-SC0025584, Dartmouth College, and Lambda AI.

References

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Martin Gauch, Maximilian Beck, Thomas Adler, Dmytro Kotsur, Stefan Fiel, Hamid Eghbal-zadeh, Johannes Brandstetter, Johannes Kofler, Markus Holzleitner, Werner Zellinger, et al. Few-shot learning by dimensionality reduction in gradient space. In *Conference on Lifelong Learning Agents*, pages 1043–1064. PMLR, 2022.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. *arXiv preprint arXiv:1901.10159*, 2019. Published in International Conference on Machine Learning (ICML).

Frithjof Gressmann, Zach Eaton-Rosen, and Carlo Luschi. Improving neural network training in low dimensional random bases. *Advances in Neural Information Processing Systems*, 33:12140–12150, 2020.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Liam Hodgkinson, Zhichao Wang, and Michael W Mahoney. Models of heavy-tailed mechanistic universality. *arXiv preprint arXiv:2506.03470*, 2025.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Tao Li, Lei Tan, Zhehao Huang, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional trajectory hypothesis is true: Dnns can be trained in tiny subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3411–3420, 2022a.

Zhiyuan Li et al. Understanding gradient descent on edge of stability in deep learning. *arXiv preprint arXiv:2205.09745*, 2022b.

Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34:20104–20117, 2021.

Rui Pan, Haishan Ye, and Tong Zhang. Eigencurve: Optimal learning rate schedule for sgd on quadratic objectives with skewed hessian spectrums. *arXiv preprint arXiv:2110.14109*, 2021.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. *arXiv preprint arXiv:1901.08244*, 2019.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in training pinns: A loss landscape perspective. *arXiv preprint arXiv:2402.01868*, 2024.

Levent Sagun, Léon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Jan Schneider, Pierre Schumacher, Simon Guist, Le Chen, Daniel Häufle, Bernhard Schölkopf, and Dieter Büchler. Identifying policy gradient subspaces. *arXiv preprint arXiv:2401.06604*, 2024.

Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. *Advances in Neural Information Processing Systems*, 34: 23914–23927, 2021.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does sgd really happen in tiny subspaces? *arXiv preprint arXiv:2405.16002*, 2024.

Minhak Song, Beomhan Baek, Kwangjun Ahn, and Chulhee Yun. Through the river: Understanding the benefit of schedule-free methods for language model training. *arXiv preprint arXiv:2507.09846*, 2025.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*, 2024.

Yikai Wu, Xingyu Zhu, Chenwei Wu, Annie Wang, and Rong Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.

Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. Taxonomizing local versus global structure in neural network loss landscapes. *Advances in Neural Information Processing Systems*, 34:18722–18733, 2021.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael W. Mahoney. Adahessian: An adaptive second order optimizer for machine learning, 2021. URL https://arxiv.org/abs/2006.00719.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Appendix A. Proof of Theorems

A.1. Preliminaries, Notation, and Assumptions (for reference)

Spectral notation and projections. We recall $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_d]$ orthonormal and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ strictly decreasing:

$$\lambda_1 \geq \cdots \geq \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_d > 0.$$

Any $\boldsymbol{x}_t \in R^d$ is expanded in the eigenbasis as

$$\boldsymbol{x}_t = \sum_{i=1}^{d} c_{i,t}\, \boldsymbol{u}_i, \qquad c_{i,t} := \langle \boldsymbol{x}_t, \boldsymbol{u}_i \rangle \in \mathbb{R}.$$

For a dominant–bulk partition $\mathcal{D} = \{1, \dots, k\}$, $\mathcal{B} = \{k+1, \dots, d\}$, define

$$\boldsymbol{P}^{\mathcal{D}} := \sum_{i \in \mathcal{D}} \boldsymbol{u}_i \boldsymbol{u}_i^\top, \qquad \boldsymbol{P}^{\mathcal{B}} := \sum_{i \in \mathcal{B}} \boldsymbol{u}_i \boldsymbol{u}_i^\top,$$

and, for $p \in \mathbb{N}$,

$$\boldsymbol{P}^{\mathcal{S}}_{\lambda^p} := \sum_{i \in \mathcal{S}} \lambda_i^p\, \boldsymbol{u}_i \boldsymbol{u}_i^\top \qquad (\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}).$$

Alignment and subspace notion. The squared alignment function $\theta(\boldsymbol{x}_t)$ of the gradient $\boldsymbol{A}\boldsymbol{x}_t$ with the dominant subspace at time t is defined as

$$\theta(\boldsymbol{x}_t) := \begin{cases} \dfrac{\|\boldsymbol{P}^{\mathcal{D}} \nabla L(\boldsymbol{x}_t)\|_2^2}{\|\nabla L(\boldsymbol{x}_t)\|_2^2} = \dfrac{\|\boldsymbol{P}^{\mathcal{D}} \boldsymbol{A}\boldsymbol{x}_t\|_2^2}{\|\boldsymbol{A}\boldsymbol{x}_t\|_2^2} = \dfrac{\sum_{i \in \mathcal{D}} \lambda_i^2 c_{i,t}^2}{\sum_{i=1}^{d} \lambda_i^2 c_{i,t}^2} \in [0,1] & \text{if } \boldsymbol{x}_t \neq \boldsymbol{0}, \\ 0 & \text{if } \boldsymbol{x}_t = \boldsymbol{0}, \end{cases}$$

The convention $\theta(\boldsymbol{0}) = 0$ ensuring well-definedness in stochastic analysis).We will use θ_t for short in the later section. Define the subspace quantities as follow:

$$\psi_{\mathcal{D}} := \sum_{i \in \mathcal{D}} \lambda_i^2, \quad \psi_{\mathcal{B}} := \sum_{i \in \mathcal{B}} \lambda_i^2,$$

$$s_t^{\mathcal{D}} := \sum_{i \in \mathcal{D}} \lambda_i^2 c_{i,t}^2, \quad s_t^{\mathcal{B}} := \sum_{i \in \mathcal{B}} \lambda_i^2 c_{i,t}^2, \quad s_t := s_t^{\mathcal{D}} + s_t^{\mathcal{B}} = \sum_{i=1}^{d} \lambda_i^2 c_{i,t}^2. \tag{9}$$

and the spectral gaps is defined as:

$$\mathrm{gap}_1 := \lambda_k - \lambda_{k+1} > 0, \qquad \mathrm{gap}_2 := \lambda_k^2 - \lambda_{k+1}^2 = (\lambda_k - \lambda_{k+1})(\lambda_k + \lambda_{k+1}) > 0.$$

Noise in the eigenbasis. The SGD update is

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t), \qquad \boldsymbol{\xi}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top \succeq 0, \quad \{\boldsymbol{\xi}_t\} \text{ i.i.d.}$$

Let $\boldsymbol{C} := \boldsymbol{U}^\top \boldsymbol{\Sigma} \boldsymbol{U}$ and define per-direction noise variances

$$\kappa_i^2 := (C)_{ii} = \boldsymbol{u}_i^\top \boldsymbol{\Sigma} \boldsymbol{u}_i \geq 0, \qquad s_{\min} := \lambda_{\min}(\boldsymbol{\Sigma}), \;\; s_{\max} := \lambda_{\max}(\boldsymbol{\Sigma}).$$

Set the block-wise noise energy be:

$$e_{\mathcal{D}} := \sum_{i \in \mathcal{D}} \lambda_i^2 \kappa_i^2, \qquad e_{\mathcal{B}} := \sum_{i \in \mathcal{B}} \lambda_i^2 \kappa_i^2,$$

and recall $\psi_{\mathcal{S}} := \sum_{i \in \mathcal{S}} \lambda_i^2$ for $\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}$. Then we have:

$$s_{\min}\psi_{\mathcal{D}} \leq e_{\mathcal{D}} \leq s_{\max}\psi_{\mathcal{D}}, s_{\min}\psi_{\mathcal{B}} \leq e_{\mathcal{B}} \leq s_{\max}\psi_{\mathcal{B}}.$$

And our assumptions are:

Assumption *1 (**Asymptotic Spectral Assumptions**) When we consider the high-dimensional regime, where both d and $k(d) \to \infty$, we assume the following conditions, summarized in Table 1:*

Table 3: Asymptotic Spectral Assumptions

Assumption	**Description**
Trajectory boundedness	The state remains bounded: $\sup_t \limsup_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} c_{i,t}^2 < \infty$.
Block proportion	The subspace dimension ratio is a fixed constant: $\rho := \frac{k}{d-k} \in (0, \infty)$. Which implies $k = \frac{\rho}{1+\rho} d$.
Block spectral moments	For $p \in \{2, 3, 4, 6, 8\}$, the block-wise spectral moments converge: $\frac{1}{k} \sum_{i \in \mathcal{D}} \lambda_i^p \to \lambda_{\mathcal{D},p} \in (0, \infty)$, $\frac{1}{d-k} \sum_{i \in \mathcal{B}} \lambda_i^p \to \lambda_{\mathcal{B},p} \in [0, \infty)$.
Noise spectral bounds	The noise covariance has a bounded trace: $Tr(\boldsymbol{\Sigma}) \in (0, +\infty)$

A.2. Next step alignment function

Exact one-step alignment transform. The SGD update projected onto the eigenbasis gives the evolution of the coefficients $c_{i,t}$:

$$c_{i,t+1} = \langle \boldsymbol{x}_{t+1}, \boldsymbol{u}_i \rangle = \langle \boldsymbol{x}_t - \eta_t(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t), \boldsymbol{u}_i \rangle = c_{i,t} - \eta_t(\lambda_i c_{i,t} + \zeta_{i,t}) = (1 - \eta_t \lambda_i) c_{i,t} - \eta_t \zeta_{i,t},$$

where $\zeta_{i,t} := \langle \boldsymbol{\xi}_t, \boldsymbol{u}_i \rangle$. At time $t+1$,$s_{t+1}^{\mathcal{S}} = \sum_{i \in \mathcal{S}} \lambda_i^2 c_{i,t+1}^2$. Thus,

$$\theta_{t+1} = \frac{\sum_{i \in \mathcal{D}} \lambda_i^2 \big((1 - \eta_t \lambda_i) c_{i,t} - \eta_t \zeta_{i,t}\big)^2}{\sum_{i=1}^{d} \lambda_i^2 \big((1 - \eta_t \lambda_i) c_{i,t} - \eta_t \zeta_{i,t}\big)^2} = \frac{1}{1 + \frac{s_{t+1}^{\mathcal{B}}}{s_{t+1}^{\mathcal{D}}}}, \tag{10}$$

where

$$s_{t+1}^{\mathcal{S}} := \sum_{i\in\mathcal{S}} \lambda_i^2\big((1-\eta_t\lambda_i)c_{i,t} - \eta_t\zeta_{i,t}\big)^2 \qquad (\mathcal{S}\in\{\mathcal{D},\mathcal{B}\}). \tag{11}$$

Remark. When we fix a $\boldsymbol{x}_t$, we can reasonably assume that $s_{t+1}\neq 0, s_{t+1}^{\mathcal{D}}\neq 0$, since the case that $\{\xi|s_{t+1}(\xi)=0\}, \{\xi|s_{t+1}^{\mathcal{D}}(\xi)=0\}$ are zero measure set for ξ.
Notice that,when we fix $\boldsymbol{x}_t$ (such as we consider the conditional expectation $\mathbb{E}[\cdot|\boldsymbol{x}_t]$ later). $\theta_{t+1}, s_{t+1} = \sum_{i=1}^d \lambda_i^2\big((1-\eta_t\lambda_i)c_{i,t} - \eta_t\zeta_{i,t}\big)^2$ is a function of ξ_t. Let $\boldsymbol{x}_t'$ be a vector that:

$$\forall i\in\mathcal{D}, <\boldsymbol{x}_t', \boldsymbol{u}_i>= \frac{(1-\eta_t\lambda_i)c_{i,t}}{\eta_t}$$

If $\xi_t = \boldsymbol{x}_t'$, $s_{t+1}^{D}=0$. Since $\mathcal{V}=\{\xi_t|\forall i\in\mathcal{D}, <\xi_t,\boldsymbol{u}_i>= \frac{(1-\eta_t\lambda_i)c_{i,t}}{\eta_t}\}$ is a $d-k$ dim hyperplane, and since ξ is a d dim gaussian. By *Sard* Theorem, $\mathcal{V}$ is a zero measure set for ξ. Since θ_{t+1} is a bounded function, then θ_{t+1} is $\xi-$ *integrable* ($\mathbb{E}_\xi[\theta_{t+1}]<\infty$). Then:

$$\mathbb{E}_\xi[\theta_{t+1}] = \int_{\mathbb{R}^d}\theta_{t+1}d\xi = \int_{\mathbb{R}^d\setminus\mathcal{V}}\theta_{t+1}d\xi + \int_{\mathcal{V}}\theta_{t+1}d\xi = \int_{\mathbb{R}^d\setminus\mathcal{V}}\theta_{t+1}d\xi.$$

Therefore, we can assume $s_{t+1}^{\mathcal{D}}\neq 0$,and it does not affect the expectation on θ_{t+1}. The case of $\{\xi|s_{t+1}^{\mathcal{D}}(\xi)=0\}$ is analogous.

Comparison functional and its meaning. To more conveniently obtain our conclusions in the asymptotic setting, we define the comparison functional:

$$f_t(\eta_t) := s_t^{\mathcal{B}}\, s_{t+1}^{\mathcal{D}} - s_{t+1}^{\mathcal{B}}\, s_t^{\mathcal{D}}, \tag{12}$$

so that

$$\theta_{t+1}>\theta_t \iff f_t(\eta_t)>0, \qquad \theta_{t+1}<\theta_t \iff f_t(\eta_t)<0.$$

The sign of $f_t(\eta_t)$ will be controlled via its conditional expectation given $\boldsymbol{x}_t$.

A.3. Probabilistic Lemmas

Lemma 17 (Variance of Gaussian Forms) *Let $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{C})$ with $\boldsymbol{C}=\boldsymbol{C}^\top\succeq 0$ and $\|\boldsymbol{C}\|_2<\infty$. For any deterministic vector $\boldsymbol{g}\in R^d$ and any deterministic diagonal matrix $\boldsymbol{D}=\mathrm{diag}(d_1,\ldots,d_d)$, the following standard results hold:*

$$\mathrm{Var}(\boldsymbol{g}^\top\boldsymbol{z}) = \boldsymbol{g}^\top\boldsymbol{C}\boldsymbol{g} \le \|\boldsymbol{C}\|_2\,\|\boldsymbol{g}\|_2^2,$$
$$\mathrm{Var}(\boldsymbol{z}^\top\boldsymbol{D}\boldsymbol{z}) = 2\,\mathrm{tr}((\boldsymbol{C}\boldsymbol{D})^2) \le 2\|\boldsymbol{C}\|_2^2\,\mathrm{tr}(\boldsymbol{D}^2).$$

Lemma 18 (A Weak Law of Large Numbers for Block Averages) *Let $\{\mathcal{S}_d\}_{d\in\mathbb{N}}$ be a sequence of index sets such that $|\mathcal{S}_d|\to\infty$ as $d\to\infty$. For each d, let $\{y_i\}_{i\in\mathcal{S}_d}$ be a collection of scalar random variables satisfying $\mathrm{Var}(\sum_{i\in\mathcal{S}_d}y_i)=O(|\mathcal{S}_d|)$. Then*

$$\frac{1}{|\mathcal{S}_d|}\sum_{i\in\mathcal{S}_d}y_i - \mathbb{E}\left[\frac{1}{|\mathcal{S}_d|}\sum_{i\in\mathcal{S}_d}y_i\right] \xrightarrow[d\to\infty]{p} 0.$$

Furthermore, if $\lim_{d\to\infty}\frac{1}{|\mathcal{S}_d|}\sum_{i\in\mathcal{S}_d}\mathbb{E}[y_i]=\ell$ for some constant ℓ, then $\frac{1}{|\mathcal{S}_d|}\sum_{i\in\mathcal{S}_d}y_i \xrightarrow[d\to\infty]{p}\ell$.

Proof Let $s_d = \sum_{i \in \mathcal{S}_d} y_i$ and $\bar{s}_d = \frac{s_d}{|\mathcal{S}_d|}$. We want to show that for any $\epsilon > 0$, $\lim_{d\to\infty} P(|\bar{s}_d - \mathbb{E}[\bar{s}_d]| \geq \epsilon) = 0$. By Chebyshev's inequality,

$$P(|\bar{s}_d - \mathbb{E}[\bar{s}_d]| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{s}_d)}{\epsilon^2}.$$

We analyze the variance term:

$$\mathrm{Var}(\bar{s}_d) = \mathrm{Var}\left(\frac{1}{|\mathcal{S}_d|}\sum_{i\in\mathcal{S}_d} y_i\right) = \frac{1}{|\mathcal{S}_d|^2}\mathrm{Var}\left(\sum_{i\in\mathcal{S}_d} y_i\right).$$

By assumption, there exists a constant $k_0 < \infty$ such that $\mathrm{Var}(\sum_{i\in\mathcal{S}_d} y_i) \leq k_0|\mathcal{S}_d|$. Substituting this into the inequality gives

$$P(|\bar{s}_d - \mathbb{E}[\bar{s}_d]| \geq \epsilon) \leq \frac{k_0|\mathcal{S}_d|}{|\mathcal{S}_d|^2\epsilon^2} = \frac{k_0}{|\mathcal{S}_d|\epsilon^2}.$$

As $d \to \infty$, we have $|\mathcal{S}_d| \to \infty$, so $\frac{k_0}{|\mathcal{S}_d|\epsilon^2} \to 0$. This proves convergence in probability. The second part of the lemma follows directly. ■

Lemma 19 (Continuous Mapping Theorem with Dominated Convergence) *Let $\{y_n\}_{n\in\mathbb{N}}$ be a sequence of scalar random variables such that $y_n \xrightarrow[n\to\infty]{p} y_0$, where y_0 is a constant. Let f be a function that is continuous at y_0 and bounded. Then $\lim_{n\to\infty}\mathbb{E}[f(y_n)] = f(y_0)$.*

Remarks. The variance formulas in Lemma 17 are standard results for Gaussian distributions. The weak law in Lemma 18 is proven directly from Chebyshev's inequality. Lemma 19 is a standard result combining the Continuous Mapping Theorem with the Bounded Convergence Theorem.

A.4. Asymptotic Property

In this section, we need to introduce some important asymptotic properties of variables under our asymptotic region settings. In this section, all probability measures p we consider are conditional measures based on $\boldsymbol{x}_t$, defined as $p(\cdot|\boldsymbol{x}_t)$.

Lemma 20 (Asymptotic ratio of expectations) *Under Assumption 1, the normalized block sums converge in probability to their conditional expectations:*

$$\frac{1}{k}\, s^{\mathcal{D}}_{t+1} \xrightarrow[d\to\infty]{p} \ell'_{\mathcal{D}}(\boldsymbol{x}_t, \boldsymbol{\Sigma}), \qquad \frac{1}{d-k}\, s^{\mathcal{B}}_{t+1} \xrightarrow[d\to\infty]{p} \ell'_{\mathcal{B}}(\boldsymbol{x}_t, \boldsymbol{\Sigma}),$$

where $\ell'_{\mathcal{S}}(\boldsymbol{x}_t, \boldsymbol{\Sigma}) = \lim_{d\to\infty}\frac{1}{|\mathcal{S}|}\mathbb{E}[s^{\mathcal{S}}_{t+1} \mid \boldsymbol{x}_t]$ are finite constants (or functions depend on $\boldsymbol{x}_t, \boldsymbol{\Sigma}$). Consequently, the limit of the expected alignment is the ratio of the limits of expected energies:

$$\lim_{d\to\infty}\mathbb{E}\big[\theta_{t+1}\,\big|\,\boldsymbol{x}_t\big] = \lim_{d\to\infty}\mathbb{E}\Big[\frac{1}{1+\frac{1}{\rho}\rho\frac{s^{\mathcal{B}}_{t+1}}{s^{\mathcal{D}}_{t+1}}}\Big] = \frac{1}{1+\frac{1}{\rho}\frac{\lim_{d\to\infty}\mathbb{E}[\frac{1}{d-k}s^{\mathcal{B}}_{t+1}|\boldsymbol{x}_t]}{\lim_{d\to\infty}\mathbb{E}[\frac{1}{k}s^{\mathcal{D}}_{t+1}|\boldsymbol{x}_t]}} = \frac{1}{1+\frac{\ell'_{\mathcal{B}}(\boldsymbol{x}_t,\boldsymbol{\Sigma})}{\rho\,\ell'_{\mathcal{D}}(\boldsymbol{x}_t,\boldsymbol{\Sigma})}}. \quad (13)$$

Proof We analyze the structure of the scalar sum $s_{t+1}^{\mathcal{S}}$. By expanding its definition, we can decompose it into terms that are constant, linear, and quadratic with respect to the noise vector $\boldsymbol{z}_t$ (whose components are $\zeta_{i,t}$).

$$s_{t+1}^{\mathcal{S}} = \chi_{\mathcal{S},t} + \boldsymbol{g}_{\mathcal{S},t}^{\top}\boldsymbol{z}_t + \boldsymbol{z}_t^{\top} D_{\mathcal{S},t}\boldsymbol{z}_t,$$

where the scalar $\chi_{\mathcal{S},t}$ (constant given $\boldsymbol{x}_t$), vector $\boldsymbol{g}_{\mathcal{S},t}$, and diagonal matrix $D_{\mathcal{S},t}$ are defined as:

$$\chi_{\mathcal{S},t} := \sum_{i\in\mathcal{S}} \lambda_i^2(1-\eta_t\lambda_i)^2 c_{i,t}^2,$$
$$(\boldsymbol{g}_{\mathcal{S},t})_i := \begin{cases} -2\eta_t\lambda_i^2(1-\eta_t\lambda_i)c_{i,t} & \text{if } i\in\mathcal{S} \\ 0 & \text{if } i\notin\mathcal{S} \end{cases},$$
$$(D_{\mathcal{S},t})_{ii} := \begin{cases} \eta_t^2\lambda_i^2 & \text{if } i\in\mathcal{S} \\ 0 & \text{if } i\notin\mathcal{S} \end{cases}.$$

The variance of $s_{t+1}^{\mathcal{S}}$ is the variance of the sum of the linear and quadratic forms. By Lemma 17 and Assumption 1, $\mathrm{Var}(s_{t+1}^{\mathcal{S}}) = O(|\mathcal{S}|)$, which satisfies the conditions for Lemma 18. This gives the convergence in probability. Let the scalar random variable $y_{k,d}$ be the ratio

$$y_{k,d} := \frac{s_{t+1}^{\mathcal{B}}}{s_{t+1}^{\mathcal{D}}} = \frac{\frac{1}{d-k}s_{t+1}^{\mathcal{B}}}{\frac{k}{d-k}\cdot\frac{1}{k}s_{t+1}^{\mathcal{D}}} \xrightarrow[k,d\to\infty]{p} \frac{\ell'_{\mathcal{B}}(\boldsymbol{x}_t,\boldsymbol{\Sigma})}{\rho\,\ell'_{\mathcal{D}}(\boldsymbol{x}_t,\boldsymbol{\Sigma})} =: y_0 \in (0,\infty).$$

With $f(y) = (1+y)^{-1}$ being bounded and continuous, Lemma 19 yields

$$\lim_{d\to\infty}\mathbb{E}[\theta_{t+1}\mid\boldsymbol{x}_t] = \lim_{d\to\infty}\mathbb{E}\Big[\frac{1}{1+y_{k,d}}\Big] = \frac{1}{1+y_0} = \frac{1}{1+\frac{\ell'_{\mathcal{B}}(\boldsymbol{x}_t,\boldsymbol{\Sigma})}{\rho\,\ell'_{\mathcal{D}}(\boldsymbol{x}_t,\boldsymbol{\Sigma})}}.$$

The last identity in (13) follows from the definition of $\ell'_{\mathcal{S}}(\boldsymbol{x}_t,\Sigma)$. ■

Lemma 20 justifies that, asymptotically, the expectation can be interchanged with the ratio defining θ_{t+1}. Which implies if we want to have a decrease for θ_t asymptotically, we have:

$$\begin{aligned} &\lim_{d\to\infty}\mathbb{E}\big[\theta_{t+1}\,\big|\,\boldsymbol{x}_t\big] = \frac{1}{1+\frac{1}{\rho}\frac{\lim_{d\to\infty}\mathbb{E}[\frac{1}{d-k}s_{t+1}^{\mathcal{B}}|\boldsymbol{x}_t]}{\lim_{d\to\infty}\mathbb{E}[\frac{1}{k}s_{t+1}^{\mathcal{D}}|\boldsymbol{x}_t]}} \quad < \lim_{d\to\infty}\theta_t = \frac{1}{1-\frac{1}{\rho}\frac{\lim_{d\to\infty}\frac{1}{d-k}s_t^{\mathcal{B}}}{\lim_{d\to\infty}\frac{1}{k}s_t^{\mathcal{D}}}} \\ \Longrightarrow\ &\lim_{d\to\infty}E\big[s_t^{\mathcal{B}}\,s_{t+1}^{\mathcal{D}}\,\big|\,\boldsymbol{x}_t\big] - \lim_{d\to\infty}E\big[s_{t+1}^{\mathcal{B}}\,s_t^{\mathcal{D}}\,\big|\,\boldsymbol{x}_t\big] < 0 \\ \Longrightarrow\ &\lim_{d\to\infty}E\big[f_t(\eta_t) := s_t^{\mathcal{B}}\,s_{t+1}^{\mathcal{D}} - s_{t+1}^{\mathcal{B}}\,s_t^{\mathcal{D}}\,\big|\,\boldsymbol{x}_t\big] < 0 \end{aligned} \tag{14}$$

Therefore, to decide the sign of the drift of θ_t in expectation, it suffices to analyze the sign of the comparison functional's expectation $\mathbb{E}[f_t(\eta_t)\mid\boldsymbol{x}_t]$ (a quadratic in η_t). This motivates computing elementwise expectations next.

