# Understanding the Generalization Benefits of Late Learning Rate Decay

**Yinuo Ren**
Stanford University

**Chao Ma**
Stanford University

**Lexing Ying**
Stanford University

## Abstract

Why do neural networks trained with large learning rates for a longer time often lead to better generalization? In this paper, we delve into this question by examining the relation between training and testing loss in neural networks. Through visualization of these losses, we note that the training trajectory with a large learning rate navigates through the minima manifold of the training loss, finally nearing the neighborhood of the testing loss minimum. Motivated by these findings, we introduce a nonlinear model whose loss landscapes mirror those observed for real neural networks. Upon investigating the training process using SGD on our model, we demonstrate that an extended phase with a large learning rate steers our model towards the minimum norm solution of the training loss, which may achieve near-optimal generalization, thereby affirming the empirically observed benefits of late learning rate decay.

## 1 INTRODUCTION

During the training of deep neural networks, one of the challenges faced by optimization algorithms arises from the intricate misalignment between the training and testing losses. This discrepancy becomes particularly evident in overparameterized settings, in which case merely minimizing the training loss does not necessarily translate to desirable testing performance. Nonetheless, in practical scenarios, neural networks often demonstrate impressive generalization when trained using stochastic gradient descent (SGD) (Allen-Zhu and Li, 2019; Kleinberg et al., 2018; Pesme et al., 2021). This relative ease of training can be attributed, at least in part, to the development and adoption of several techniques over the years, such as normalization (Salimans and Kingma, 2016; Ba et al., 2016), adaptive optimization (Duchi et al., 2011; Kingma and Ba, 2014), and learning rate schemes (Smith, 2017; Loshchilov and Hutter, 2017).

It is a widely accepted observation that implementing late learning rate decay, *i.e.* maintaining a large learning rate with SGD for an extended period even after the stabilization of the training loss, can enhance generalization performance (Li et al., 2019b; Wu et al., 2020a; Wang et al., 2021; Beugnot et al., 2022). Numerous theoretical studies have explored and interpreted related phenomena are rich, and we direct readers to Section 1.2 for a brief review. For example, Wu et al. (2020a) studies linear regression and argues that the "directional bias" of SGD with a large learning rate boosts final testing performance, while Li et al. (2019b) emphasizes the mismatch between learnability and generalizability can explain the necessity of an initial large learning rate. However, there is still a noticeable gap in the literature regarding the interplay between training and testing losses and their landscape formations.

Our study is motivated by specific observations from our visualization of the training and testing loss landscapes of neural networks, as depicted in Figure 2:

- The training loss landscape displays a *minima manifold* characterized by open level sets, while the testing loss landscape presents an isolated minimum with closed level sets.
- With a large learning rate, the training trajectory navigates through the minima manifold of the training loss towards the neighborhood of the testing loss minimum.

Several previous works contend that the traversal of the minima manifold is driven by the flatness of the training loss landscape (Wu et al., 2018; Mulayoff

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

et al., 2021; Nacson et al., 2022). Yet, the question lingers as to why the training loss landscape is inherently flatter around the testing loss minimum.

In this work, we scrutinize the relation between training and testing loss landscapes with neural networks as nonlinear overparameterized models and the implication for training behaviors. We propose a simple nonlinear model whose loss landscapes mirror our visualizations from neural networks. Our model can be interpreted as follows: Starting with an overparametrized linear regression model, the testing loss, as an expectation of quadratic functions, yields an isolated minimum, while the null space of the training data produces a minima manifold. Then, a transformation motivated by the depth of the neural network is applied to both losses within the parameter space, resulting in non-quadratic landscapes exhibiting varying flatness.

We then study the training process of our model using SGD via a continuous-time analysis. Our findings suggest that this process can be divided into three phases: (I) an initial phase with a large learning rate, (II) an extended phase maintaining the large learning rate, and (III) a final phase with a decayed learning rate. We prove that with high probability, Phase II propels the model towards the minimum $L^2$-norm solution of the training loss, which has long been believed to be the near-optimal solution for overparametrized models (Wu and Xu, 2020; Bartlett et al., 2020), thereby affirming the empirically observed benefits of late learning rate decay.

### 1.1 Contribution

Our main contributions in this paper are summarized as follows:

- Through experiments, we empirically demonstrate the generalization advantages of late learning rate decay. Further, we offer visualizations of the training and testing loss landscapes of neural networks, illustrating the interrelation between these two loss landscapes.
- We introduce a nonlinear overparameterized model that recovers the loss landscape behaviors observed in real neural networks. Our insights suggest that the flatness of the training loss landscape near the testing loss minimum is intrinsically linked to the depth of the neural networks.
- We systematically dissect the training process of our model into three phases and show that extended training using a large learning rate helps find the minimum $L^2$-norm solution of the training loss and thus corroborates the provable benefits of late learning rate decay.

### 1.2 Related Works

**Implicit Regularization.** The implicit regularization effect of optimization with SGD has been studied extensively in previous works (Mandt et al., 2016; Hoffer et al., 2017; Kleinberg et al., 2018). Many works argue that SGD picks flat minima (Keskar et al., 2016; Du et al., 2019; Wu et al., 2022), which boosts the generalization performance (Hochreiter and Schmidhuber, 1997; Zhou et al., 2020).

To understand the behavior of SGD, several mathematical models have been proposed and studied, including the stochastic differential equations (SDEs) (Li et al., 2017, 2019a, 2021a; Mori et al., 2022), and Langevin dynamics (Welling and Teh, 2011; Raginsky et al., 2017; Zhang et al., 2017; Chen et al., 2020). Recent works including (Blanc et al., 2020; Pesme et al., 2021; HaoChen et al., 2021; Damian et al., 2021; Even et al., 2023) adopts the *diagonal linear networks* to study implicit regularization.