A.5. Comparison Lemma

Lemma 21 (Elementwise expectation of $\boldsymbol{x}_{t+1}$) *For each $i \in \{1, \dots, d\}$ and for a given $\boldsymbol{x}_t$,*

$$\mathbb{E}\big[\lambda_i^2 c_{i,t+1}^2 \mid \boldsymbol{x}_t\big] = (1-\eta_t\lambda_i)^2\lambda_i^2 c_{i,t}^2 + \eta_t^2\lambda_i^2\,\kappa_i^2.$$

Consequently,

$$\mathbb{E}\big[s_{t+1}^{\mathcal{D}} \mid \boldsymbol{x}_t\big] = s_t^{\mathcal{D}} - 2\eta_t \sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2 + \eta_t^2\Big(\sum_{i\in\mathcal{D}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{D}}\Big), \tag{15}$$

$$\mathbb{E}\big[s_{t+1}^{\mathcal{B}} \mid \boldsymbol{x}_t\big] = s_t^{\mathcal{B}} - 2\eta_t \sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2 + \eta_t^2\Big(\sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{B}}\Big). \tag{16}$$

Proof We expand the square for $c_{i,t+1}$ and take the expectation conditional on $\boldsymbol{x}_t$:

$$\mathbb{E}\big[\lambda_i^2 c_{i,t+1}^2 \mid \boldsymbol{x}_t\big] = \mathbb{E}\big[\lambda_i^2\left((1-\eta_t\lambda_i)c_{i,t} - \eta_t\zeta_{i,t}\right)^2 \mid \boldsymbol{x}_t\big].$$

Expanding the squared term gives

$$\mathbb{E}\big[\lambda_i^2\left((1-\eta_t\lambda_i)^2 c_{i,t}^2 - 2\eta_t(1-\eta_t\lambda_i)c_{i,t}\zeta_{i,t} + \eta_t^2\zeta_{i,t}^2\right) \mid \boldsymbol{x}_t\big].$$

Using $\mathbb{E}[\zeta_{i,t} \mid \boldsymbol{x}_t] = 0$ and $\mathbb{E}[\zeta_{i,t}^2 \mid \boldsymbol{x}_t] = \kappa_i^2$, the cross-term vanishes. We are left with

$$\mathbb{E}\big[\lambda_i^2 c_{i,t+1}^2 \mid \boldsymbol{x}_t\big] = \lambda_i^2(1-\eta_t\lambda_i)^2 c_{i,t}^2 + \eta_t^2\lambda_i^2\kappa_i^2 \tag{17}$$

$$= (1 - 2\eta_t\lambda_i + \eta_t^2\lambda_i^2)\lambda_i^2 c_{i,t}^2 + \eta_t^2\lambda_i^2\kappa_i^2. \tag{18}$$

Summing over $i \in \mathcal{D}$ (or $\mathcal{B}$) and regrouping terms yields

$$\mathbb{E}\big[s_{t+1}^{\mathcal{S}} \mid \boldsymbol{x}_t\big] = \sum_{i\in\mathcal{S}} \lambda_i^2 c_{i,t}^2 - 2\eta_t \sum_{i\in\mathcal{S}} \lambda_i^3 c_{i,t}^2 + \eta_t^2\left(\sum_{i\in\mathcal{S}} \lambda_i^4 c_{i,t}^2 + \sum_{i\in\mathcal{S}} \lambda_i^2\kappa_i^2\right).$$

This simplifies to the expressions in (15) and (16) using the definitions of $s_t^{\mathcal{S}}$ and $e_{\mathcal{S}}$. ■

Lemma 22 (Quadratic form for step size condition) *With $f_t(\eta_t)$ as in (12), for a given $\boldsymbol{x}_t$, its conditional expectation is a quadratic function of η_t with no constant term:*

$$\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] = p_t\,\eta_t^2 + q_t\,\eta_t, \tag{19}$$

where the scalars p_t and q_t are defined as

$$q_t = 2\Big(s_t^{\mathcal{D}} \sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2 - s_t^{\mathcal{B}} \sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2\Big) \le 2 s_t^{\mathcal{B}} s_t^{\mathcal{D}}(\lambda_{k+1} - \lambda_k) < 0, \tag{20}$$

$$p_t = s_t^{\mathcal{B}}\Big(\sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 + e_{\mathcal{D}}\Big) \;-\; s_t^{\mathcal{D}}\Big(\sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{B}}\Big). \tag{21}$$

If $p_t \neq 0$, we define the adaptive critical step size as the non-zero root $\eta_t^* := -q_t/p_t$.

Proof We explicitly compute the conditional expectation of $f_t(\eta_t)$ by substituting the results from Lemma 21 into its definition and collecting terms by powers of η_t.

$$\begin{aligned}
\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] &= s_t^{\mathcal{B}}\mathbb{E}[s_{t+1}^{\mathcal{D}} \mid \boldsymbol{x}_t] - s_t^{\mathcal{D}}\mathbb{E}[s_{t+1}^{\mathcal{B}} \mid \boldsymbol{x}_t] \\
&= s_t^{\mathcal{B}}\left(s_t^{\mathcal{D}} - 2\eta_t \sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2 + \eta_t^2\Big(\sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 + e_{\mathcal{D}}\Big)\right) \\
&\quad - s_t^{\mathcal{D}}\left(s_t^{\mathcal{B}} - 2\eta_t \sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2 + \eta_t^2\Big(\sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{B}}\Big)\right) \\
&= (s_t^{\mathcal{B}} s_t^{\mathcal{D}} - s_t^{\mathcal{D}} s_t^{\mathcal{B}}) \\
&\quad + \eta_t\left(-2s_t^{\mathcal{B}} \sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2 + 2s_t^{\mathcal{D}} \sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2\right) \\
&\quad + \eta_t^2\left(s_t^{\mathcal{B}}\Big(\sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 + e_{\mathcal{D}}\Big) - s_t^{\mathcal{D}}\Big(\sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{B}}\Big)\right) \\
&= \eta_t \underbrace{2\Big(s_t^{\mathcal{D}} \sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2 - s_t^{\mathcal{B}} \sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2\Big)}_{q_t} \\
&\quad + \eta_t^2 \underbrace{\left(s_t^{\mathcal{B}}\Big(\sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 + e_{\mathcal{D}}\Big) - s_t^{\mathcal{D}}\Big(\sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{B}}\Big)\right)}_{p_t}.
\end{aligned}$$

This confirms the quadratic form $\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] = p_t\eta_t^2 + q_t\eta_t$. The bound on q_t follows from $\sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2 \le \lambda_{k+1} s_t^{\mathcal{B}}$ and $\sum_{j\in\mathcal{D}} \lambda_j^3 c_{j,t}^2 \ge \lambda_k s_t^{\mathcal{D}}$, which makes q_t strictly negative since $\lambda_k > \lambda_{k+1}$. ■

Remark. The expected one-step alignment change functional $\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t]$ is a quadratic function of the step size η_t that always passes through the origin ($\eta_t = 0, \mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] = 0$). The behavior for any positive step size $\eta_t > 0$ is entirely determined by the sign of the quadratic coefficient p_t, as illustrated in Figure 4.

When $p_t > 0$ (Figure 4a), the parabola opens upwards. Since $q_t < 0$, there exists a positive critical step size $\eta_t^* = -q_t/p_t > 0$. For a small step size $0 < \eta_t < \eta_t^*$, we have $\mathbb{E}[f_t(\eta_t)] < 0$, implying the expected alignment $\mathbb{E}[\theta_{t+1}]$ decreases. For a large step size $\eta_t > \eta_t^*$, we have $\mathbb{E}[f_t(\eta_t)] > 0$, implying the expected alignment increases.

When $p_t < 0$ (Figure 4b), the parabola opens downwards. Since both coefficients p_t and q_t are negative, the functional $\mathbb{E}[f_t(\eta_t)]$ is negative for all positive step sizes $\eta_t > 0$. This implies that, in this regime, any choice of step size will lead to a decrease in the expected alignment.

A.6. Proofs of Step Size Condition Theory

Theorem 2*(**Decreases Condition**) Under Assumption 1, if* $0 < \eta_t < \eta_t^*(\boldsymbol{x}_t)$*, then*

$$\lim_{d\to\infty} \mathbb{E}\big[\theta_{t+1} \,\big|\, \boldsymbol{x}_t\big] \;<\; \theta_t.$$

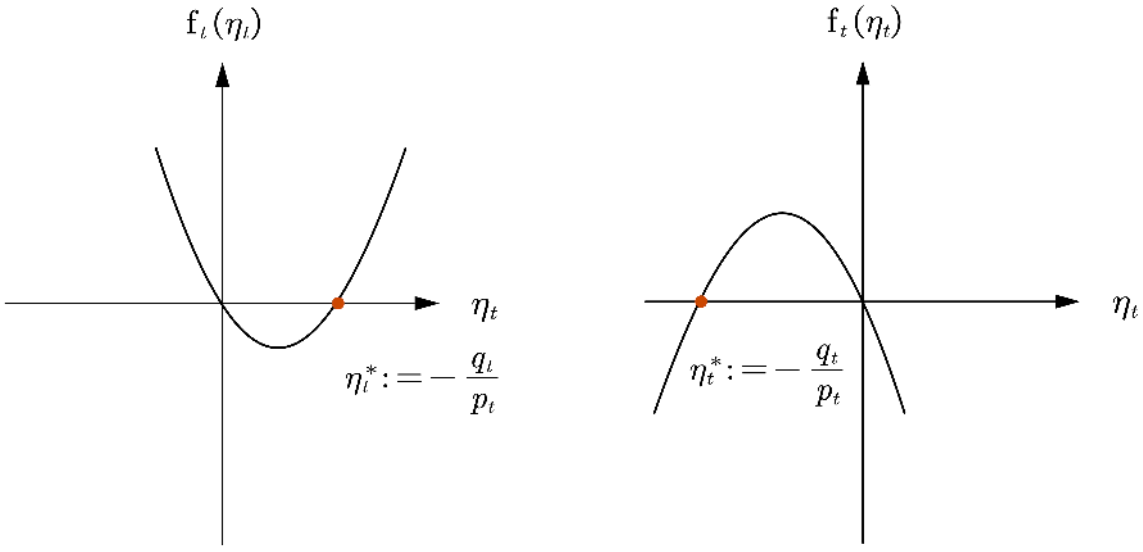


Figure 4: The sign of $\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t]$ as a function of η_t. (a) shows the behavior when $p_t > 0$, where the parabola opens upwards. (b) shows the behavior when $p_t < 0$, where the parabola opens downwards.

Proof By Lemma 22, $\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] = p_t\eta_t^2 + q_t\eta_t$ with $q_t < 0$. If $p_t > 0$, then for $0 < \eta_t < \eta_t^* = -q_t/p_t$, we have $\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] < 0$. Lemma 20 transfers the sign to $\lim_{d\to\infty} \mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] < \theta_t$. ■

Theorem *3(**Increase Condition**) Under Assumption 1, let*

$$g_{\text{gap}} := \frac{1}{1 + \frac{s_{\max}}{s_{\min}} \cdot \frac{1}{\rho}\left(\frac{\lambda_{k+1}}{\lambda_k}\right)^2} \in (0,1).$$

For t and $\boldsymbol{x}_t$. If $\theta_t \le g_{\text{gap}}$ and $\eta_t > \eta_t^*(\boldsymbol{x}_t)$, then

$$\lim_{d\to\infty} \mathbb{E}\big[\theta_{t+1} \,\big|\, \boldsymbol{x}_t\big] \;>\; \theta_t.$$

Proof We will show $p_t > 0$ when $\theta_t \le g_{\text{gap}}$, then apply Lemma 20 and 22.
First, recall the expression for p_t (formula 21). Let's consider the term which do not contain $e_{\mathcal{S}}$ in p_t:

$$\begin{aligned} s_t^{\mathcal{B}} \sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 - s_t^{\mathcal{D}} \sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 &= \sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{D}} \lambda_i^2\lambda_j^2 c_{i,t}^2 c_{j,t}^2 (\lambda_j^2 - \lambda_i^2) \\ &\ge s_t^{\mathcal{B}} s_t^{\mathcal{D}} (\lambda_k^2 - \lambda_{k+1}^2) > 0. \end{aligned} \tag{22}$$

Next, using $e_{\mathcal{D}} \ge s_{\min}\psi_{\mathcal{D}}$ and $e_{\mathcal{B}} \le s_{\max}\psi_{\mathcal{B}}$,

$$s_t^{\mathcal{B}} e_{\mathcal{D}} - s_t^{\mathcal{D}} e_{\mathcal{B}} \;\ge\; s_t\Big((1-\theta_t)s_{\min}\psi_{\mathcal{D}} - \theta_t s_{\max}\psi_{\mathcal{B}}\Big). \tag{23}$$

If $\theta_t \le g_{\text{noise}} := \left(1+\frac{s_{\max}\psi_{\mathcal{B}}}{s_{\min}\psi_{\mathcal{D}}}\right)^{-1}$, then the bracket is non-negative and so the term which do contain $e_{\mathcal{S}}$ is ≥ 0. Using $\psi_{\mathcal{D}} \ge k\lambda_k^2$ and $\psi_{\mathcal{B}} \le (d-k)\lambda_{k+1}^2$,we can have $g_{\text{gap}} \le g_{\text{noise}}$. Hence $p_t > 0$ whenever $\theta_t \le g_{\text{gap}}$. Therefore, if additionally $\eta_t > \eta_t^*$, then $\mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] > 0$ and Lemma 20 yields

$$\lim_{d\to\infty} \mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] > \theta_t.$$

As $\lambda_k/\lambda_{k+1} \to \infty$, $(\lambda_{k+1}/\lambda_k)^2 \to 0$ implies $g_{\text{gap}} \to 1$. ■

Theorem *4(**Large Alignment Regime Condition**) Under Assumption 1, there exists a critical alignment threshold $\theta_t^* \in (0,1)$ such that if the current alignment $\theta_t \ge \theta_t^*$, then for any positive step size $\eta_t > 0$, the expected alignment decreases:*

$$\lim_{d\to\infty} \mathbb{E}\big[\theta_{t+1} \,\big|\, \boldsymbol{x}_t\big] \;<\; \theta_t.$$

The threshold is given by $\theta_t^* := r_{0,t}/(1+r_{0,t})$, where $r_{0,t}$ is the positive root of the quadratic equation:

$$a_t^{\text{aux}} r^2 + (a_t^{\text{aux}} - m_t^{\text{aux}} - h_t^{\text{aux}})r - h_t^{\text{aux}} = 0.$$

The coefficients are defined using the second-order spectral masses $\psi_{\mathcal{S}} := \sum_{i\in\mathcal{S}} \lambda_i^2$ and the total energy $s_t := \sum_{i=1}^d \lambda_i^2 c_{i,t}^2$ as:

$$\begin{aligned} a_t^{\text{aux}} &:= s_{\min}\psi_{\mathcal{B}}, \\ h_t^{\text{aux}} &:= s_{\max}\psi_{\mathcal{D}}, \\ m_t^{\text{aux}} &:= s_t(\lambda_1^2 - \lambda_d^2). \end{aligned}$$

As $\psi_{\mathcal{D}}/\psi_{\mathcal{B}} \to \infty$, we have $\theta_t^* \to 1$.

Proof We derive a sufficient condition for $p_t \le 0$. Using

$$\sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 \le \lambda_1^2 s_t^{\mathcal{D}}, \quad \sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 \ge \lambda_d^2 s_t^{\mathcal{B}}, \quad e_{\mathcal{D}} \le s_{\max}\psi_{\mathcal{D}}, \quad e_{\mathcal{B}} \ge s_{\min}\psi_{\mathcal{B}},$$

we get

$$\begin{aligned} p_t &\le s_t^{\mathcal{B}}(\lambda_1^2 s_t^{\mathcal{D}} + s_{\max}\psi_{\mathcal{D}}) - s_t^{\mathcal{D}}(\lambda_d^2 s_t^{\mathcal{B}} + s_{\min}\psi_{\mathcal{B}}) \\ &= s_t^2\theta_t(1-\theta_t)(\lambda_1^2 - \lambda_d^2) + s_t\big((1-\theta_t)s_{\max}\psi_{\mathcal{D}} - \theta_t s_{\min}\psi_{\mathcal{B}}\big). \end{aligned} \tag{24}$$

Let $r_t = \theta_t/(1-\theta_t)$ and define $a_t^{\text{aux}} := s_{\min}\psi_{\mathcal{B}}$, $h_t^{\text{aux}} := s_{\max}\psi_{\mathcal{D}}$, $m_t^{\text{aux}} := s_t(\lambda_1^2 - \lambda_d^2)$. Then $p_t \le 0$ is implied by $r_t m_t^{\text{aux}} + (1+r_t)h_t^{\text{aux}} - r_t(1+r_t)a_t^{\text{aux}} \le 0$, which is equivalent to

$$a_t^{\text{aux}} r_t^2 + (a_t^{\text{aux}} - m_t^{\text{aux}} - h_t^{\text{aux}})r_t - h_t^{\text{aux}} \;\ge\; 0.$$

Hence, if θ_t is large enough so that the quadratic inequality holds, $p_t \le 0$ and, for any $\eta_t > 0$, combine with Lemma 20:

$$\lim_{d\to\infty} \mathbb{E}[f_t(\eta_t) \mid \boldsymbol{x}_t] \le 0 \;\Rightarrow\; \lim_{d\to\infty} \mathbb{E}[\theta_{t+1} \mid \boldsymbol{x}_t] < \theta_t,$$

■

Theorem *6 (**Separation of Alignment Regimes**) Under Assumption 1, for any nontrivial problem with a non-zero spectral gap* ($\mathrm{gap}_2 > 0$) *and bounded noise* ($s_{\max} < \infty$), *the low-alignment threshold* g_{gap} *is strictly less than the high-alignment threshold* θ_t^**:*

$$g_{\mathrm{gap}} < \theta_t^*.$$

Proof According to the proof of Theorem 3:

$$\theta_t < g_{gap} \implies p_t > 0,$$

if $g_{gap} > \theta_t^*, \exists \theta_t \in (\theta_t^*, g_{gap})$, by the proof of Theorem 4. we have:

$$\theta_t > \theta_t^* \implies p_t < 0$$

which contradicts the previous conclusion. Therefore $g_{gap} < \theta_t^*$ ■

Theorem *5(**Asymptotic rate of** θ_t^*) Under Assumption 1, with* $m := \lambda_k/\lambda_{k+1} > 1$, *there exist constants* $\alpha, \beta > 0$ *such that*

$$1 - \frac{\beta}{m^2} \;\le\; \theta_t^* \;\le\; 1 - \frac{\alpha}{m^2},$$

and hence

$$\theta_t^* = 1 - \Theta\left(\frac{1}{m^2}\right).$$

Proof By Theorem 4 we have

$$\theta_t^* \;=\; \frac{r_{0,t}}{1 + r_{0,t}},$$

where $r_{0,t} > 0$ is the positive root of the quadratic

$$a_t^{\mathrm{aux}}\, r^2 + \left(a_t^{\mathrm{aux}} - m_t^{\mathrm{aux}} - h_t^{\mathrm{aux}}\right) r - h_t^{\mathrm{aux}} = 0.$$

Here

$$a_t^{\mathrm{aux}} = s_{\min} \sum_{i \in \mathcal{B}} \lambda_i^2, \quad h_t^{\mathrm{aux}} = s_{\max} \sum_{i \in \mathcal{D}} \lambda_i^2, \quad m_t^{\mathrm{aux}} = s_t\left(\lambda_1^2 - \lambda_d^2\right),$$

are all nonnegative scalars. Using the quadratic formula, the positive root can be written as

$$r_{0,t} = \frac{m_t^{\mathrm{aux}} + h_t^{\mathrm{aux}} - a_t^{\mathrm{aux}} + \sqrt{\left(m_t^{\mathrm{aux}} + h_t^{\mathrm{aux}} - a_t^{\mathrm{aux}}\right)^2 + 4\, a_t^{\mathrm{aux}}\, h_t^{\mathrm{aux}}}}{2\, a_t^{\mathrm{aux}}}.$$

Under Assumption 1 the eigenvalues satisfy

$$\lambda_{k+1}^2 \le \lambda_i^2 \le \lambda_k^2 \quad \text{for } i \in \mathcal{D},$$

and

$$\lambda_{k+1}^2 \le \lambda_i^2 \quad \text{for } i \in \mathcal{B},$$

so that, with the shorthand

$$\ell_{\mathcal{B}} := |\mathcal{B}| \cdot \lambda_{k+1}^2, \qquad u_{\mathcal{B}} := |\mathcal{B}| \cdot m^2 \lambda_{k+1}^2,$$

we have $\ell_{\mathcal{B}} \le \sum_{i\in\mathcal{B}} \lambda_i^2 \le u_{\mathcal{B}}$ and $\rho\,\ell_{\mathcal{B}} \le \sum_{i\in\mathcal{D}} \lambda_i^2 \le \rho\,u_{\mathcal{B}}$, where $\rho = k/(d-k)$. Using the noise spectral bounds $s_{\min} \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le s_{\max}$, it follows that

$$s_{\min}\,\ell_{\mathcal{B}} \;\le\; a_t^{\text{aux}} \;\le\; s_{\min}\,u_{\mathcal{B}}, \qquad s_{\max}\,\rho\,\ell_{\mathcal{B}} \;\le\; h_t^{\text{aux}} \;\le\; s_{\max}\,\rho\,u_{\mathcal{B}}.$$

Define the fixed scalar $\beta := \rho\,(s_{\max}/s_{\min}) > 0$. Using $m_t^{\text{aux}} \ge 0$, we obtain

$$r_{0,t} \;\ge\; \frac{s_{\max}\,\rho\,\ell_{\mathcal{B}} - s_{\min}\,u_{\mathcal{B}}}{s_{\min}\,u_{\mathcal{B}}} = \frac{\beta}{m^2} - 1,$$

and hence

$$\theta_t^* = \frac{r_{0,t}}{1+r_{0,t}} \;\ge\; \frac{\beta/m^2 - 1}{1 + (\beta/m^2 - 1)} = 1 - \frac{1}{\beta/m^2}.$$

For the upper bound, since $a_t^{\text{aux}} \ge s_{\min}\ell_{\mathcal{B}}$, $h_t^{\text{aux}} \le s_{\max}\rho\,u_{\mathcal{B}}$, and $m_t^{\text{aux}} \le s_t(\lambda_1^2 - \lambda_d^2)$, there exists a fixed scalar $\alpha > 0$ such that

$$r_{0,t} \;\le\; \alpha\,m^2.$$

Then

$$\theta_t^* = \frac{r_{0,t}}{1+r_{0,t}} \;\le\; \frac{\alpha\,m^2}{1+\alpha\,m^2} = 1 - \frac{1}{1+\alpha\,m^2},$$

which gives

$$1 - \frac{\beta}{m^2} \;\le\; \theta_t^* \;\le\; 1 - \frac{\alpha}{m^2}.$$

The $\Theta(1/m^2)$ form follows from these two-sided inequalities. ■

Theorem *7(**State- and gap-aware bounds on** η_t^* **(with** $\|\boldsymbol{x}_t\|_2$**))** Under Assumption 1, for any x_t, we have*

$$\eta_t^* \;\ge\; \frac{2\,\text{gap}_1}{(\lambda_1^2 - \lambda_d^2) \;+\; \dfrac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2\,\|\boldsymbol{x}_t\|_2^2\,\theta_t}},$$

Proof The bounds for $\eta_t^* = -q_t/p_t$ are derived by finding lower and upper bounds for $-q_t$ and p_t. By formula (20), we have:

$$-q_t \ge 2s_t^2\theta_t(1-\theta_t)\text{gap}_1$$

.

And by formula (24), we have:

$$p_t \le s_t^2\theta_t(1-\theta_t)(\lambda_1^2 - \lambda_d^2) + s_t(1-\theta_t)s_{\max}\psi_{\mathcal{D}}$$

. Therefore

$$\eta_t^* = \frac{-q_t}{p_t} \;\ge\; \frac{2\,\text{gap}_1}{(\lambda_1^2 - \lambda_d^2) + \dfrac{s_{\max}\psi_{\mathcal{D}}}{s_t\,\theta_t}}.$$

Using $s_t \ge \lambda_d^2\|\boldsymbol{x}_t\|_2^2$ yields the stated lower bounds.
■

Theorem *8(**State- and gap-aware upper bounds on** η^*) Under Assumption 1, providing that the alignment satisfies $\theta_t \geq \frac{e_{\mathcal{B}}}{e_{\mathcal{B}}+e_{\mathcal{D}}}$, where $e_{\mathcal{B}}, e_{\mathcal{D}}$ represent the noise magnitudes in the principal and tail subspaces respectively, we have:*

$$\eta_t^* \;\leq\; \frac{2(\lambda_1-\lambda_d)}{\lambda_k\lambda_1-\lambda_{k+1}\lambda_d}.$$

Proof Recall that the optimal step size is given by the ratio $\eta_t^* = -q_t/p_t$. We first derive an upper bound for $-q_t$. By definition, and noting that q_t contains no noise terms under the orthogonality assumption, we have:

$$\begin{aligned} -q_t \;&=\; s_t^{\mathcal{B}} \sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2 - s_t^{\mathcal{D}} \sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2 \\ &\leq\; s_t^{\mathcal{B}}\lambda_1 \left(\sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t}^2\right) - s_t^{\mathcal{D}}\lambda_d \left(\sum_{i\in\mathcal{B}} \lambda_i^2 c_{i,t}^2\right) \\ &=\; \lambda_1 s_t^{\mathcal{B}} s_t^{\mathcal{D}} - \lambda_d s_t^{\mathcal{D}} s_t^{\mathcal{B}} \\ &=\; (\lambda_1-\lambda_d) s_t^{\mathcal{D}} s_t^{\mathcal{B}}. \end{aligned} \tag{25}$$

Next, we derive a lower bound for p_t. Expanding the expression for p_t yields:

$$\begin{aligned} p_t \;&=\; s_t^{\mathcal{B}}\Big(\sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 + e_{\mathcal{D}}\Big) - s_t^{\mathcal{D}}\Big(\sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{B}}\Big) \\ &=\; \left(s_t^{\mathcal{B}} \sum_{j\in\mathcal{D}} \lambda_j^4 c_{j,t}^2 - s_t^{\mathcal{D}} \sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2\right) + (s_t^{\mathcal{B}} e_{\mathcal{D}} - s_t^{\mathcal{D}} e_{\mathcal{B}}). \end{aligned} \tag{26}$$

Applying the spectral bounds $\lambda_j \geq \lambda_k$ for $j \in \mathcal{D}$ and $\lambda_i \leq \lambda_{k+1}$ for $i \in \mathcal{B}$ to the signal components:

$$p_t \;\geq\; (\lambda_k\lambda_1 - \lambda_{k+1}\lambda_d) s_t^{\mathcal{D}} s_t^{\mathcal{B}} + (s_t^{\mathcal{B}} e_{\mathcal{D}} - s_t^{\mathcal{D}} e_{\mathcal{B}}). \tag{27}$$

The given condition $\theta_t \geq \frac{e_{\mathcal{B}}}{e_{\mathcal{B}}+e_{\mathcal{D}}}$ implies:

$$\frac{s_t^{\mathcal{D}}}{s_t^{\mathcal{D}}+s_t^{\mathcal{B}}} \geq \frac{e_{\mathcal{B}}}{e_{\mathcal{B}}+e_{\mathcal{D}}} \iff s_t^{\mathcal{D}} e_{\mathcal{D}} + s_t^{\mathcal{D}} e_{\mathcal{B}} \geq s_t^{\mathcal{D}} e_{\mathcal{B}} + s_t^{\mathcal{B}} e_{\mathcal{B}} \iff s_t^{\mathcal{B}} e_{\mathcal{D}} - s_t^{\mathcal{D}} e_{\mathcal{B}} \geq 0.$$

Since the noise residual term is non-negative, we may lower bound p_t by omitting it:

$$p_t \;\geq\; (\lambda_k\lambda_1 - \lambda_{k+1}\lambda_d) s_t^{\mathcal{D}} s_t^{\mathcal{B}}. \tag{28}$$

Substituting the bounds for $-q_t$ and p_t back into the expression for η_t^*:

$$\eta_t^* \;\leq\; \frac{(\lambda_1-\lambda_d) s_t^{\mathcal{D}} s_t^{\mathcal{B}}}{(\lambda_k\lambda_1 - \lambda_{k+1}\lambda_d) s_t^{\mathcal{D}} s_t^{\mathcal{B}}} \;=\; \frac{\lambda_1-\lambda_d}{\lambda_k\lambda_1-\lambda_{k+1}\lambda_d}. \tag{29}$$

■

Corollary *(**The comparison between η_t^* and $\frac{2}{\lambda_1}$**) Suppose the conditions of Theorem 8 hold. We have the following relation between η_t^* and the convergence step size for Quadratic Programming:*

$$\eta_t^* \leq \frac{2}{\lambda_1}.$$

Proof Using the tight bound derived in the proof of Theorem 8:

$$\eta_t^* \leq \frac{(\lambda_1 - \lambda_d)}{\lambda_k \lambda_1 - \lambda_{k+1}\lambda_d}.$$

We first verify $\eta_t^* \leq \frac{1}{\lambda_1}$, it suffices to show:

$$\lambda_1(\lambda_1 - \lambda_d) \leq \lambda_k \lambda_1 - \lambda_{k+1}\lambda_d.$$

Noting that $\lambda_k \leq \lambda_1$, the inequality holds if $\lambda_1^2 - \lambda_1\lambda_d \leq \lambda_1^2 - \lambda_{k+1}\lambda_d$, which simplifies to $\lambda_{k+1} \leq \lambda_1$. This is naturally satisfied by the definition of principal and tail eigenvalues. Therefore:

$$\eta_t^* \leq \frac{1}{\lambda_1} < \frac{2}{\lambda_1}$$

■

A.7. Proofs for Projected SGD Theorem

This section provides the detailed proofs for the theorems and lemma presented in the main text. For clarity, we first recall the central definitions for a given subspace $\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}$:

$$s_t^{\mathcal{S}} := \sum_{i\in\mathcal{S}} \lambda_i^2 c_{i,t}^2, \qquad \tau_t^{\mathcal{S}} := \sum_{i\in\mathcal{S}} \lambda_i^3 c_{i,t}^2, \qquad u_t^{\mathcal{S}} := \sum_{i\in\mathcal{S}} \lambda_i^4 c_{i,t}^2,$$

$$n_{\mathcal{S}}^{\text{loss}} := \sum_{i\in\mathcal{S}} \lambda_i \kappa_i^2, \qquad \mu_t^{\mathcal{S}}(p,q) := \frac{\sum_{i\in\mathcal{S}} \lambda_i^p c_{i,t}^2}{\sum_{i\in\mathcal{S}} \lambda_i^q c_{i,t}^2}, \qquad \theta_t := \frac{s_t^{\mathcal{D}}}{s_t^{\mathcal{D}} + s_t^{\mathcal{B}}}.$$

We begin with the proof of the first theorem, which establishes the one-step expected loss condition.

Theorem *10(**Condition Differences on Different Alignment Regime**) For a given state $\boldsymbol{x}_t$, the one-step expected loss decreases if and only if the step size η_t satisfies the following conditions for the dominant and bulk projected updates, respectively:*

$$\textit{For } \mathcal{D}: \quad \mathbb{E}[L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) \mid \boldsymbol{x}_t] < 0 \iff 0 < \eta_t < \eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) := \frac{2\, s_t^{\mathcal{D}}}{\tau_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}}}$$

$$\textit{For } \mathcal{B}: \quad \mathbb{E}[L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) \mid \boldsymbol{x}_t] < 0 \iff 0 < \eta_t < \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t) := \frac{2\, s_t^{\mathcal{B}}}{\tau_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}}$$

Proof Let the projected stochastic gradient be $\boldsymbol{g}_t^{\mathcal{S}} = \boldsymbol{P}^{\mathcal{S}}(\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t)$. The update rule is $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{g}_t^{\mathcal{S}}$. The one-step change in the loss function $L(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x}$ is:

$$\begin{aligned}\Delta L = L(\boldsymbol{x}_{t+1}) - L(\boldsymbol{x}_t) &= \frac{1}{2}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t^{\mathcal{S}})^\top \boldsymbol{A}(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t^{\mathcal{S}}) - \frac{1}{2}\boldsymbol{x}_t^\top \boldsymbol{A}\boldsymbol{x}_t \\ &= -\eta_t (\boldsymbol{g}_t^{\mathcal{S}})^\top \boldsymbol{A}\boldsymbol{x}_t + \frac{1}{2}\eta_t^2 (\boldsymbol{g}_t^{\mathcal{S}})^\top \boldsymbol{A}\boldsymbol{g}_t^{\mathcal{S}}.\end{aligned}$$

We now take the expectation with respect to the noise $\boldsymbol{\xi}_t$, conditioned on $\boldsymbol{x}_t$. The linear term becomes $\mathbb{E}[(\boldsymbol{g}_t^{\mathcal{S}})^\top \boldsymbol{A}\boldsymbol{x}_t] = (\boldsymbol{P}^{\mathcal{S}}\boldsymbol{A}\boldsymbol{x}_t)^\top A\boldsymbol{x}_t = \boldsymbol{x}_t^\top \boldsymbol{A}\boldsymbol{P}^{\mathcal{S}}\boldsymbol{A}\boldsymbol{x}_t = s_t^{\mathcal{S}}$. The quadratic term becomes $\mathbb{E}[(\boldsymbol{g}_t^{\mathcal{S}})^\top \boldsymbol{A}\boldsymbol{g}_t^{\mathcal{S}}] = (\boldsymbol{P}^{\mathcal{S}}\boldsymbol{A}\boldsymbol{x}_t)^\top \boldsymbol{A}(\boldsymbol{P}^{\mathcal{S}}\boldsymbol{A}\boldsymbol{x}_t) + \mathbb{E}[(\boldsymbol{P}^{\mathcal{S}}\boldsymbol{\xi}_t)^\top \boldsymbol{A}(\boldsymbol{P}^{\mathcal{S}}\boldsymbol{\xi}_t)] = \tau_t^{\mathcal{S}} + n_{\mathcal{S}}^{\text{loss}}$. Combining these, the expected change in loss is $\mathbb{E}[\Delta L \mid \boldsymbol{x}_t] = -\eta_t s_t^{\mathcal{S}} + \frac{1}{2}\eta_t^2(\tau_t^{\mathcal{S}} + n_{\mathcal{S}}^{\text{loss}})$. Setting this to be less than zero holds for $0 < \eta_t < \frac{2s_t^{\mathcal{S}}}{\tau_t^{\mathcal{S}} + n_{\mathcal{S}}^{\text{loss}}}$. ■

The proof of Theorem 11 relies on the fundamental properties of its governing quadratic function, which we first prove in the following lemma.

Additional Notion To give a further analysis of the projected SGD, we need to introduce the following quantities $\alpha(\boldsymbol{x}_t), \beta(\boldsymbol{x}_t), \gamma(\boldsymbol{x}_t)$ which depend on $\boldsymbol{x}_t$. (We will use $\alpha_t, \beta_t, \gamma_t$ later for short.):

$$\alpha_t = s_t\left(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\right) \tag{30}$$

$$\beta_t = -\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}} \tag{31}$$

$$\gamma_t = -n_{\mathcal{D}}^{\text{loss}} \tag{32}$$

Lemma 23 *Let $h(\theta) := \alpha_t\theta^2 + \beta_t\theta + \gamma_t$ be the crossover quadratic, with coefficients as defined above. The following properties hold:*

1. $\forall\, \boldsymbol{x}_t,\ \alpha_t > 0$ *(positive definite in $\boldsymbol{x}_t$),*
2. $\forall\, \boldsymbol{x}_t,\ \gamma_t < 0$ *(negative definite in $\boldsymbol{x}_t$).*

Additionally, $\forall\, \boldsymbol{x}_t$, $h(0) = \gamma_t < 0$ and $h(1) = \alpha_t + \beta_t + \gamma_t > 0$.

Proof First, we prove that the leading coefficient α_t is positive definite. By definition,

$$\alpha_t = s_t\left(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\right).$$

For any $\boldsymbol{x}_t$, $s_t > 0$, so the sign of α_t is determined by the term in the parenthesis. This term is positive if and only if the following inequality holds:

$$\mu_t^{\mathcal{D}}(3,2) > \mu_t^{\mathcal{B}}(3,2).$$

For the dominant subspace, $\forall i \in \mathcal{D}, \lambda_i \geq \lambda_k$, which provides a lower bound:

$$\mu_t^{\mathcal{D}}(3,2) = \frac{\sum_{i\in\mathcal{D}} \lambda_i \cdot (\lambda_i^2 c_{i,t}^2)}{\sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t}^2} \geq \frac{\sum_{i\in\mathcal{D}} \lambda_k \cdot (\lambda_i^2 c_{i,t}^2)}{\sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t}^2} = \lambda_k.$$

For the bulk subspace, $\forall j \in \mathcal{B}, \lambda_j \leq \lambda_{k+1}$, which provides an upper bound:

$$\mu_t^{\mathcal{B}}(3,2) = \frac{\sum_{j\in\mathcal{B}} \lambda_j \cdot (\lambda_j^3 c_{j,t}^2)}{\sum_{j\in\mathcal{B}} \lambda_j^2 c_{j,t}^2} \leq \frac{\sum_{j\in\mathcal{B}} \lambda_{k+1} \cdot (\lambda_j^2 c_{j,t}^2)}{\sum_{j\in\mathcal{B}} \lambda_j^2 c_{j,t}^2} = \lambda_{k+1}.$$

The spectral gap condition $\lambda_k > \lambda_{k+1}$ combines these bounds into the strict inequality chain:

$$\mu_t^{\mathcal{D}}(3,2) \geq \lambda_k > \lambda_{k+1} \geq \mu_t^{\mathcal{B}}(3,2).$$

This proves $\mu_t^{\mathcal{D}}(3,2) > \mu_t^{\mathcal{B}}(3,2)$, which implies that $\alpha_t > 0$.

Second, we prove that the constant coefficient γ_t is negative definite. By definition,

$$\gamma_t = -n_{\mathcal{D}}^{\text{loss}}.$$

For any $\boldsymbol{x}_t$ and noise $\boldsymbol{\xi}$, $n_{\mathcal{D}}^{\text{loss}} = \sum_{i\in\mathcal{D}} \lambda_i \kappa_i^2 > 0$. Since γ_t is the negative product of two positive definite quantities, it follows that $\gamma_t < 0$.

Finally, we verify the function's values at the boundaries. At $\theta = 0$, the value is the constant term, so $h(0) = \gamma_t$. As just shown, this is negative definite. At $\theta = 1$, the value is the sum of the coefficients. Substituting their definitions and simplifying yields:

$$\begin{aligned} h(1) &= \alpha_t + \beta_t + \gamma_t \\ &= \alpha_t + \left(-\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}\right) + \left(-n_{\mathcal{D}}^{\text{loss}}\right) \\ &= n_{\mathcal{B}}^{\text{loss}}. \end{aligned}$$

Since $n_{\mathcal{B}}^{\text{loss}} > 0$, we have $h(1) > 0$. This completes the proof of all claims. ■

Lemma 24 *Under the conditions of Lemma 23, there exists a unique critical function $\theta^{crit}(\boldsymbol{x}_t) : \mathbb{R}^d \to (0,1)$. For brevity, we will use θ_t^{crit} for $\theta^{crit}(\boldsymbol{x}_t)$ in the following. For $h(\theta) = \alpha_t\theta^2 + \beta_t\theta + \gamma_t$, the following hold:*

1. $\forall\, \boldsymbol{x}_t,\ \forall\, \theta \in \left(0, \theta_t^{crit}\right),\ h(\theta) < 0,$
2. $\forall\, \boldsymbol{x}_t,\ \forall\, \theta \in \left(\theta_t^{crit}, 1\right),\ h(\theta) > 0.$

Moreover, θ_t^{crit} is the unique root of $h(\theta) = 0$ in $(0,1)$, given explicitly by:

$$\theta_t^{crit} = \frac{-\beta_t + \sqrt{\beta_t^2 - 4\alpha_t\gamma_t}}{2\alpha_t}.$$

Proof The proof relies on the properties of the quadratic function $h(\theta) = \alpha_t\theta^2 + \beta_t\theta + \gamma_t$ established in Lemma 23 for any given $\boldsymbol{x}_t$.

First, we analyze the roots of the equation $h(\theta) = 0$. The discriminant is $\Delta_t = \beta_t^2 - 4\alpha_t\gamma_t$. From Lemma 23, we have $\alpha_t > 0$ and $\gamma_t < 0$. This implies that the term $-4\alpha_t\gamma_t$ is strictly positive. Therefore,

$$\Delta_t = \beta_t^2 - 4\alpha_t\gamma_t > 0.$$

Since $\Delta_t > 0$, the equation $h(\theta) = 0$ has two distinct real roots.

Next, we analyze the location of these roots using Vieta's formulas. The product of the roots is given by:

$$\theta_1\theta_2 = \frac{\gamma_t}{\alpha_t}.$$

Since $\gamma_t < 0$ and $\alpha_t > 0$, their ratio is strictly negative, $\frac{\gamma_t}{\alpha_t} < 0$. This proves that one root is positive and the other is negative.

Let the unique positive root be denoted by θ_t^{crit}. We now prove that this root must lie in the interval $(0, 1)$. Lemma 23 states that $h(0) < 0$ and $h(1) > 0$. Since $h(\theta)$ is a continuous function, the Intermediate Value Theorem guarantees that there must be at least one root in $(0, 1)$. As we have already established that there is only one positive root, this positive root must be the one that lies in $(0, 1)$. Thus, there exists a unique root θ_t^{crit} in the interval $(0, 1)$.

Now, we derive the explicit formula for this root. The two roots of the quadratic equation are given by $\frac{-\beta_t \pm \sqrt{\Delta_t}}{2\alpha_t}$. We must identify which sign corresponds to the positive root. We analyze the magnitude of the square root term:

$$\Delta_t = \beta_t^2 - 4\alpha_t\gamma_t > \beta_t^2 \implies \sqrt{\Delta_t} > \sqrt{\beta_t^2} = |\beta_t|.$$

Consider the root with the minus sign: $-\beta_t - \sqrt{\Delta_t}$. Since $\sqrt{\Delta_t} > |\beta_t| \geq -\beta_t$, this term is always negative. Thus, $\frac{-\beta_t - \sqrt{\Delta_t}}{2\alpha_t}$ is the negative root. Consider the root with the plus sign: $-\beta_t + \sqrt{\Delta_t}$. Since $\sqrt{\Delta_t} > |\beta_t| \geq \beta_t$, this term is always positive. Thus, $\frac{-\beta_t + \sqrt{\Delta_t}}{2\alpha_t}$ is the unique positive root. We conclude:

$$\theta_t^{\text{crit}} = \frac{-\beta_t + \sqrt{\beta_t^2 - 4\alpha_t\gamma_t}}{2\alpha_t}.$$

Finally, we prove the sign of $h(\theta)$ on either side of θ_t^{crit}. From Lemma 23, we know that $h(\theta)$ is an upward-opening parabola ($\alpha_t > 0$). A continuous, upward-opening parabola with a single root in an interval must be negative before that root and positive after it within that interval.

- For any $\theta \in (0, \theta_t^{\text{crit}})$, since $h(0) < 0$ and the only root in this range is at the endpoint, $h(\theta)$ must remain negative.
- For any $\theta \in (\theta_t^{\text{crit}}, 1)$, since $h(1) > 0$ and the only root in this range is at the startpoint, $h(\theta)$ must remain positive.

This completes the proof of all claims. ■

Theorem *[11](**Condition Differences on Different Alignment Regime**) Under Assumption 1, the relative ordering of the loss thresholds is determined by the alignment θ_t of a given $\boldsymbol{x}_t$ relative to the unique critical threshold, $\theta_t^{crit} \in (0, 1)$. This threshold is the root of the quadratic equation $h(\theta) = \alpha_t\theta^2 + \beta_t\theta + \gamma_t = 0$, with coefficients defined by:*

$$\alpha_t = s_t \left(\mu_t^{\mathcal{D}}(3, 2) - \mu_t^{\mathcal{B}}(3, 2)\right) \tag{33}$$

$$\beta_t = -\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}} \tag{34}$$

$$\gamma_t = -n_{\mathcal{D}}^{\text{loss}} \tag{35}$$

The ordering of the thresholds in the two resulting regimes is as follows:

$\boxed{\theta_t < \theta_t^{crit}}$	$\boxed{\theta_t > \theta_t^{crit}}$
$\underbrace{\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) < \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t)}$	$\underbrace{\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) > \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t)}$
Low-Alignment Regime	*High-Alignment Regime*

Proof We begin by analyzing the sign of the difference between the two loss thresholds.

$$\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) - \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t) = \frac{2s_t^{\mathcal{D}}}{\tau_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}}} - \frac{2s_t^{\mathcal{B}}}{u_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}} = \frac{2\left(s_t^{\mathcal{D}}(\tau_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}) - s_t^{\mathcal{B}}(\tau_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}})\right)}{(\tau_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}})(\tau_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}})}.$$

Since the denominator is a product of positive definite terms, the sign of this difference is strictly determined by the sign of the numerator.

$$\text{sign}\left(\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) - \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t)\right) = \text{sign}\left(s_t^{\mathcal{D}}(u_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}) - s_t^{\mathcal{D}}(u_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}})\right). \tag{36}$$

We now perform a detailed expansion of the numerator expression. Notice that $\tau_t^{\mathcal{S}} = \mu_t^{\mathcal{S}}(3,2)s_t^{\mathcal{S}}$ and $\theta_t = \frac{s_t^{\mathcal{D}}}{s_t} = 1 - \frac{s_t^{\mathcal{B}}}{s_t}$:

$$\begin{aligned}
&s_t^{\mathcal{D}}(\tau_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}) - s_t^{\mathcal{B}}(\tau_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}}) \\
&= s_t^{\mathcal{D}}\left(\mu_t^{\mathcal{B}}(3,2)s_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}\right) - s_t^{\mathcal{B}}\left(\mu_t^{\mathcal{D}}(3,2)s_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}}\right) \\
&= s_t^{\mathcal{D}}s_t^{\mathcal{B}}\left(\mu_t^{\mathcal{B}}(3,2) - \mu_t^{\mathcal{D}}(3,2)\right) + s_t^{\mathcal{D}}n_{\mathcal{B}}^{\text{loss}} - s_t^{\mathcal{B}}n_{\mathcal{D}}^{\text{loss}} \\
&= s_t^2\theta_t(1-\theta_t)\left(\mu_t^{\mathcal{B}}(3,2) - \mu_t^{\mathcal{D}}(3,2)\right) \\
&\quad + s_t\theta_t n_{\mathcal{B}}^{\text{loss}} - s_t(1-\theta_t)n_{\mathcal{D}}^{\text{loss}}.
\end{aligned}$$

For non-zero $\boldsymbol{x}_t$, we have $s_t > 0$. We can factor out s_t from the entire expression and collect terms by powers of θ_t:

$$\begin{aligned}
= s_t \cdot \Bigg(&\theta_t^2\Big[s_t\left(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\right)\Big] \\
+ \quad &\theta_t\Big[s_t\left(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\right) + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}\Big] \\
+ \quad &\Big[-n_{\mathcal{D}}^{\text{loss}}\Big]\Bigg).
\end{aligned}$$

The expression inside the parentheses is a quadratic function of θ_t. We define this normalized quadratic as our crossover function $h(\theta_t)$:

$$h(\theta_t) := \alpha_t\theta_t^2 + \beta_t\theta_t + \gamma_t,$$

Where the coefficients $\alpha_t, \beta_t, \gamma_t$ are defined as:

$$\alpha_t = s_t\left(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\right) \tag{37}$$

$$\beta_t = -\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}} \tag{38}$$

$$\gamma_t = -n_{\mathcal{D}}^{\text{loss}} \tag{39}$$

Which precisely match the coefficients of Lemma 23. We have thus established the central equivalence:

$$\operatorname{sign}\left(\eta_{\mathcal{D}}^{\text{loss}}(\boldsymbol{x}_t) - \eta_{\mathcal{B}}^{\text{loss}}(\boldsymbol{x}_t)\right) = \operatorname{sign}\left(h(\theta_t(\boldsymbol{x}_t))\right).$$

The remainder of the proof follows directly from Lemma 23 and Lemma 24. These lemmas establish that for any given $\boldsymbol{x}_t$, the function $h(\theta)$ has a unique root $\theta_t^{\text{crit}} \in (0,1)$, is strictly negative for all $\theta \in (0, \theta_t^{\text{crit}})$, and is strictly positive for all $\theta \in (\theta_t^{\text{crit}}, 1)$. This proves the two regimes stated in the theorem. ■

Theorem *12(**Asymptotic Limit of the Alignment Threshold**) Under Assumption 1, let the spectral gap be denoted by $m := \lambda_k/\lambda_{k+1}$. In the limit as the gap grows infinitely large, the critical threshold converges to 1:*

$$\lim_{m\to\infty} \theta_{\text{crit}}^{\eta}(\boldsymbol{x}_t) = 1.$$

Proof Recall that the critical threshold θ_t^{crit} is defined as the alignment where the loss thresholds are equal. Citing the condition from Eq. (36), any state $\boldsymbol{x}_t$ at this threshold must satisfy the exact equality:

$$s_t^{\mathcal{D}}(\tau_t^{\mathcal{B}} + n_{\mathcal{B}}^{\text{loss}}) - s_t^{\mathcal{B}}(\tau_t^{\mathcal{D}} + n_{\mathcal{D}}^{\text{loss}}) = 0.$$

This equality can be rearranged by separating $\tau_t^{\mathcal{S}}$ and $n_{\mathcal{S}}^{\text{loss}}$:

$$s_t^{\mathcal{B}}\tau_t^{\mathcal{D}} - s_t^{\mathcal{D}}\tau_t^{\mathcal{B}} = s_t^{\mathcal{D}}n_{\mathcal{B}}^{\text{loss}} - s_t^{\mathcal{B}}n_{\mathcal{D}}^{\text{loss}}.$$

Dividing by $s_t^{\mathcal{D}}s_t^{\mathcal{B}}$ gives the equivalent :

$$\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2) = \frac{n_{\mathcal{B}}^{\text{loss}}}{s_t^{\mathcal{B}}} - \frac{n_{\mathcal{D}}^{\text{loss}}}{s_t^{\mathcal{D}}}.$$

We now analyze this equality in the limit as the spectral gap $m := \lambda_k/\lambda_{k+1} \to \infty$. We can establish a lower bound for the left-hand side (LHS) and an upper bound for the right-hand side (RHS):

$$\text{LHS} = \mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2) \geq \lambda_k - \lambda_{k+1},$$

$$\text{RHS} = \frac{n_{\mathcal{B}}^{\text{loss}}}{s_t^{\mathcal{B}}} - \frac{n_{\mathcal{D}}^{\text{loss}}}{s_t^{\mathcal{D}}} < \frac{n_{\mathcal{B}}^{\text{loss}}}{s_t^{\mathcal{B}}}.$$

For the equality to hold, the lower bound of the LHS must be less than the upper bound of the RHS. This provides a necessary condition that any state at the crossover must satisfy:

$$\lambda_k - \lambda_{k+1} < \frac{n_{\mathcal{B}}^{\text{loss}}}{s_t^{\mathcal{B}}}.$$

This inequality provides a strict upper bound on the third-order bulk energy, $s_t^{\mathcal{B}}$:

$$0 \leq s_t^{\mathcal{B}} < \frac{n_{\mathcal{B}}^{\text{loss}}}{\lambda_k - \lambda_{k+1}}.$$

Remark *It is crucial to note that the term on the left-hand side of the inequality, $s_t^{\mathcal{B}}$, is intrinsically dependent on the vector $\boldsymbol{x}_t$, since its definition $s_t^{\mathcal{B}} = \sum_{j\in\mathcal{B}} \lambda_j^3 c_{j,t}^2$ explicitly involves the state's components $c_{j,t}$. Conversely, the upper bound on the right-hand side is independent of $\boldsymbol{x}_t$, as both the numerator $n_{\mathcal{B}}^{\text{loss}} = \sum_{j\in\mathcal{B}} \lambda_j \kappa_j^2$ and the denominator $\lambda_k - \lambda_{k+1}$ are determined solely by the constant Hessian eigenvalues and noise covariance structure. This inequality therefore provides a state-independent upper bound for a state-dependent quantity.*

We now show that this upper bound converges to zero as $m \to \infty$, independently of the state components $c_{j,t}$. The numerator is bounded by $n_{\mathcal{B}}^{\text{loss}} = \sum_{j\in\mathcal{B}} \lambda_j \kappa_j^2 \le \lambda_{k+1} \sum_{j\in\mathcal{B}} \kappa_j^2$. The denominator is $\lambda_{k+1}(m-1)$. Combining these gives:

$$\frac{n_{\mathcal{B}}^{\text{loss}}}{\lambda_k - \lambda_{k+1}} \le \frac{\lambda_{k+1} \sum_{j\in\mathcal{B}} \kappa_j^2}{\lambda_{k+1}(m-1)} = \frac{1}{m-1} \sum_{j\in\mathcal{B}} \kappa_j^2 \le \frac{1}{m-1} Tr(\boldsymbol{\Sigma}).$$

The term $Tr(\boldsymbol{\Sigma})$ is a finite constant dependent only on the noise covariance, not on the state $\boldsymbol{x}_t$. Thus, the upper bound for $s_t^{\mathcal{B}}$ converges to zero as $m \to \infty$. By the Squeeze Theorem, we rigorously conclude that $\lim_{m\to\infty} s_t^{\mathcal{B}} = 0$.

Finally, the limit of the critical alignment $\theta_t^{\text{crit}} = s_t^{\mathcal{D}}/(s_t^{\mathcal{D}} + s_t^{\mathcal{B}})$ can be determined. For a non-trivial state where the total energy does not vanish, we conclude:

$$\lim_{m\to\infty} \theta_t^{\text{crit}} = \frac{\lim_{m\to\infty} s_t^{\mathcal{D}}}{\lim_{m\to\infty} s_t^{\mathcal{D}} + \lim_{m\to\infty} s_t^{\mathcal{B}}} = \frac{s_t^{\mathcal{D}}}{s_t^{\mathcal{D}} + 0} = 1.$$

■

Theorem *13(**Asymptotic rate of** θ_t^{crit}) Under Assumption 1, let $m := \frac{\lambda_k}{\lambda_{k+1}} > 1$. Then the critical alignment threshold $\theta_t^{\text{crit}} \in (0,1)$ satisfies*

$$\frac{n_{\mathcal{B}}^{\text{loss}}}{s_t\,\lambda_{k+1}(m-1) + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}} \;\le\; 1 - \theta_t^{\text{crit}} \;\le\; \frac{n_{\mathcal{B}}^{\text{loss}}}{s_t\,\lambda_{k+1}(m-1)}.$$

If $\lambda_{k+1} = \Theta(1)$, consequently,

$$\lambda_k = \Theta(m), 1 - \theta_t^{\text{crit}} \in \Theta\left(\frac{1}{s_t\,(m-1)}\right).$$

Proof By Theorem 11, the critical threshold $\theta_t^{\text{crit}} \in (0,1)$ is the unique solution in $(0,1)$ of

$$\alpha_t \theta^2 + \beta_t \theta + \gamma_t = 0,$$

where

$$\alpha_t = s_t\big(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\big), \quad \beta_t = -\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}, \quad \gamma_t = -\,n_{\mathcal{D}}^{\text{loss}}.$$

Introduce the change of variables $\delta := 1 - \theta$. Substituting $\theta = 1 - \delta$ into the quadratic equation yields

$$\alpha_t(1-\delta)^2 + \big(-\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}\big)(1-\delta) - n_{\mathcal{D}}^{\text{loss}} = 0.$$

Expanding and collecting terms gives

$$\alpha_t - 2\alpha_t\delta + \alpha_t\delta^2 - \alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}} - \big(-\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}\big)\delta - n_{\mathcal{D}}^{\text{loss}} = 0,$$

which simplifies to

$$\alpha_t\delta^2 + \big(n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}\big)\delta - n_{\mathcal{B}}^{\text{loss}} = 0.$$

Rearranging yields the equivalent identity

$$\delta\big(\alpha_t\delta + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}\big) = n_{\mathcal{B}}^{\text{loss}}.$$

Since $\delta > 0$, this identity immediately implies the bounds

$$\frac{n_{\mathcal{B}}^{\text{loss}}}{\alpha_t + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}} \;\le\; \delta \;\le\; \frac{n_{\mathcal{B}}^{\text{loss}}}{\alpha_t}.$$

Next, by definition,

$$\mu_t^{\mathcal{D}}(3,2) = \frac{\sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2}{\sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t}^2} = \sum_{i\in\mathcal{D}} \left(\frac{\lambda_i^2 c_{i,t}^2}{\sum_{j\in\mathcal{D}} \lambda_j^2 c_{j,t}^2}\right) \lambda_i,$$

which is a convex combination of $\{\lambda_i : i \in \mathcal{D}\}$ and therefore satisfies $\mu_t^{\mathcal{D}}(3,2) \ge \lambda_k$. Similarly, $\mu_t^{\mathcal{B}}(3,2) \le \lambda_{k+1}$. Consequently,

$$\alpha_t = s_t\big(\mu_t^{\mathcal{D}}(3,2) - \mu_t^{\mathcal{B}}(3,2)\big) \ge s_t(\lambda_k - \lambda_{k+1}) = s_t\,\lambda_{k+1}(m-1).$$

Substituting this bound into the previous inequalities and recalling that $\delta = 1 - \theta_t^{\text{crit}}$ yields

$$\frac{n_{\mathcal{B}}^{\text{loss}}}{s_t\,\lambda_{k+1}(m-1) + n_{\mathcal{B}}^{\text{loss}} + n_{\mathcal{D}}^{\text{loss}}} \;\le\; 1 - \theta_t^{\text{crit}} \;\le\; \frac{n_{\mathcal{B}}^{\text{loss}}}{s_t\,\lambda_{k+1}(m-1)}.$$

If $\lambda_{k+1} = \Theta(1)$, then $\lambda_k = \Theta(m)$ and $\lambda_k - \lambda_{k+1} = \Theta(m-1)$, and the above bounds imply

$$1 - \theta_t^{\text{crit}} \in \Theta\left(\frac{1}{s_t\,(m-1)}\right),$$

which completes the proof. ■

A.8. Lemma Of CSGD

Let $\mathbb{E}_t[\cdot \mid \boldsymbol{x}_j], \mathrm{Var}_t[\cdot \mid \boldsymbol{x}_j]$ denote expectation,variance over the noise sequence $\{\boldsymbol{\xi}_s\}_{s=j+1}^{t-1}$.

Lemma 25 (Asymptotic representation of conditional alignment) *Under Assumption 1, define the normalized block statistics at time t for $\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}$ as*

$$s_t'^{,\mathcal{S}} := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \lambda_i^2 c_{i,t}^2, \quad u_t'^{,\mathcal{S}} := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \lambda_i^4 c_{i,t}^2, \quad w_t'^{,\mathcal{S}} := \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \lambda_i^3 c_{i,t}^2,$$

and let $\bar{e}_{\mathcal{D}} := \frac{1}{k} e_{\mathcal{D}}$, $\bar{e}_{\mathcal{B}} := \frac{1}{d-k} e_{\mathcal{B}}$. *Define the deterministic function* $k : (0,\infty)^6 \to [0,1]$ *by*

$$k(s_1, s_2, u_1, u_2, w_1, w_2) = \frac{1}{1 + \frac{1}{\rho} \cdot \frac{s_2 - 2\eta_t w_2 + \eta_t^2 (u_2 + \bar{e}_{\mathcal{B}})}{s_1 - 2\eta_t w_1 + \eta_t^2 (u_1 + \bar{e}_{\mathcal{D}})}}, \quad \textit{where } \rho = \lim_{d\to\infty} \frac{k}{d-k}.$$

Then, under the spectral and noise assumptions, we have:

$$\begin{aligned} &k(s_1, s_2, u_1, u_2, w_1, w_2) \\ &= \lim_{d\to\infty} \mathbb{E}_{t+1} \left[\theta_{t+1} \,\middle|\, s_t^{\prime,\mathcal{D}} = s_1,\ s_t^{\prime,\mathcal{B}} = s_2,\ u_t^{\prime,\mathcal{D}} = u_1,\ u_t^{\prime,\mathcal{B}} = u_2,\ w_t^{\prime,\mathcal{D}} = w_1,\ w_t^{\prime,\mathcal{B}} = w_2 \right] \end{aligned} \tag{40}$$

Notice that for any deterministic $\boldsymbol{x}$ *satisfy :*

$$s^{\prime,\mathcal{D}} = s_1,\ s^{\prime,\mathcal{B}} = s_2,\ u^{\prime,\mathcal{D}} = u_1,\ u^{\prime,\mathcal{B}} = u_2,\ w^{\prime,\mathcal{D}} = w_1,\ w^{\prime,\mathcal{B}} = w_2$$

We have:

$$k(s_1, s_2, u_1, u_2, w_1, w_2) = \lim_{d\to\infty} \mathbb{E}_{t+1} \left[\theta_{t+1} \,|\, \boldsymbol{x}_t = \boldsymbol{x} \right]$$

Proof Recall that $\theta_{t+1} = \dfrac{s_{t+1}^{\mathcal{D}}}{s_{t+1}^{\mathcal{D}} + s_{t+1}^{\mathcal{B}}}$, where the unnormalized energies are

$$s_{t+1}^{\mathcal{D}} = \sum_{i\in\mathcal{D}} \lambda_i^2 c_{i,t+1}^2, \quad s_{t+1}^{\mathcal{B}} = \sum_{i\in\mathcal{B}} \lambda_i^2 c_{i,t+1}^2.$$

From Lemma 21, their conditional expectations (with respect to the noise $\boldsymbol{\xi}_t$ at step t) are

$$\begin{aligned} \mathbb{E}_{t+1}[s_{t+1}^{\mathcal{D}} \mid \boldsymbol{x}_t] &= s_t^{\mathcal{D}} - 2\eta_t \sum_{i\in\mathcal{D}} \lambda_i^3 c_{i,t}^2 + \eta_t^2 \left(\sum_{i\in\mathcal{D}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{D}} \right), \\ \mathbb{E}_{t+1}[s_{t+1}^{\mathcal{B}} \mid \boldsymbol{x}_t] &= s_t^{\mathcal{B}} - 2\eta_t \sum_{i\in\mathcal{B}} \lambda_i^3 c_{i,t}^2 + \eta_t^2 \left(\sum_{i\in\mathcal{B}} \lambda_i^4 c_{i,t}^2 + e_{\mathcal{B}} \right). \end{aligned}$$

Dividing by block sizes yields

$$\begin{aligned} \frac{1}{k} \mathbb{E}_{t+1}[s_{t+1}^{\mathcal{D}} \mid \boldsymbol{x}_t] &= s_t^{\prime,\mathcal{D}} - 2\eta_t w_t^{\prime,\mathcal{D}} + \eta_t^2 (u_t^{\prime,\mathcal{D}} + \bar{e}_{\mathcal{D}}), \\ \frac{1}{d-k} \mathbb{E}_{t+1}[s_{t+1}^{\mathcal{B}} \mid \boldsymbol{x}_t] &= s_t^{\prime,\mathcal{B}} - 2\eta_t w_t^{\prime,\mathcal{B}} + \eta_t^2 (u_t^{\prime,\mathcal{B}} + \bar{e}_{\mathcal{B}}). \end{aligned}$$

Define the deterministic scalars

$$\vartheta_1 := s_t^{\prime,\mathcal{D}} - 2\eta_t w_t^{\prime,\mathcal{D}} + \eta_t^2 (u_t^{\prime,\mathcal{D}} + \bar{e}_{\mathcal{D}}), \quad \vartheta_2 := s_t^{\prime,\mathcal{B}} - 2\eta_t w_t^{\prime,\mathcal{B}} + \eta_t^2 (u_t^{\prime,\mathcal{B}} + \bar{e}_{\mathcal{B}}).$$

We now analyze the concentration of the normalized energies. Expanding the SGD update

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t (\boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\xi}_t)$$

in the eigenbasis $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d]$, we have for each i:

$$c_{i,t+1} = (1 - \eta_t \lambda_i) c_{i,t} - \eta_t \zeta_{i,t},$$

where $\boldsymbol{\zeta}_t := \boldsymbol{U}^\top \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$ with $\boldsymbol{C} = \boldsymbol{U}^\top \boldsymbol{\Sigma} \boldsymbol{U}$ and $\kappa_i^2 = (\boldsymbol{C})_{ii}$. Hence,

$$s_{t+1}^{\mathcal{S}} = \sum_{i \in \mathcal{S}} \lambda_i^2 c_{i,t+1}^2 = (\text{deterministic}) + (\text{linear in } \boldsymbol{\zeta}_t) + (\text{quadratic in } \boldsymbol{\zeta}_t).$$

By Lemma 17 and Assumption 1, the conditional variance satisfies

$$\mathrm{Var}_{t+1}\left(\frac{1}{|\mathcal{S}|} s_{t+1}^{\mathcal{S}} \,\middle|\, \boldsymbol{x}_t \right) = O\left(\frac{1}{|\mathcal{S}|} \right).$$

Thus, by Chebyshev's inequality,

$$\lim_{d \to \infty} \mathbb{P}\left(\left| \frac{1}{|\mathcal{S}|} s_{t+1}^{\mathcal{S}} - \vartheta_{\mathcal{S}} \right| > \epsilon \,\middle|\, \boldsymbol{x}_t \right) = 0,$$

where $\vartheta_{\mathcal{D}} = \vartheta_1$ and $\vartheta_{\mathcal{B}} = \vartheta_2$. Consequently,

$$\lim_{d \to \infty} \frac{s_{t+1}^{\mathcal{B}}}{s_{t+1}^{\mathcal{D}}} = \frac{1}{\rho} \cdot \frac{\vartheta_2}{\vartheta_1} \quad \text{in probability (conditional on } \boldsymbol{x}_t).$$

Since $f(x) = 1/(1+x)$ is continuous and bounded on $(0, \infty)$, Lemma 19 implies

$$\lim_{d \to \infty} \mathbb{E}_{t+1}\big[\theta_{t+1} \mid \boldsymbol{x}_t\big] = f\left(\frac{1}{\rho} \cdot \frac{\vartheta_2}{\vartheta_1} \right) = k\big(s_t^{\prime,\mathcal{D}}, s_t^{\prime,\mathcal{B}}, u_t^{\prime,\mathcal{D}}, u_t^{\prime,\mathcal{B}}, w_t^{\prime,\mathcal{D}}, w_t^{\prime,\mathcal{B}}\big),$$

which completes the proof. ■

Lemma 26 (Concentration of macroscopic statistics at time t under CSGD) *Under Assumptions 1 and 14, with deterministic initialization $\boldsymbol{x}_0$ (so that $c_{i,0}$ are fixed scalars), define for $\mathcal{S} \in \{\mathcal{D}, \mathcal{B}\}$ the normalized statistics at time t:*

$$s_t^{\prime,\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \lambda_i^2 c_{i,t}^2, \quad u_t^{\prime,\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \lambda_i^4 c_{i,t}^2, \quad w_t^{\prime,\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \lambda_i^3 c_{i,t}^2.$$

Then, as $d \to \infty$, each statistic converges in probability to its expectation, which is a deterministic function of the initial eigencoordinates $\{c_{i,0}\}$:

$$s_t^{\prime,\mathcal{S}} \xrightarrow{p} \bar{s}_t^{\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \lambda_i^2 \left[(1 - \eta\lambda_i)^{2t} (c_{i,0}^2 - \beta_i) + \beta_i \right],$$

$$u_t^{\prime,\mathcal{S}} \xrightarrow{p} \bar{u}_t^{\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \lambda_i^4 \left[(1 - \eta\lambda_i)^{2t} (c_{i,0}^2 - \beta_i) + \beta_i \right],$$

$$w_t^{\prime,\mathcal{S}} \xrightarrow{p} \bar{w}_t^{\mathcal{S}} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \lambda_i^3 \left[(1 - \eta\lambda_i)^{2t} (c_{i,0}^2 - \beta_i) + \beta_i \right],$$

where $\beta_i = \dfrac{\eta\kappa_i^2}{2\lambda_i - \eta\lambda_i^2} > 0.$

Proof We prove the result for $s_t'^{,\mathcal{D}}$; the other five statistics follow identically by replacing the weight λ_i^2 with λ_i^4 or λ_i^3, and/or changing the block to $\mathcal{B}$. Under constant step size SGD, the eigencoordinate evolves as

$$c_{i,t} = (1-\eta\lambda_i)^t c_{i,0} - \eta \sum_{s=0}^{t-1} (1-\eta\lambda_i)^{t-1-s} \zeta_{i,s},$$

where $\zeta_{i,s} = \boldsymbol{u}_i^\top \boldsymbol{\xi}_s \sim \mathcal{N}(0, \kappa_i^2)$ are independent across s. Since $c_{i,0}$ is deterministic, $c_{i,t}$ is Gaussian with mean $\mu_{i,t} = (1-\eta\lambda_i)^t c_{i,0}$ and variance

$$\sigma_{i,t}^2 = \eta^2\kappa_i^2 \sum_{s=0}^{t-1} (1-\eta\lambda_i)^{2(t-1-s)} = \eta^2\kappa_i^2 \frac{1-(1-\eta\lambda_i)^{2t}}{1-(1-\eta\lambda_i)^2} = \beta_i \left(1-(1-\eta\lambda_i)^{2t}\right),$$

where we used the definition $\beta_i = \eta\kappa_i^2/(2\lambda_i - \eta\lambda_i^2)$ and the identity $1-(1-\eta\lambda_i)^2 = \eta\lambda_i(2-\eta\lambda_i)$. For a Gaussian random variable, $\mathbb{E}_t[c_{i,t}^2] = \mu_{i,t}^2 + \sigma_{i,t}^2$, so

$$\mathbb{E}_t[c_{i,t}^2] = (1-\eta\lambda_i)^{2t} c_{i,0}^2 + \beta_i \left(1-(1-\eta\lambda_i)^{2t}\right) = (1-\eta\lambda_i)^{2t}(c_{i,0}^2 - \beta_i) + \beta_i.$$

Thus, the expectation of $s_t'^{,\mathcal{D}}$ is

$$\mathbb{E}_t[s_t'^{,\mathcal{D}}] = \frac{1}{k}\sum_{i\in\mathcal{D}} \lambda_i^2 \mathbb{E}_t[c_{i,t}^2] = \bar{s}_t^{\mathcal{D}},$$

which is deterministic and depends only on $\{c_{i,0}\}$. Since the noise is independent across eigendirections, the random variables $\{c_{i,t}^2\}_{i=1}^d$ are independent. Hence,

$$\mathrm{Var}_t(s_t'^{,\mathcal{D}}) = \frac{1}{k^2}\sum_{i\in\mathcal{D}} \lambda_i^4 \mathrm{Var}_t(c_{i,t}^2).$$

For a Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mathrm{Var}(X^2) = 2\sigma^4 + 4\mu^2\sigma^2 \le 2(\mu^2+\sigma^2)^2 = 2(\mathbb{E}[X^2])^2$. Therefore,

$$\mathrm{Var}_t(c_{i,t}^2) \le 2\left(\mathbb{E}_t[c_{i,t}^2]\right)^2 \le 2M^2,$$

for some constant $M < \infty$, because $c_{i,0}^2$ is bounded (by trajectory boundedness in Assumption 1), $\beta_i \le \frac{\eta s_{\max}}{2\lambda_d - \eta\lambda_d^2} < \infty$ (since $\lambda_d > 0$ and $\eta < 2/\lambda_1 \le 2/\lambda_d$), and $(1-\eta\lambda_i)^{2t} \in [0,1]$. Thus,

$$\mathrm{Var}_t(s_t'^{,\mathcal{D}}) \le \frac{1}{k^2}\sum_{i\in\mathcal{D}} \lambda_i^4 \cdot 2M^2 = \frac{2M^2}{k} \cdot \left(\frac{1}{k}\sum_{i\in\mathcal{D}} \lambda_i^4\right).$$

By Assumption 1, $\frac{1}{k}\sum_{i\in\mathcal{D}} \lambda_i^4 \to \lambda_{\mathcal{D},4} < \infty$, and $k \to \infty$ as $d \to \infty$. Hence,

$$\mathrm{Var}_t(s_t'^{,\mathcal{D}}) = O\left(\frac{1}{k}\right) \xrightarrow{d\to\infty} 0.$$

By Chebyshev's inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|s_t'^{,\mathcal{D}} - \mathbb{E}_t[s_t'^{,\mathcal{D}}]\right| > \epsilon\right) \le \frac{\mathrm{Var}_t(s_t'^{,\mathcal{D}})}{\epsilon^2} \xrightarrow{d\to\infty} 0.$$

Therefore, $s_t'^{,\mathcal{D}} \xrightarrow{p} \bar{s}_t^{\mathcal{D}}$. The same argument applies to $s_t'^{,\mathcal{B}}$ (with $|\mathcal{B}| = d-k \to \infty$), and to $u_t'^{,\mathcal{S}}, w_t'^{,\mathcal{S}}$ by replacing λ_i^2 with λ_i^4 or λ_i^3 (the boundedness of spectral moments for $p = 3, 4$ is assumed in Assumption 1). This completes the proof. ■

Lemma 27 (Interchange of expectation and $k(\cdot)$ at time t) *Let $k : (0,\infty)^6 \to [0,1]$ be the deterministic function defined in Lemma 25. Under the conditions of Lemma 26, we have*

$$\lim_{d\to\infty} \mathbb{E}_t\left[k\left(s_t'^{,\mathcal{D}}, s_t'^{,\mathcal{B}}, u_t'^{,\mathcal{D}}, u_t'^{,\mathcal{B}}, w_t'^{,\mathcal{D}}, w_t'^{,\mathcal{B}}\right)\right] = k\left(\bar{s}_t^{\mathcal{D}}, \bar{s}_t^{\mathcal{B}}, \bar{u}_t^{\mathcal{D}}, \bar{u}_t^{\mathcal{B}}, \bar{w}_t^{\mathcal{D}}, \bar{w}_t^{\mathcal{B}}\right),$$

where $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \boldsymbol{x}_0]$ denotes expectation over the noise up to time $t-1$, and the limits $\bar{s}_t^{\mathcal{S}}, \bar{u}_t^{\mathcal{S}}, \bar{w}_t^{\mathcal{S}}$ are as defined in Lemma 26.

Proof Define the random vector

$$\boldsymbol{Z}_d := \left(s_t'^{,\mathcal{D}}, s_t'^{,\mathcal{B}}, u_t'^{,\mathcal{D}}, u_t'^{,\mathcal{B}}, w_t'^{,\mathcal{D}}, w_t'^{,\mathcal{B}}\right).$$

By Lemma 26, $\boldsymbol{Z}_d \xrightarrow{p} \bar{\boldsymbol{z}}_t := (\bar{s}_t^{\mathcal{D}}, \bar{s}_t^{\mathcal{B}}, \bar{u}_t^{\mathcal{D}}, \bar{u}_t^{\mathcal{B}}, \bar{w}_t^{\mathcal{D}}, \bar{w}_t^{\mathcal{B}})$ as $d \to \infty$. The function $k(\cdot)$ is continuous at $\bar{\boldsymbol{z}}_t$ because its explicit form

$$k(s_1, s_2, u_1, u_2, w_1, w_2) = \frac{1}{1 + \frac{1}{\rho} \cdot \frac{s_2 - 2\eta w_2 + \eta^2(u_2 + \bar{e}_{\mathcal{B}})}{s_1 - 2\eta w_1 + \eta^2(u_1 + \bar{e}_{\mathcal{D}})}}$$

involves only continuous operations, and the denominator is strictly positive under Assumption 14 (since $\eta < 2/\lambda_1$ ensures $(1-\eta\lambda_i)^2 > 0$ for all i, and the noise terms $\bar{e}_{\mathcal{D}}, \bar{e}_{\mathcal{B}} \geq 0$ add positivity). Moreover, $k(\cdot)$ is bounded between 0 and 1 since it represents the asymptotic limit of the alignment $\theta_{t+1} \in [0,1]$. Therefore, by Lemma 19 (the Continuous Mapping Theorem combined with the Bounded Convergence Theorem), the convergence in probability of $\boldsymbol{Z}_d$ to $\bar{\boldsymbol{z}}_t$ implies that

$$\lim_{d\to\infty} \mathbb{E}_t\left[k(\boldsymbol{Z}_d)\right] = k(\bar{\boldsymbol{z}}_t),$$

which is precisely the claimed equality. ■

Lemma 28 (Equivalence of conditional and unconditional alignment expectation) *Under Assumptions 1 and 14, with deterministic initialization $\boldsymbol{x}_0$, define the deterministic limits*

$$\bar{s}_t^{\mathcal{D}} := \lim_{d\to\infty} \mathbb{E}_t[s_t'^{,\mathcal{D}}], \qquad \bar{s}_t^{\mathcal{B}} := \lim_{d\to\infty} \mathbb{E}_t[s_t'^{,\mathcal{B}}],$$
$$\bar{u}_t^{\mathcal{D}} := \lim_{d\to\infty} \mathbb{E}_t[u_t'^{,\mathcal{D}}], \qquad \bar{u}_t^{\mathcal{B}} := \lim_{d\to\infty} \mathbb{E}_t[u_t'^{,\mathcal{B}}],$$
$$\bar{w}_t^{\mathcal{D}} := \lim_{d\to\infty} \mathbb{E}_t[w_t'^{,\mathcal{D}}], \qquad \bar{w}_t^{\mathcal{B}} := \lim_{d\to\infty} \mathbb{E}_t[w_t'^{,\mathcal{B}}].$$

Then the following equality holds:

$$\lim_{d\to\infty} \mathbb{E}_{t+1}\left[\theta_{t+1} \,\middle|\, \begin{array}{l} s_t'^{,\mathcal{D}} = \bar{s}_t^{\mathcal{D}},\ s_t'^{,\mathcal{B}} = \bar{s}_t^{\mathcal{B}}, \\ u_t'^{,\mathcal{D}} = \bar{u}_t^{\mathcal{D}},\ u_t'^{,\mathcal{B}} = \bar{u}_t^{\mathcal{B}}, \\ w_t'^{,\mathcal{D}} = \bar{w}_t^{\mathcal{D}},\ w_t'^{,\mathcal{B}} = \bar{w}_t^{\mathcal{B}} \end{array}\right] = \lim_{d\to\infty} \mathbb{E}_{t+1}[\theta_{t+1}].$$

Proof By Lemma 25, for any fixed values of the six statistics at time t, the conditional expectation satisfies

$$\lim_{d\to\infty} \mathbb{E}_{t+1}[\theta_{t+1} \mid s_t'^{,\mathcal{D}}, s_t'^{,\mathcal{B}}, u_t'^{,\mathcal{D}}, u_t'^{,\mathcal{B}}, w_t'^{,\mathcal{D}}, w_t'^{,\mathcal{B}}] = k(s_t'^{,\mathcal{D}}, s_t'^{,\mathcal{B}}, u_t'^{,\mathcal{D}}, u_t'^{,\mathcal{B}}, w_t'^{,\mathcal{D}}, w_t'^{,\mathcal{B}}).$$

Substituting the deterministic limits $\bar{s}_t^{\mathcal{D}}, \ldots, \bar{w}_t^{\mathcal{B}}$ into this identity yields

$$\lim_{d\to\infty} \mathbb{E}_{t+1}[\theta_{t+1} \mid s_t'^{,\mathcal{D}} = \bar{s}_t^{\mathcal{D}}, \ldots, w_t'^{,\mathcal{B}} = \bar{w}_t^{\mathcal{B}}] = k(\bar{s}_t^{\mathcal{D}}, \bar{s}_t^{\mathcal{B}}, \bar{u}_t^{\mathcal{D}}, \bar{u}_t^{\mathcal{B}}, \bar{w}_t^{\mathcal{D}}, \bar{w}_t^{\mathcal{B}}).$$

On the other hand, by the law of total expectation,

$$\mathbb{E}_{t+1}[\theta_{t+1}] = \mathbb{E}_t\left[\mathbb{E}_{t+1}[\theta_{t+1} \mid s_t'^{,\mathcal{D}}, \ldots, w_t'^{,\mathcal{B}}]\right].$$

Taking the limit as $d \to \infty$ on both sides gives

$$\lim_{d\to\infty} \mathbb{E}_{t+1}[\theta_{t+1}] = \lim_{d\to\infty} \mathbb{E}_t\left[\mathbb{E}_{t+1}[\theta_{t+1} \mid s_t'^{,\mathcal{D}}, \ldots, w_t'^{,\mathcal{B}}]\right].$$

Applying Lemma 25 inside the expectation, the right-hand side becomes

$$\lim_{d\to\infty} \mathbb{E}_t\left[k(s_t'^{,\mathcal{D}}, s_t'^{,\mathcal{B}}, u_t'^{,\mathcal{D}}, u_t'^{,\mathcal{B}}, w_t'^{,\mathcal{D}}, w_t'^{,\mathcal{B}})\right].$$

Finally, by Lemma 27, this limit equals

$$k(\bar{s}_t^{\mathcal{D}}, \bar{s}_t^{\mathcal{B}}, \bar{u}_t^{\mathcal{D}}, \bar{u}_t^{\mathcal{B}}, \bar{w}_t^{\mathcal{D}}, \bar{w}_t^{\mathcal{B}}).$$

Thus, both sides of the claimed equality converge to the same deterministic value, and the result follows. ■

A.9. Proofs of theorems in CSGD

Recall our assumption and notion:

Assumption *14 Our analysis is based on the following assumptions:*

Table 4: Assumptions for CSGD

Assumption	**Description**
Constant Step Size	$\eta_t = \eta, \quad 0 < \eta < \min\left\{\frac{2}{\lambda_1}, \frac{2(\lambda_k - \lambda_{k+1})}{\lambda_1^2 - \lambda_d^2}\right\}$
Initialization for $\boldsymbol{x}_0$	$\forall i \in \mathcal{D}, \quad c_{i,0}^2 > \beta_i,$ $\varrho_{\mathcal{D}} > \delta - \sum_{i\in\mathcal{D}} \beta_i$

For the analysis, we define several key quantities:

$$\beta_i := \frac{\eta\,\kappa_i^2}{2\lambda_i - \eta\lambda_i^2} > 0,\ \varrho_{\mathcal{D}} := \sum_{i\in\mathcal{D}}(c_{i,0}^2 - \beta_i).$$

$$\delta := \frac{s_{\max}\psi_{\mathcal{D}}\lambda_1^2}{\lambda_d^2\lambda_k^2\left(\frac{2(\lambda_k-\lambda_{k+1})}{\eta} - (\lambda_1^2-\lambda_d^2)\right)}.$$

Theorem *15(**Initial Decrease Phase**) Under Assumption 1 and Assumption 14, let the t^* be defined as*

$$t^* := \left\lfloor \frac{\log\left(\frac{\varrho_{\mathcal{D}}}{\delta - \sum_{i\in\mathcal{D}}\beta_i}\right)}{-2\log(1-\eta\lambda_1)} \right\rfloor.$$

Then for all time steps $t \in \{0, 1, \ldots, t^*-1\}$, the expected alignment is strictly decreasing:

$$\lim_{d\to\infty}\mathbb{E}[\theta_{t+1}] < \lim_{d\to\infty}\mathbb{E}[\theta_t].$$

Proof By Theorem 2, the desired inequality follows if we can show that

$$\lim_{d\to\infty}\mathbb{E}_t[\eta_t^* \mid s_t'^{,\mathcal{D}} = \bar{s}_t^{\mathcal{D}}, \ldots, w_t'^{,\mathcal{B}} = \bar{w}_t^{\mathcal{B}}] > \eta,$$

where $\bar{s}_t^{\mathcal{S}} = \lim_{d\to\infty}\mathbb{E}_t[s_t'^{,\mathcal{S}}]$, etc., are the deterministic limits from Lemma 26.

To lower-bound this expectation, we use the state-dependent bound from Theorem 7:

$$\eta_t^* \geq \frac{2\,\mathrm{gap}_1}{(\lambda_1^2-\lambda_d^2) + \frac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2\|\boldsymbol{x}_t\|^2\theta_t}}.$$

Taking conditional expectation given the six macroscopic statistics at time t, we obtain

$$\mathbb{E}_t[\eta_t^* \mid \bar{\boldsymbol{z}}_t] \geq \mathbb{E}_t\left[\frac{2\,\mathrm{gap}_1}{(\lambda_1^2-\lambda_d^2) + \frac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2\|\boldsymbol{x}_t\|^2\theta_t}} \,\middle|\, \bar{\boldsymbol{z}}_t\right].$$

Now, observe that $\|\boldsymbol{x}_t\|^2\theta_t$ is a continuous function of the six statistics. Specifically,

$$\|\boldsymbol{x}_t\|^2\theta_t = \left(\sum_{i=1}^d c_{i,t}^2\right)\cdot\frac{\sum_{j\in\mathcal{D}}\lambda_j^2 c_{j,t}^2}{\sum_{i=1}^d\lambda_i^2 c_{i,t}^2}.$$

By Lemma 26, each block average $\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}c_{i,t}^2$, $\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\lambda_i^2 c_{i,t}^2$, etc., concentrates around its expectation as $d\to\infty$. Therefore, $\|\boldsymbol{x}_t\|^2\theta_t$ concentrates around its deterministic limit

$$\overline{\|\boldsymbol{x}_t\|^2\theta_t} := \lim_{d\to\infty}\mathbb{E}_t[\|\boldsymbol{x}_t\|^2\theta_t].$$

Since the function $x \mapsto 1/(a + b/x)$ is continuous and bounded for $x > 0$, Lemma 19 implies

$$\lim_{d\to\infty} \mathbb{E}_t \left[\frac{2\,\mathrm{gap}_1}{(\lambda_1^2 - \lambda_d^2) + \dfrac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2 \|\boldsymbol{x}_t\|^2 \theta_t}} \,\middle|\, \bar{\boldsymbol{z}}_t \right] = \frac{2\,\mathrm{gap}_1}{(\lambda_1^2 - \lambda_d^2) + \dfrac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2 \overline{\|\boldsymbol{x}_t\|^2 \theta_t}}}.$$

We now lower-bound $\overline{\|\boldsymbol{x}_t\|^2 \theta_t}$. Using $\lambda_i \geq \lambda_k$ for $i \in \mathcal{D}$ and $\lambda_i \leq \lambda_1$ for all i, we have

$$\overline{\|\boldsymbol{x}_t\|^2 \theta_t} \geq \frac{\lambda_k^2}{\lambda_1^2} \lim_{d\to\infty} \mathbb{E}_t \left[\sum_{i\in\mathcal{D}} c_{i,t}^2 \right] = \frac{\lambda_k^2}{\lambda_1^2} \left((1 - \eta\lambda_1)^{2t} \varrho_{\mathcal{D}} + \sum_{i\in\mathcal{D}} \beta_i \right).$$

By the definition of t^*, for all $t < t^*$,

$$(1 - \eta\lambda_1)^{2t} \varrho_{\mathcal{D}} + \sum_{i\in\mathcal{D}} \beta_i > \delta.$$

Therefore,

$$\overline{\|\boldsymbol{x}_t\|^2 \theta_t} > \frac{\lambda_k^2}{\lambda_1^2} \delta.$$

Substituting the definition of δ yields

$$\frac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2 \overline{\|\boldsymbol{x}_t\|^2 \theta_t}} < \frac{s_{\max}\psi_{\mathcal{D}}\lambda_1^2}{\lambda_d^2 \lambda_k^2 \delta} = \frac{2(\lambda_k - \lambda_{k+1})}{\eta} - (\lambda_1^2 - \lambda_d^2).$$

Hence,

$$\frac{2\,\mathrm{gap}_1}{(\lambda_1^2 - \lambda_d^2) + \dfrac{s_{\max}\psi_{\mathcal{D}}}{\lambda_d^2 \overline{\|\boldsymbol{x}_t\|^2 \theta_t}}} > \frac{2(\lambda_k - \lambda_{k+1})}{(\lambda_1^2 - \lambda_d^2) + \left(\dfrac{2(\lambda_k - \lambda_{k+1})}{\eta} - (\lambda_1^2 - \lambda_d^2) \right)} = \eta.$$

Combining the above, we conclude that

$$\lim_{d\to\infty} \mathbb{E}_t[\eta_t^* \mid \bar{\boldsymbol{z}}_t] > \eta,$$

which implies

$$\lim_{d\to\infty} \mathbb{E}_{t+1}[\theta_{t+1} \mid \bar{\boldsymbol{z}}_t] < \lim_{d\to\infty} \theta_t.$$

Finally, by Lemma 28 and the concentration of θ_t, we obtain

$$\lim_{d\to\infty} \mathbb{E}_{t+1}[\theta_{t+1}] < \lim_{d\to\infty} \mathbb{E}_t[\theta_t]$$

for all $t \in \{0, 1, \ldots, t^* - 1\}$, completing the proof. ■

Theorem *16(**Late Phase**) Under Assumption 1 and Assumption 14, the late-time asymptotic expected alignment is given by*

$$\theta_\infty := \lim_{t\to\infty}\lim_{d\to\infty}\mathbb{E}_t[\theta_t] = \frac{\lim_{d\to\infty}\sum_{i\in\mathcal{D}}\lambda_i^2\beta_i}{\lim_{d\to\infty}\sum_{i=1}^{d}\lambda_i^2\beta_i},$$

where $\beta_i = \dfrac{\eta\kappa_i^2}{2\lambda_i - \eta\lambda_i^2} > 0$. *Equivalently,*

$$\theta_\infty = \frac{\lim_{d\to\infty}\sum_{i\in\mathcal{D}}\dfrac{\eta\lambda_i^2\kappa_i^2}{2\lambda_i-\eta\lambda_i^2}}{\lim_{d\to\infty}\sum_{i=1}^{d}\dfrac{\eta\lambda_i^2\kappa_i^2}{2\lambda_i-\eta\lambda_i^2}}.$$

Proof From the CSGD dynamics, the second moment of each eigencoordinate satisfies

$$\mathbb{E}_t[c_{i,t}^2] = (1-\eta\lambda_i)^{2t}(c_{i,0}^2 - \beta_i) + \beta_i, \quad \text{where } \beta_i = \frac{\eta\kappa_i^2}{2\lambda_i - \eta\lambda_i^2}.$$

Since $0 < \eta < 2/\lambda_1$, we have $|1-\eta\lambda_i| < 1$ for all i, so $(1-\eta\lambda_i)^{2t} \to 0$ as $t\to\infty$. Therefore,

$$\lim_{t\to\infty}\mathbb{E}_t[c_{i,t}^2] = \beta_i.$$

The expected unnormalized energies are

$$\mathbb{E}_t[s_t^{\mathcal{D}}] = \sum_{i\in\mathcal{D}}\lambda_i^2\mathbb{E}_t[c_{i,t}^2], \quad \mathbb{E}_t[s_t] = \sum_{i=1}^{d}\lambda_i^2\mathbb{E}_t[c_{i,t}^2].$$

Taking the limit as $t\to\infty$, we obtain

$$\lim_{t\to\infty}\mathbb{E}_t[s_t^{\mathcal{D}}] = \sum_{i\in\mathcal{D}}\lambda_i^2\beta_i, \quad \lim_{t\to\infty}\mathbb{E}_t[s_t] = \sum_{i=1}^{d}\lambda_i^2\beta_i.$$

By Lemma 26, for each fixed t, the normalized statistics $s_t'^{,\mathcal{D}} = s_t^{\mathcal{D}}/k$ and $s_t' = s_t/d$ concentrate around their expectations as $d\to\infty$:

$$\lim_{d\to\infty}\mathbb{E}_t[\theta_t] = \frac{\lim_{d\to\infty}\mathbb{E}_t[s_t^{\mathcal{D}}]}{\lim_{d\to\infty}\mathbb{E}_t[s_t]}.$$

Moreover, the convergence $\mathbb{E}_t[c_{i,t}^2]\to\beta_i$ is uniform in i under Assumption 1 (bounded spectrum and noise). Therefore, we can interchange the limits $t\to\infty$ and $d\to\infty$, yielding

$$\theta_\infty = \lim_{t\to\infty}\lim_{d\to\infty}\mathbb{E}_t[\theta_t] = \lim_{t\to\infty}\lim_{d\to\infty}\frac{\mathbb{E}_t[s_t^{\mathcal{D}}]}{\mathbb{E}_t[s_t]} = \frac{\lim_{d\to\infty}\sum_{i\in\mathcal{D}}\lambda_i^2\beta_i}{\lim_{d\to\infty}\sum_{i=1}^{d}\lambda_i^2\beta_i}.$$

This completes the proof. ■

Appendix B. Numerical Simulation Results

In this subsection, we detail the parameter configurations for our numerical simulations. We consider constant-step-size SGD on a deep linear network with the following global settings:

1. Input dimension: $d = 500$, target rank $k = 50$.
2. Learning rate: $\eta = 0.003$, total steps $T = 30,000$.
3. Initialization: The matrices A are randomly initialized as positive definite symmetric matrices A with different spectral gap $m = \lambda_k/\lambda_{k+1}$. We conduct experiments across a range of values: $m \in \{5, 10, 20, 50, 100, 200, 300, 400, 500\}$.
4. Random seeds: We choose the various random seeds $\{42, 87, 568, \ldots, 4008001\}$.

The code is available via link.

B.1. General Trends across Spectral Gaps

The alignment results for different values of m are presented in Figure 5. These results demonstrate how the spectral gap influences the speed and stability of the alignment process. For each experiment group, the left panel plots loss as a function of training steps, and the right panel plots alignment as a function of training steps.

B.2. Expectation and standard deviation of the alignment for different seeds

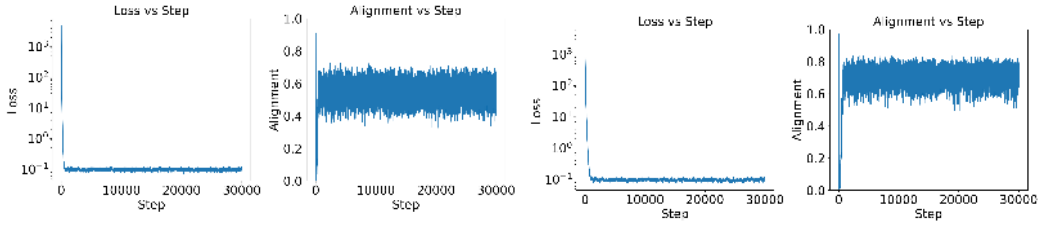
To account for randomness in initialization and SGD trajectories, we repeated the experiments across various random seeds $\{42, 87, 568, \ldots, 4008001\}$. Figure 6 illustrates the expectation and standard deviation of the alignment, confirming the consistency of our theoretical predictions.

B.3. Analysis of Phase I Alignment Decay Rate

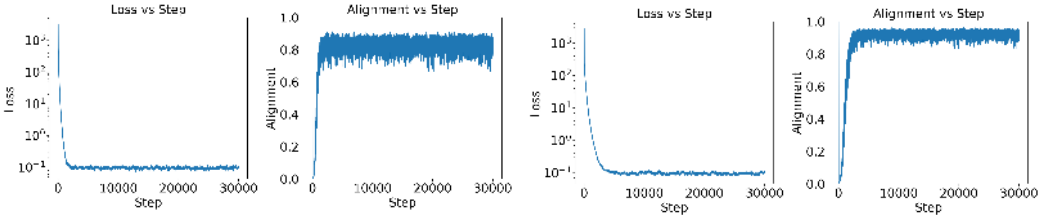
As illustrated in Figures 7 through 13, we analyze the decay rate of alignment for various seeds. The empirical results consistently indicate that the alignment exhibits a polynomial decay rate during Phase I.

B.4. Convergence Rates and Spectral Gap Scaling

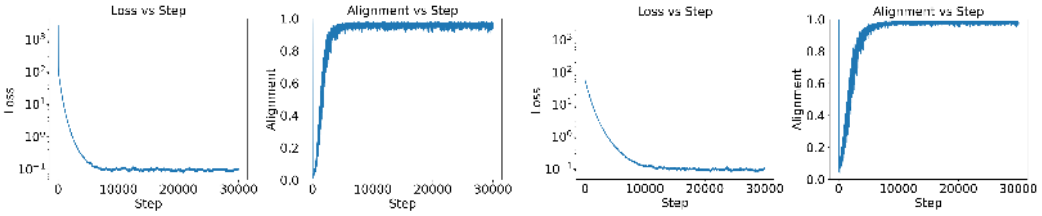
Finally, we investigate how the spectral gap m influences the final alignment state. We plot the relationship between the alignment and the parameter m in Figure 14, which shows that the alignment scales logarithmically with respect to m. This logarithmic dependency suggests that while a larger spectral gap facilitates alignment, the marginal benefit diminishes as m increases.



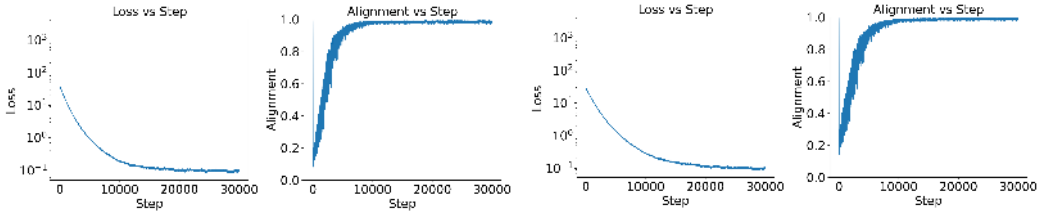
(*a*) m=5 (*b*) m=10



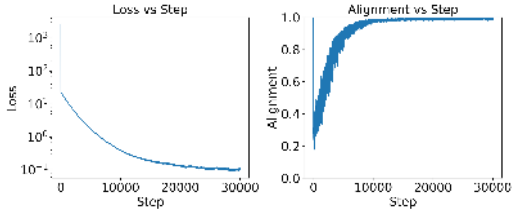
(*c*) m=20 (*d*) m=50



(*e*) m=100 (*f*) m=200

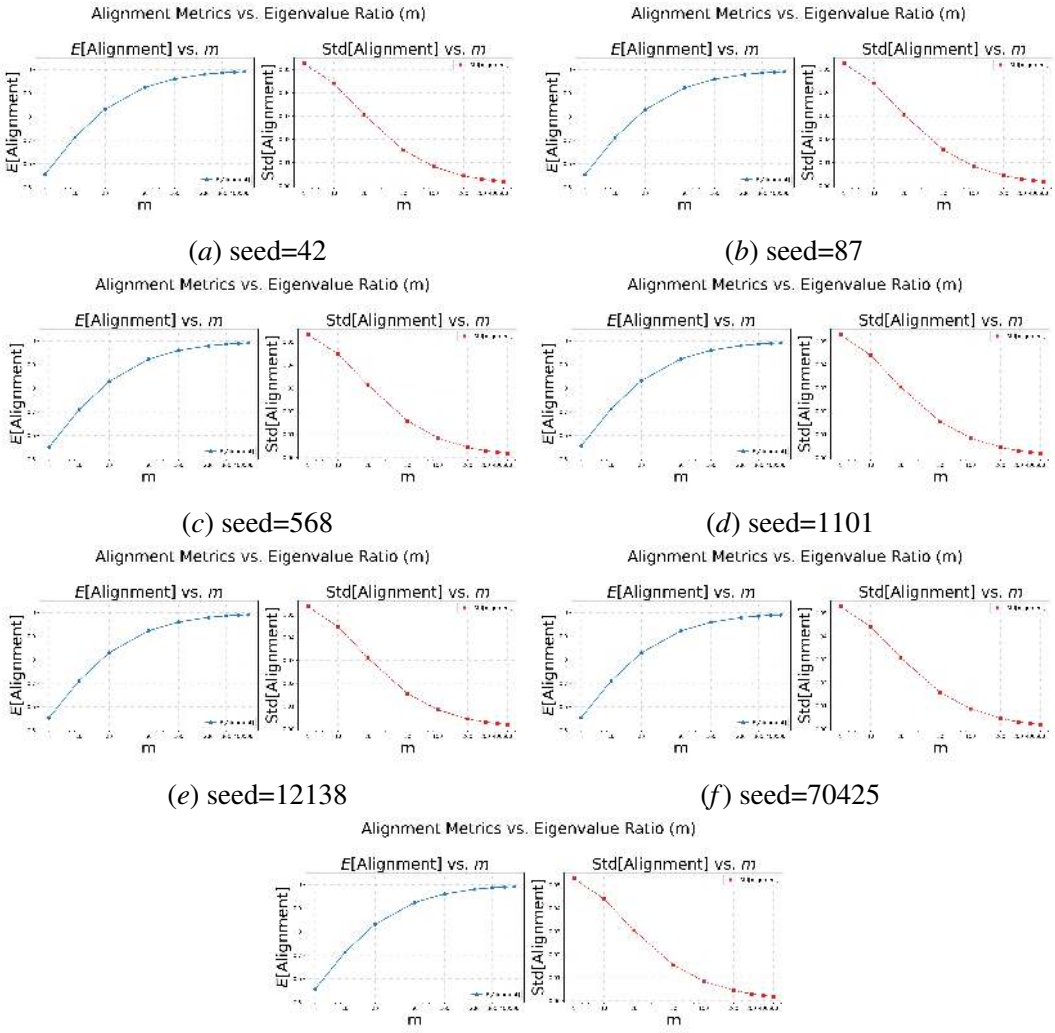


(*g*) m=300 (*h*) m=400



(*i*) m=500

Figure 5: Numerical simulation experiments with different spectral gaps ($m = \lambda_k/\lambda_{k+1}$)



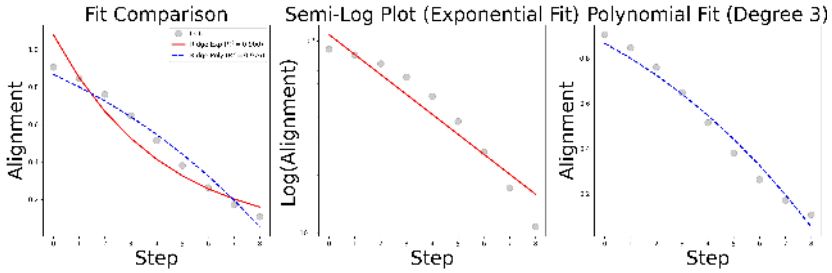
(*a*) seed=42 (*b*) seed=87

(*c*) seed=568 (*d*) seed=1101

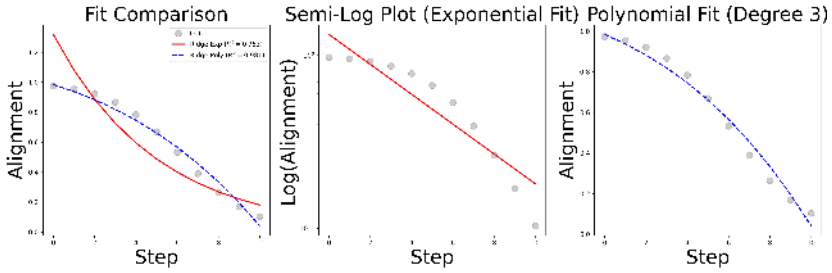
(*e*) seed=12138 (*f*) seed=70425

(*g*) seed=4008001

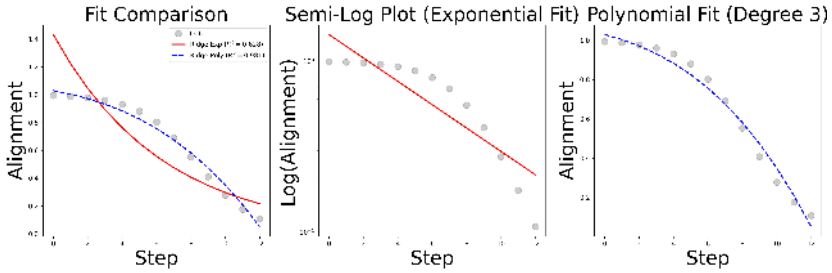
Figure 6: The limitation of the expectation and standard deviation of alignment with respect to $m = \frac{\lambda_k}{\lambda_{k+1}}$.



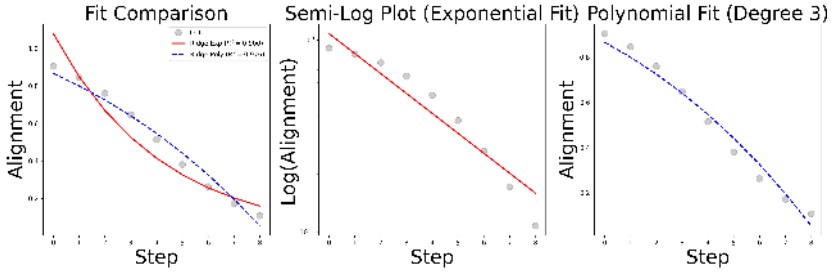
(*a*) m=5



(*b*) m=10

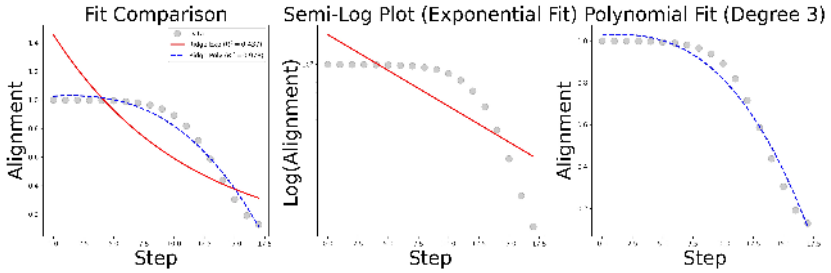


(*c*) m=20

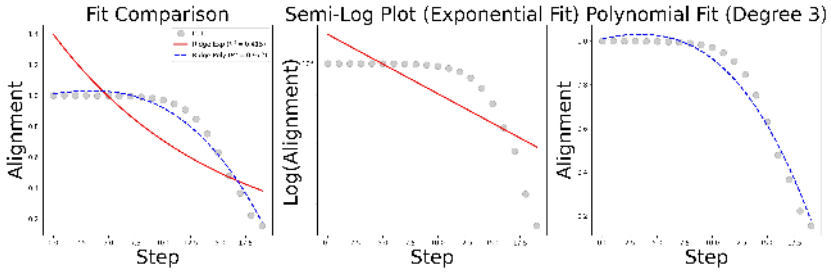


(*d*) m=50

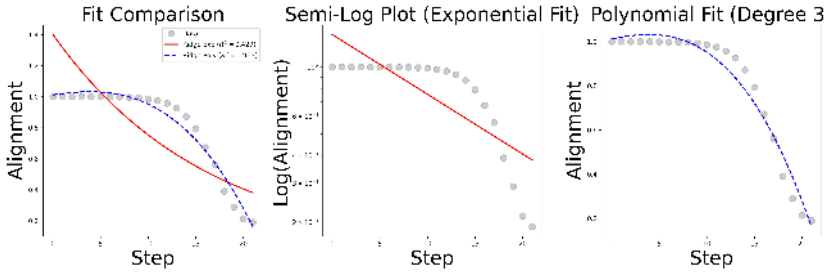
Figure 7: The decay rate of phase I when seed=42 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1).



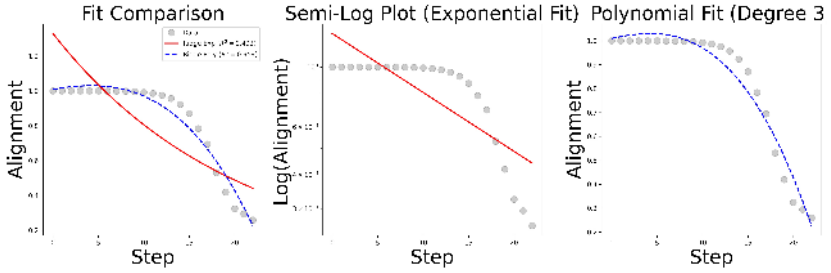
(*a*) m=100



(*b*) m=200

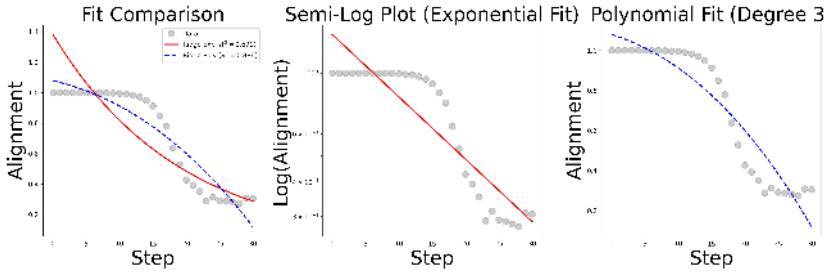


(*c*) m=300



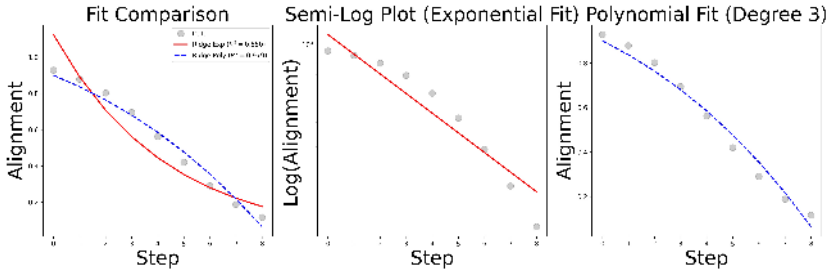
(*d*) m=400

Figure 7: The decay rate of phase I when seed=42 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 2).

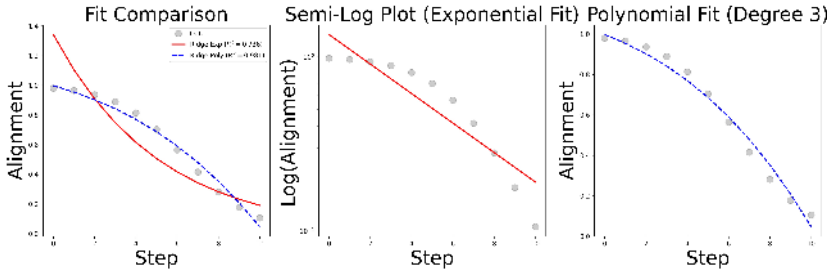


(*e*) m=500

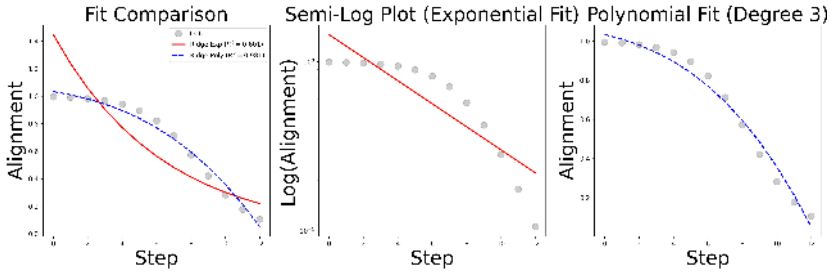
Figure 7: The decay rate of phase I when seed=42 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 3).



(*f*) m=5

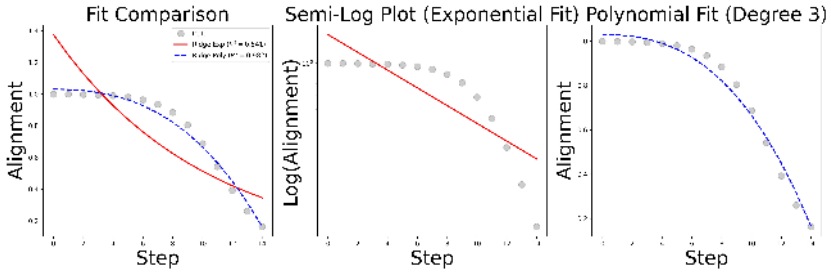


(*g*) m=10

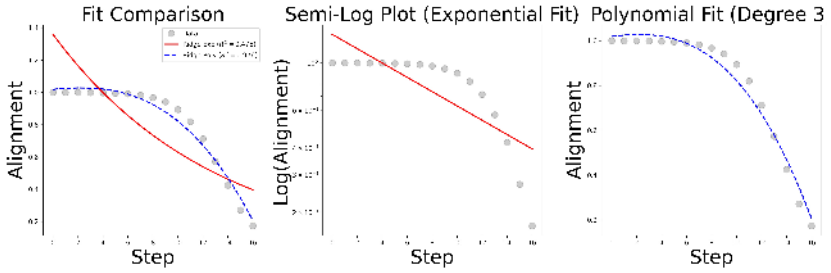


(*h*) m=20

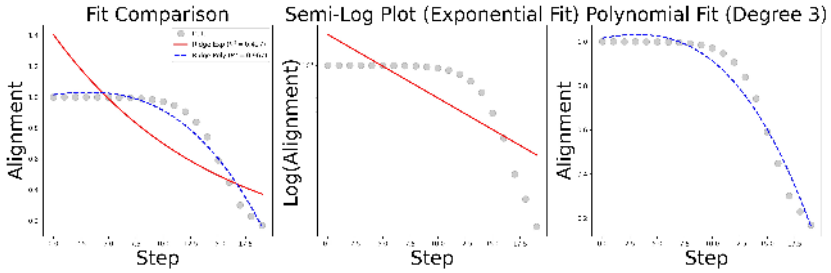
Figure 8: The decay rate of phase I when seed=87 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1)



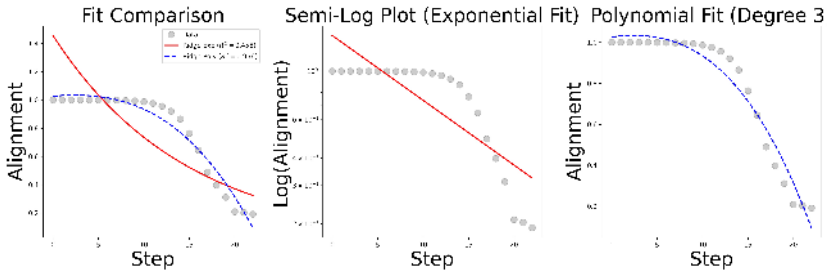
(*a*) m=50



(*b*) m=100

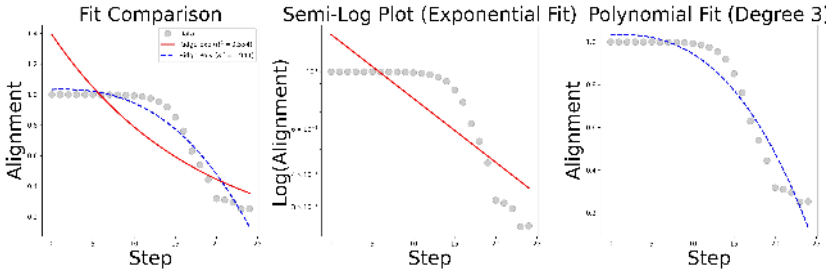


(*c*) m=200

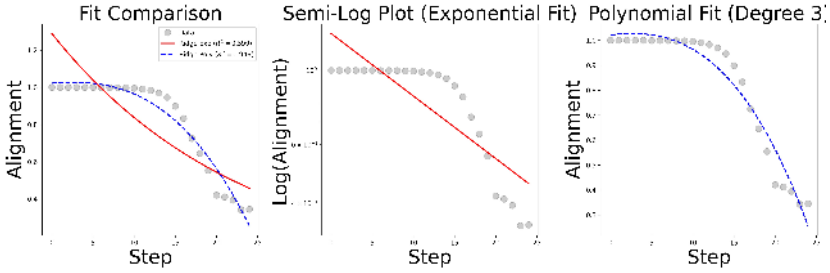


(*d*) m=300

Figure 8: The decay rate of phase I when seed=87 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 2)

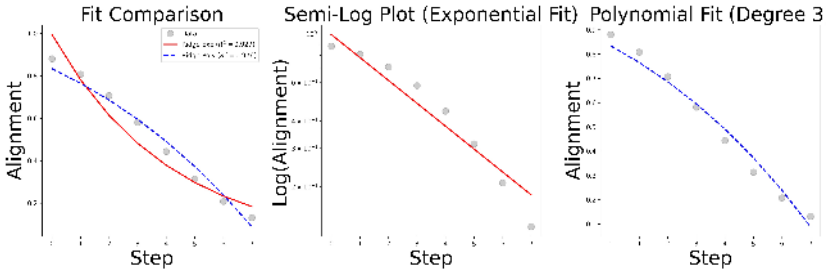


(*e*) m=400

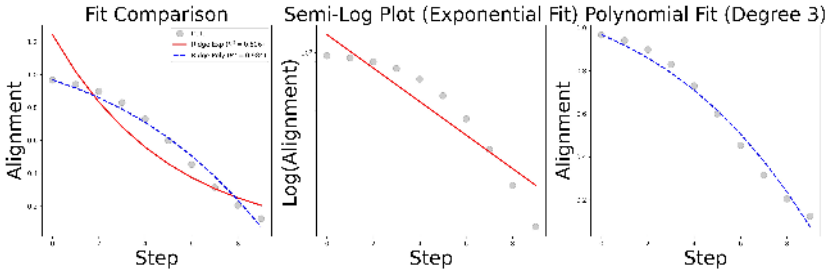


(*f*) m=500

Figure 8: The decay rate of phase I when seed=87 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 3)

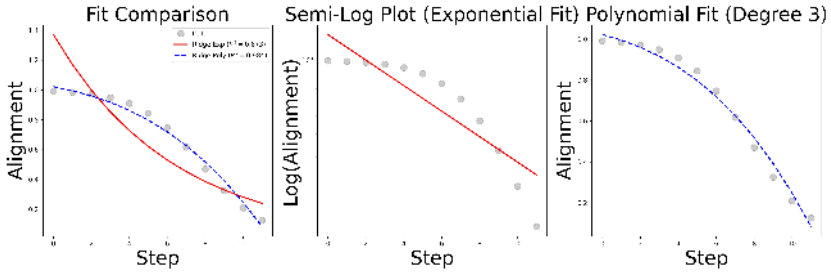


(*g*) m=5

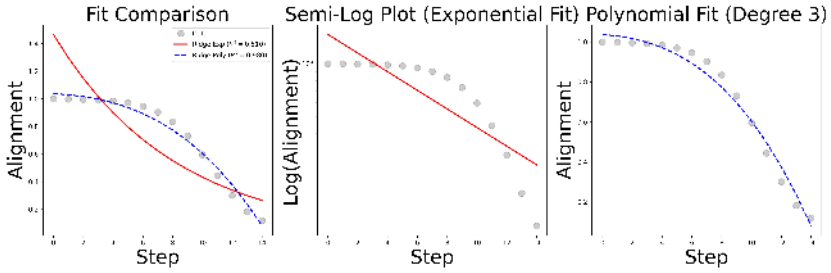


(*h*) m=10

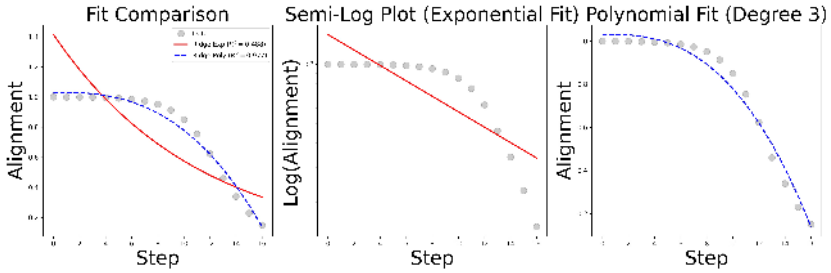
Figure 9: The decay rate of phase I when seed=568 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1)



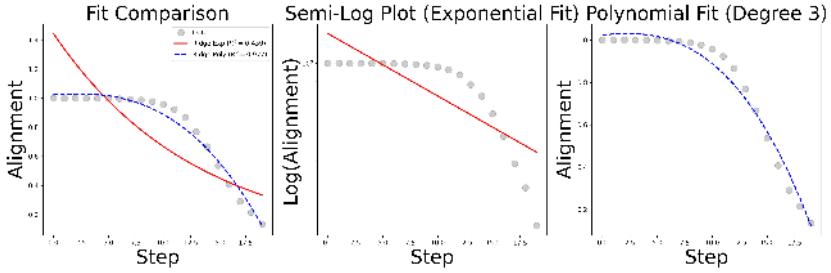
(*a*) m=20



(*b*) m=50

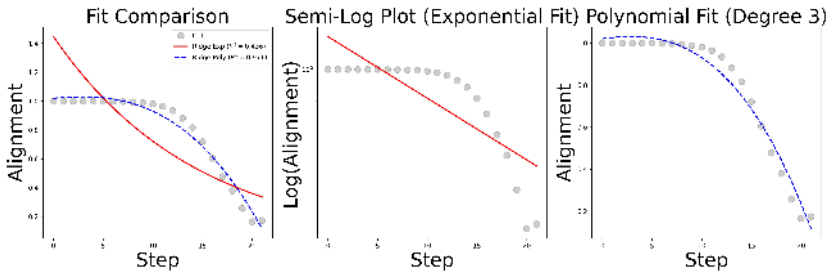


(*c*) m=100

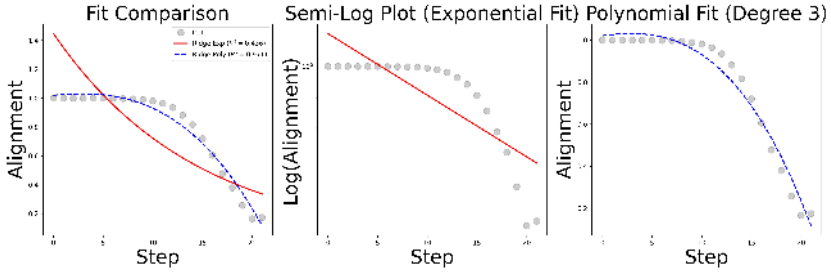


(*d*) m=200

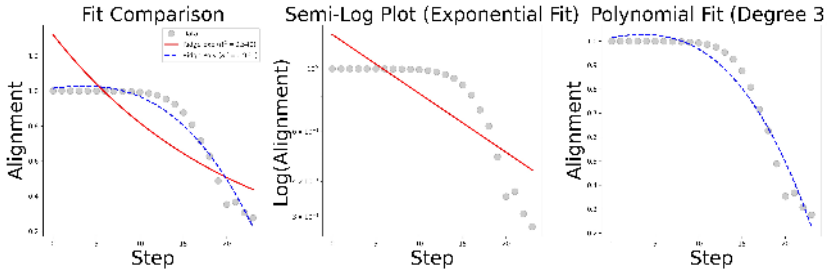
Figure 9: The decay rate of phase I when seed=568 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 2)



(*e*) m=300

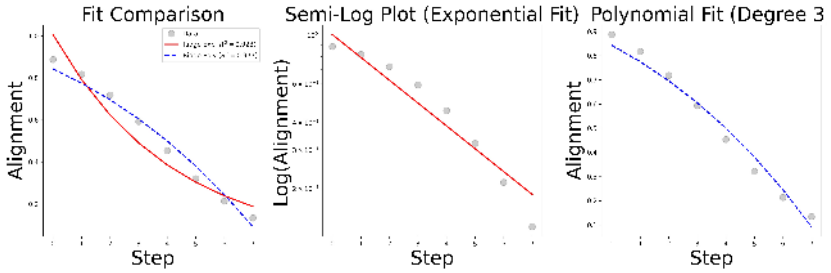


(*f*) m=400



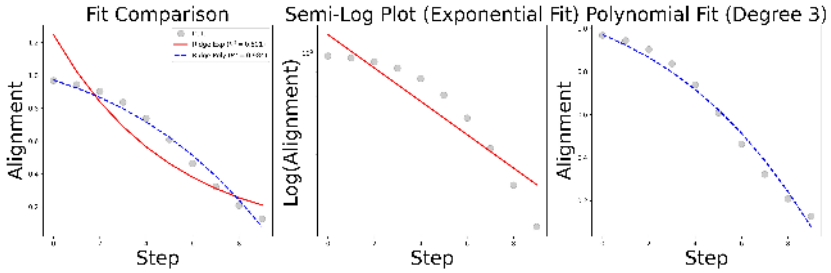
(*g*) m=500

Figure 9: The decay rate of phase I when seed=568 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 3)

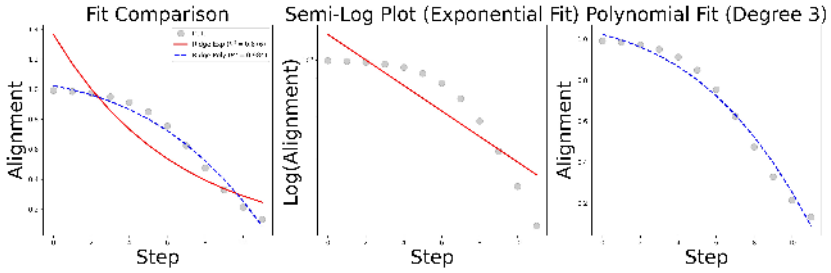


(*h*) m=5

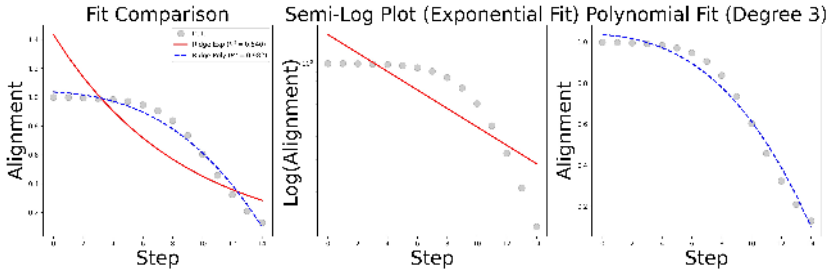
Figure 10: The decay rate of phase I when seed=1101 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1)



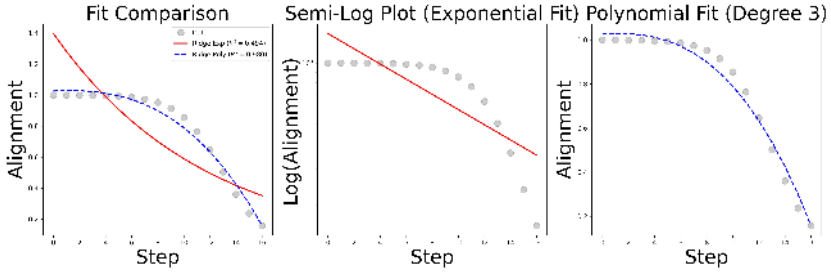
(*a*) m=10



(*b*) m=20

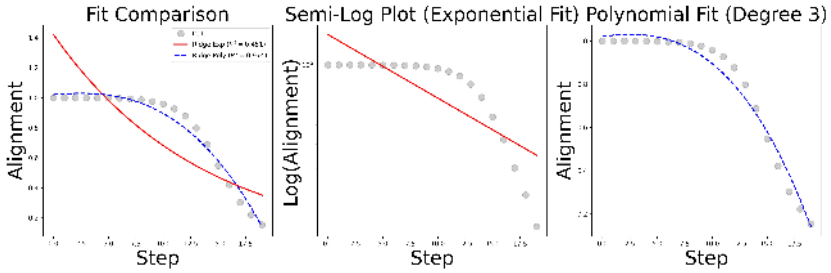


(*c*) m=50

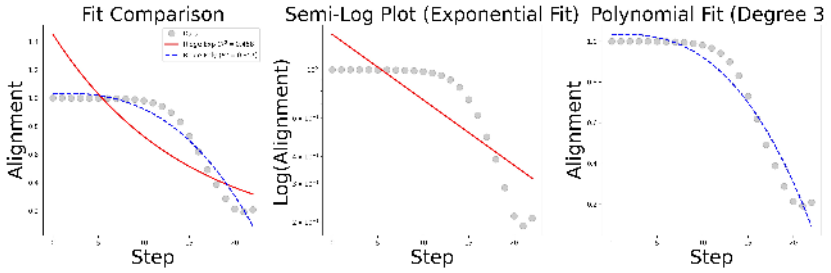


(*d*) m=100

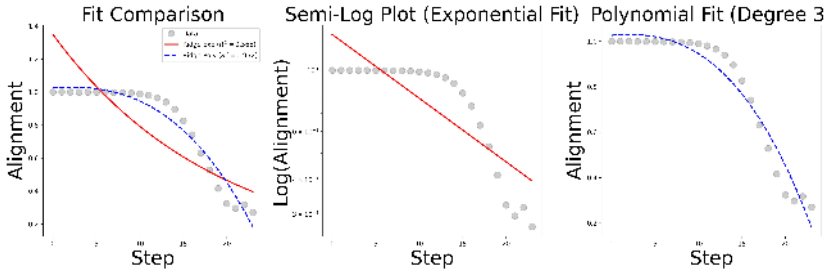
Figure 10: The decay rate of phase I when seed=1101 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 2)



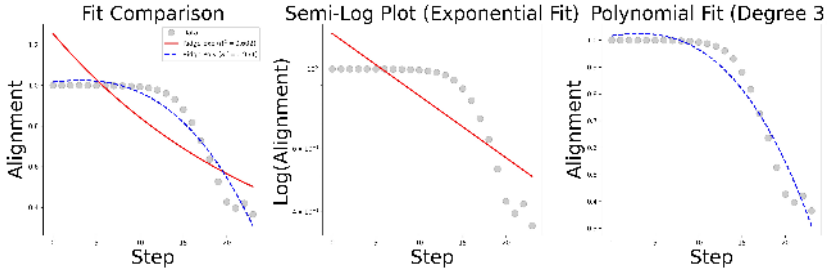
(*e*) m=200



(*f*) m=300

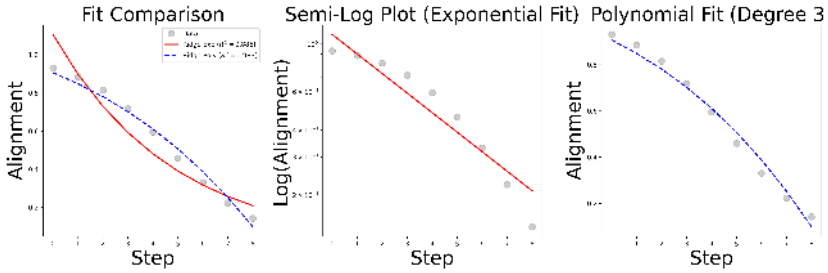


(*g*) m=400

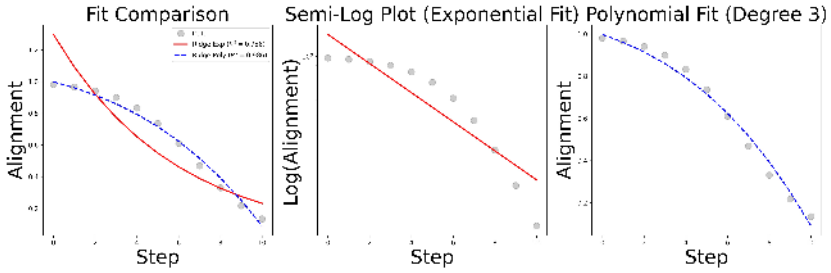


(*h*) m=500

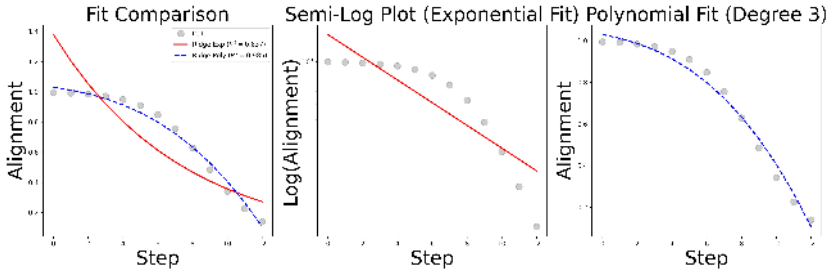
Figure 10: The decay rate of phase I when seed=1101 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 3)



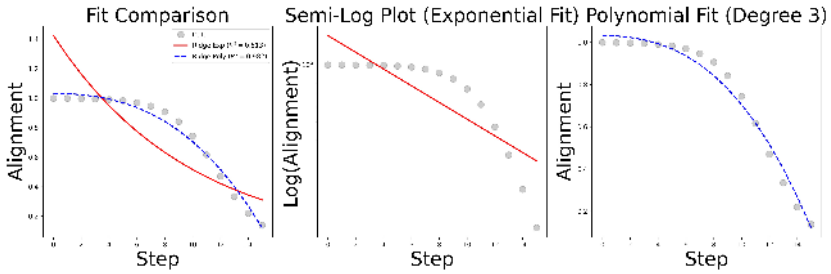
(*i*) m=5



(*j*) m=10

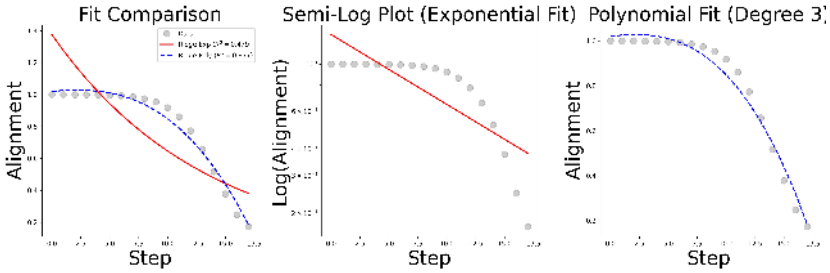


(*k*) m=20

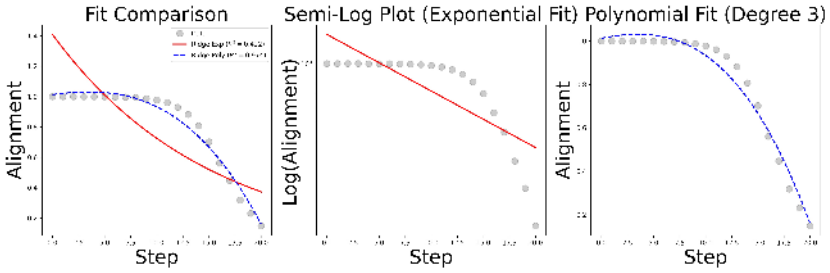


(*l*) m=50

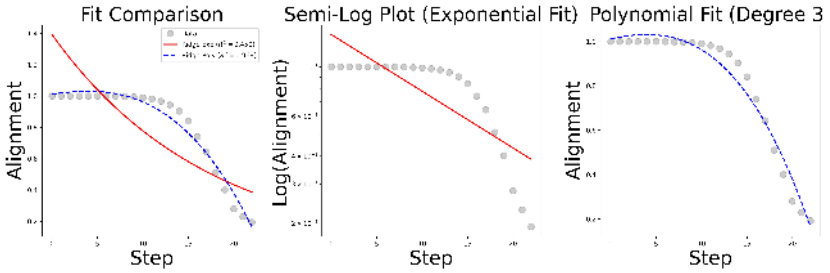
Figure 11: The decay rate of phase I when seed=12138 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1)



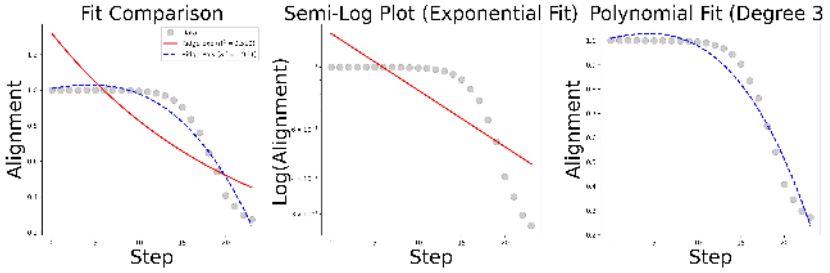
(*a*) m=100



(*b*) m=200

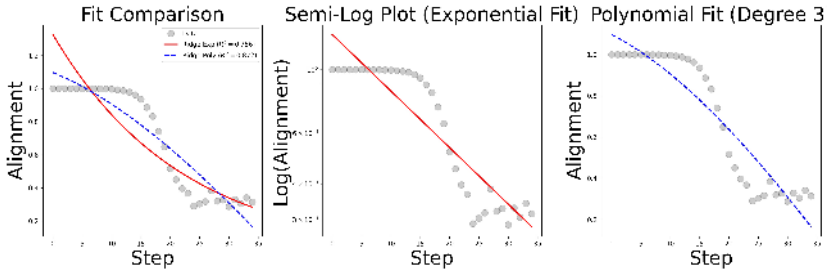


(*c*) m=300



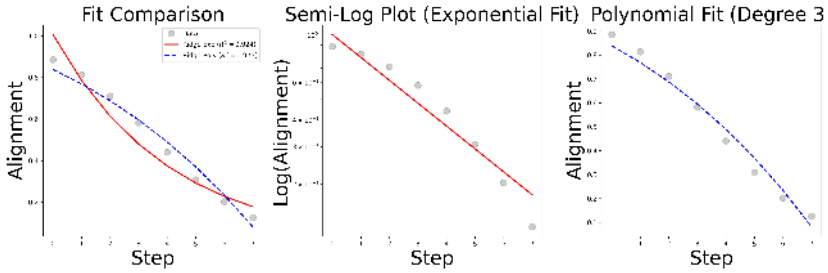
(*d*) m=400

Figure 11: The decay rate of phase I when seed=12138 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 2)

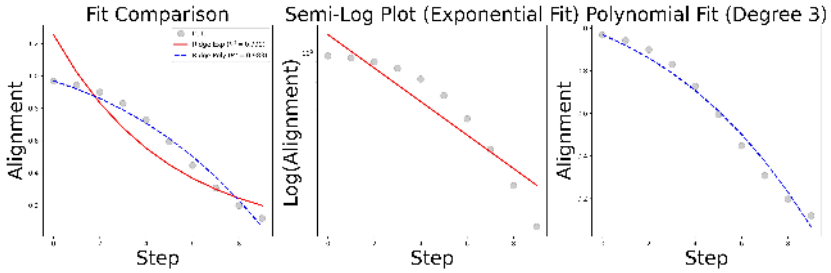


(*e*) m=500

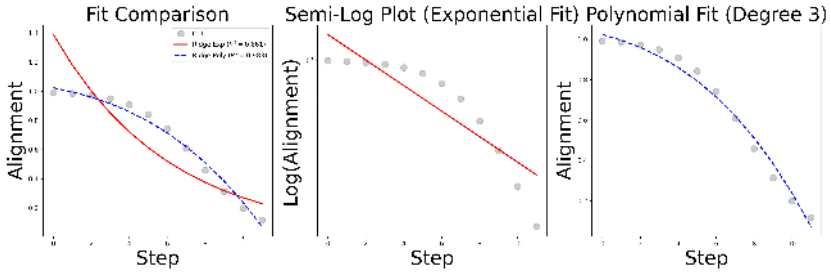
Figure 11: The decay rate of phase I when seed=12138 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 3)



(*f*) m=5

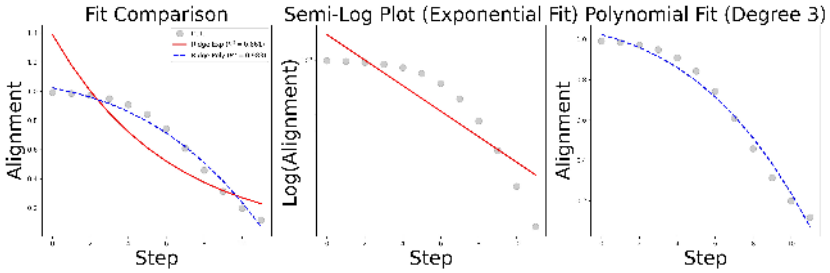


(*g*) m=10

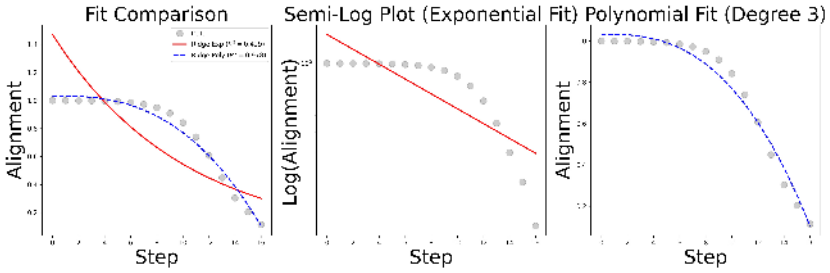


(*h*) m=20

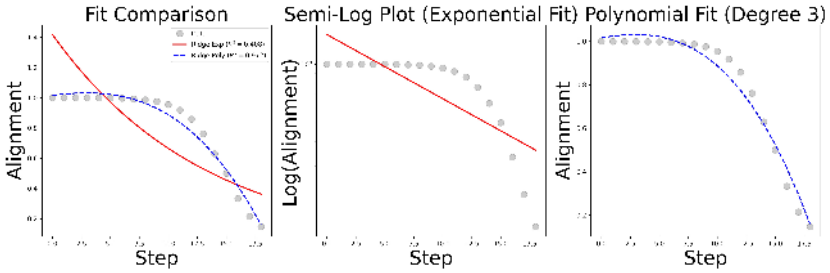
Figure 12: The decay rate of phase I when seed=70425 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1)



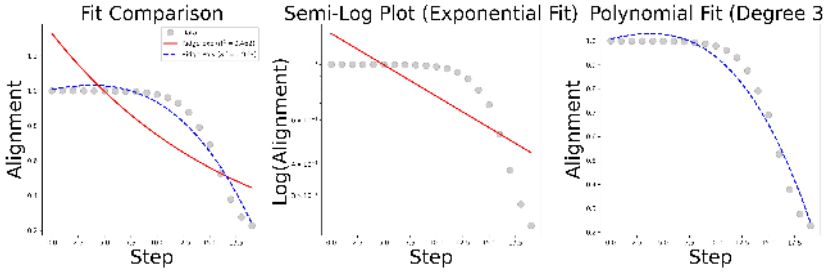
(*a*) m=50



(*b*) m=100

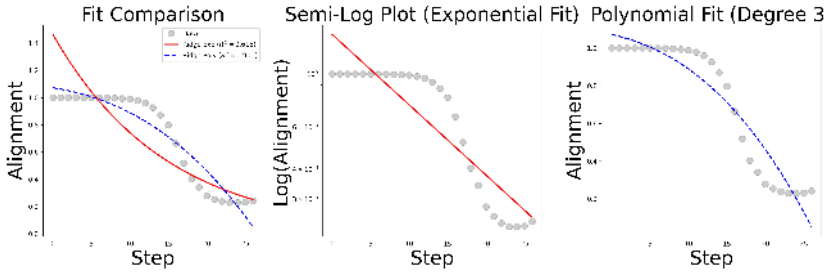


(*c*) m=200

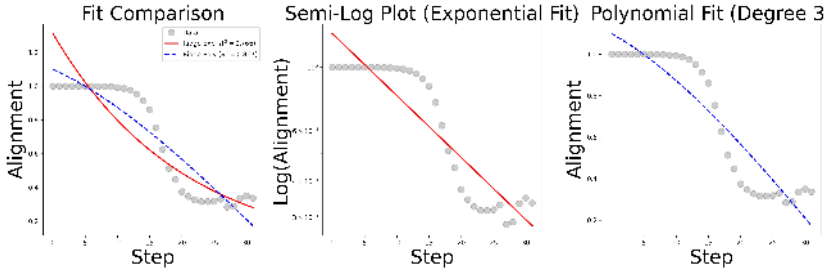


(*d*) m=300

Figure 12: The decay rate of phase I when seed=70425 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 2)

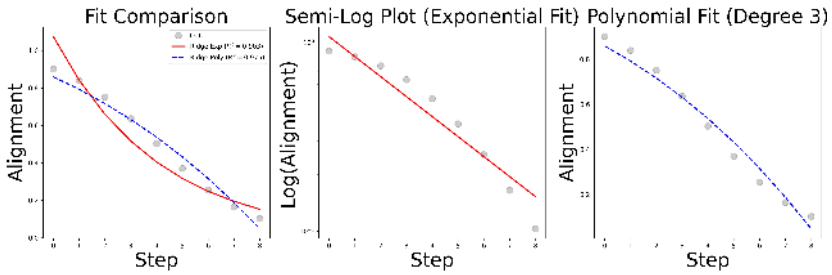


(*e*) m=400

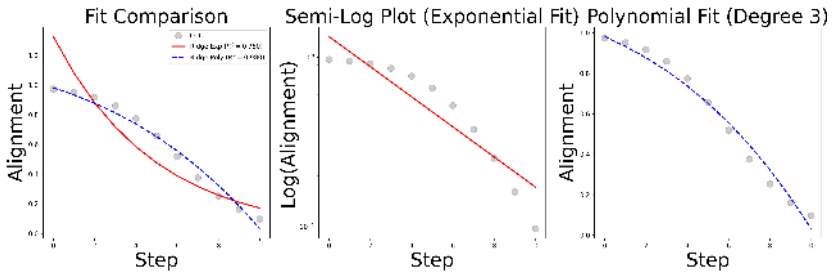


(*f*) m=500

Figure 12: The decay rate of phase I when seed=70425 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 3)

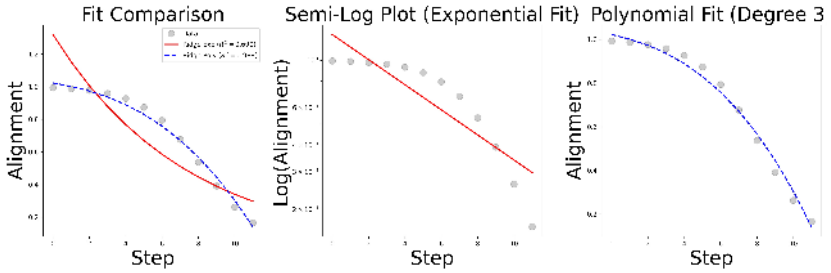


(*g*) m=5

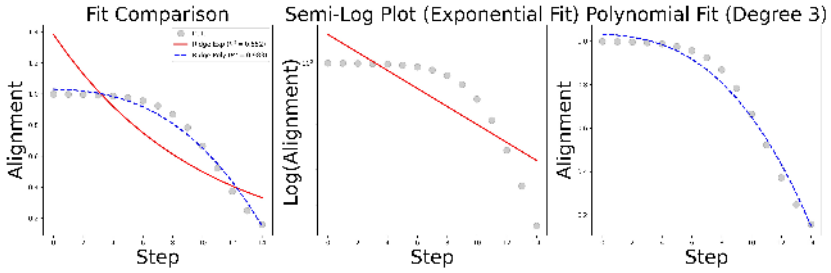


(*h*) m=10

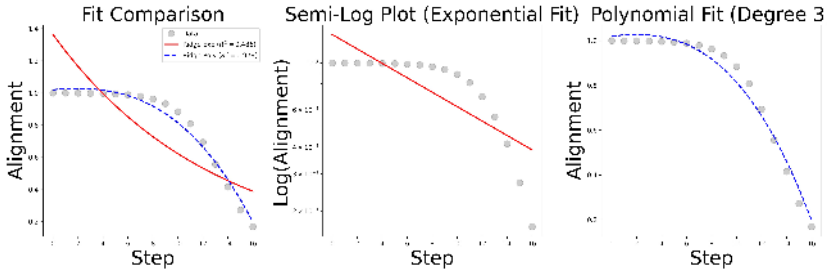
Figure 13: The decay rate of phase I when seed=4008001 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1)



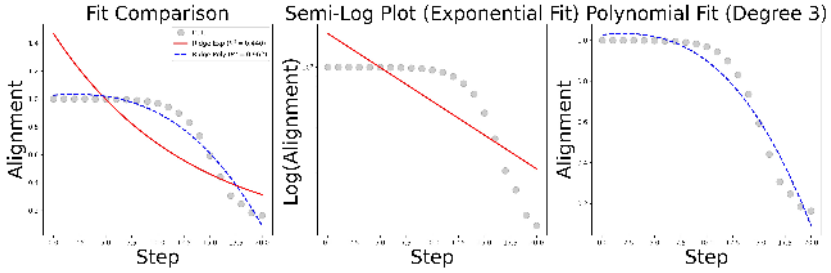
(*a*) m=20



(*b*) m=50

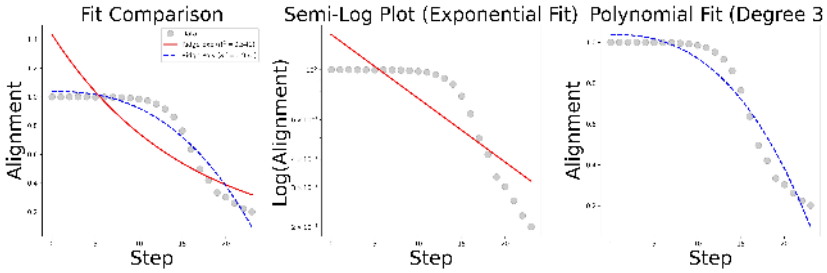


(*c*) m=100

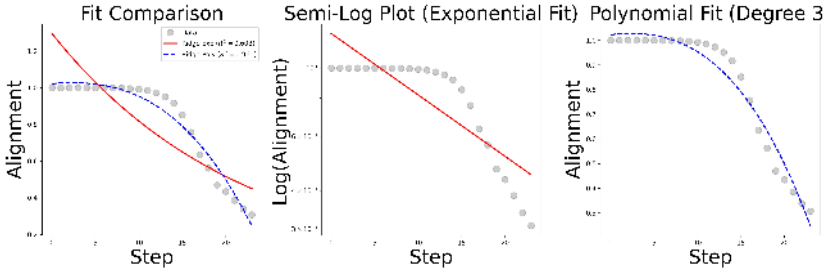


(*d*) m=200

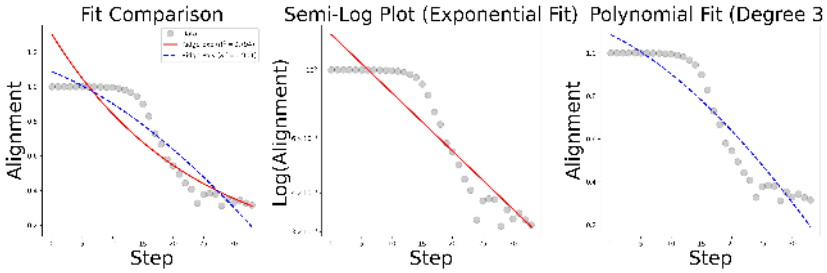
Figure 13: The decay rate of phase I when seed=4008001 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 2)



(*e*) m=300

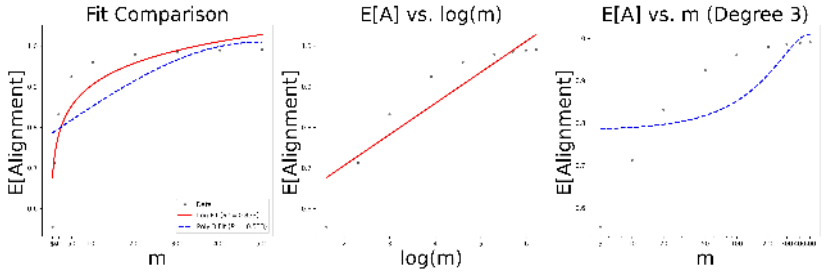


(*f*) m=400



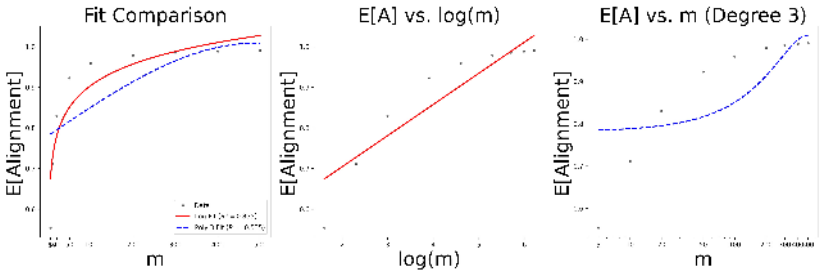
(*g*) m=500

Figure 13: The decay rate of phase I when seed=4008001 for various $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 3)

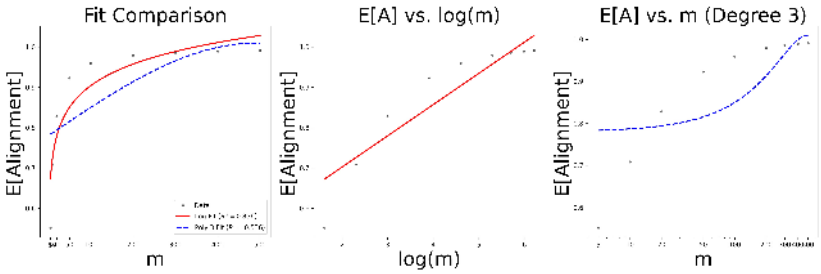


(*h*) seed=42

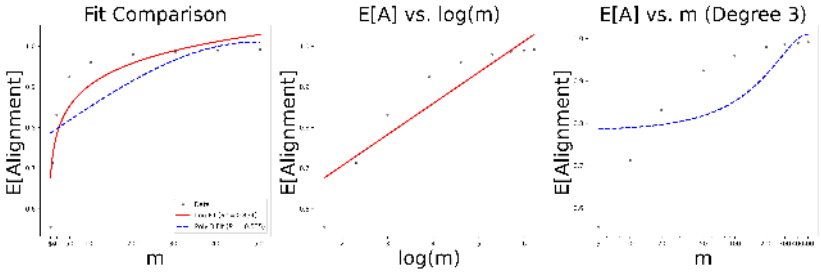
Figure 14: The rate of the expected alignment convergence with respect to $m = \frac{\lambda_k}{\lambda_{k+1}}$ (Part 1)



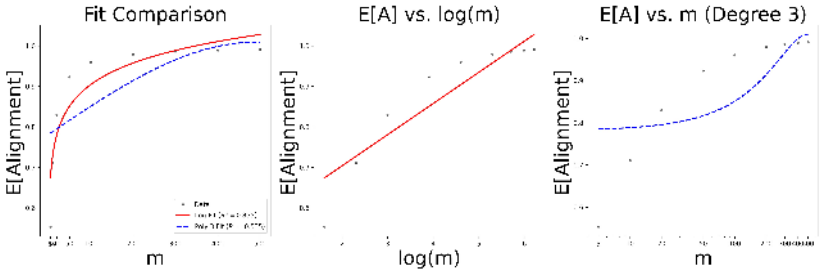
(*a*) seed=87



(*b*) seed=568

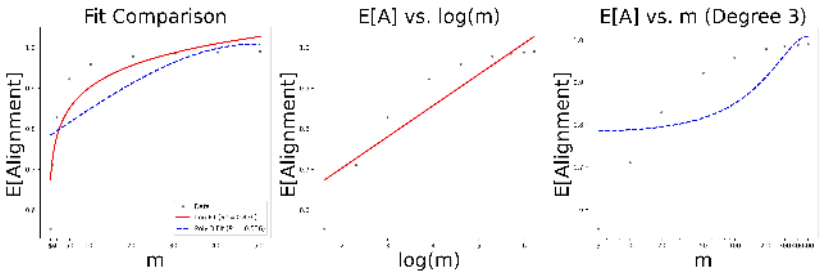


(*c*) seed=1101

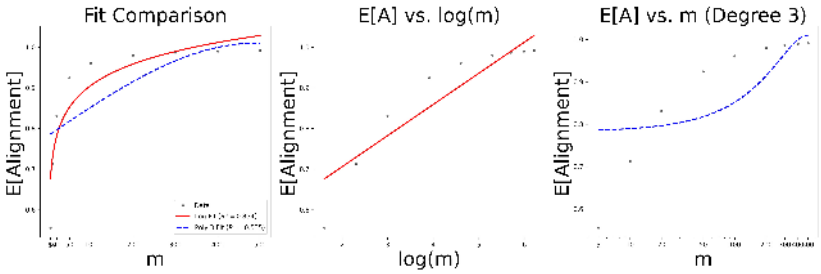


(*d*) seed=12138

Figure 14: The rate of the expected alignment convergence (Part 2)



(*e*) seed=70425



(*f*) seed=4008001

Figure 14: The rate of the expected alignment convergence (Part 3)