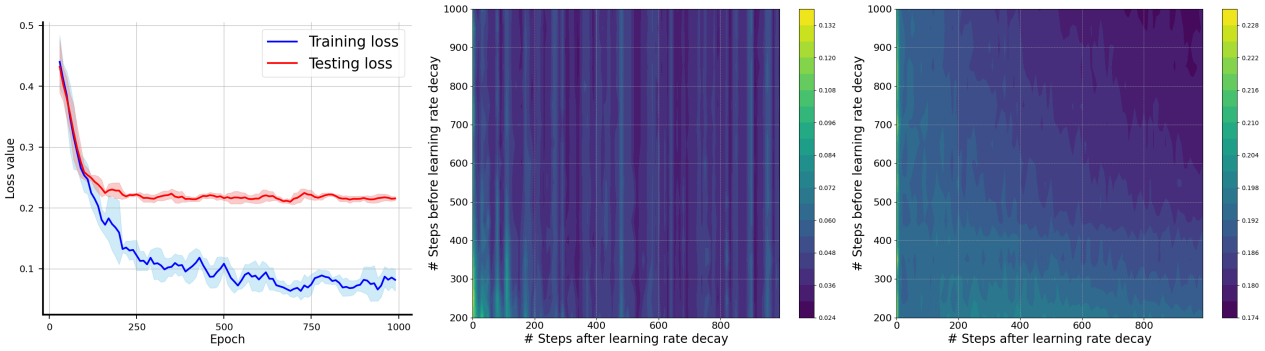
**SGD Scheduling.** Another line of the research on SGD attempts to gain a deeper understanding of how the choice and scheduling of the learning rate affects the performance of SGD (Smith and Le, 2017; Jastrzębski et al., 2018; Wu et al., 2018; Li et al., 2020; Lyu and Li, 2019; Mulayoff et al., 2021; Nacson et al., 2022; Li et al., 2022). Large learning rates are shown to be beneficial for generalization both empirically and theoretically (Li et al., 2019b; Wu et al., 2020a; Wang et al., 2021; Andriushchenko et al., 2022). Intricate designs of learning rate schedules have also been proved to achieve faster convergence rate for gradient descent (Smith, 2017; Agarwal et al., 2021; Grimmer, 2023).

The study of the effect of large learning rates is also closely related to the topic of the *Edge of Stability (EoS)* phenomenon (Cohen et al., 2021; Arora et al., 2022; Damian et al., 2022; Zhu et al., 2022; Beugnot et al., 2022; Ma et al., 2022; Chen and Bruna, 2022), and *grokking* (Power et al., 2022; Liu et al., 2022; Žunkovič and Ilievski, 2022).

## 2 MOTIVATING EMPIRICAL OBSERVATIONS

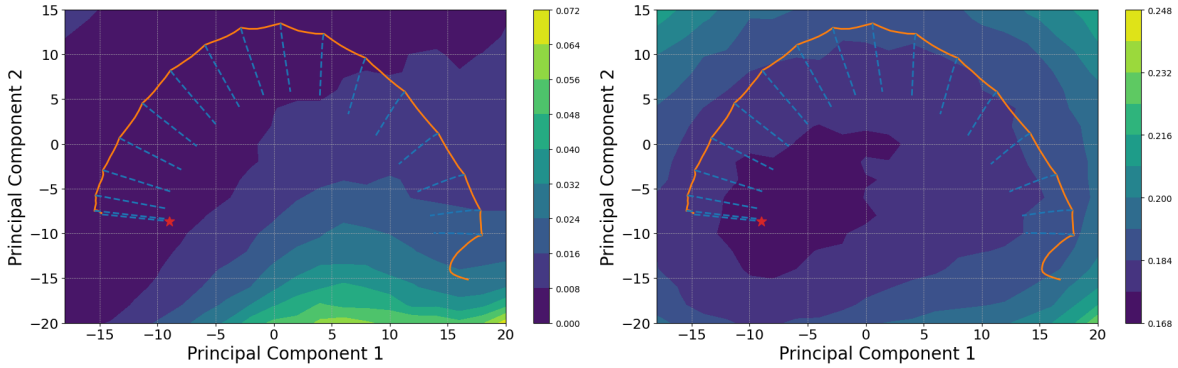
In this section, we present experiment results to observe the training behaviors in neural networks. We further visualize the training and testing loss landscapes obtained by performing PCA to the parameters on the training trajectories and discuss its implications on the training process[1].

[1]Code is accessible at `https://github.com/yinuoren/landscape`




(a) Learning curves of initial phase (b) Training loss (c) Test loss

Figure 1: Bahaviors of the training and testing losses for a VGG-11 model trained on the CIFAR-10 dataset under various learning rate schedules. Panel (a) showcases the learning curves of the main path with a learning rate of 0.1. In panels (b) and (c), the $y$-axis represents the number of epochs before the learning rate decay, and the $x$-axis indicates the number of epochs after the decay. Each slice parallel to the $x$-axis illustrates the learning curve of a subpath originating from the same main path as shown in (a) with a learning rate of 0.01.




(a) Training loss landscape (b) Test loss landscape

Figure 2: Visualization of the training and testing loss landscapes for a VGG-11 model trained on the CIFAR-10 dataset. The main path with the initial learning rate of 0.1 is represented by an orange line, while the subpaths with the reduced learning rate of 0.01 are depicted in blue dashed lines. The final point of the most extended training trajectory that spans 2000 epochs is marked with a red star.

### 2.1 Observations from Training Behaviors

We focus on the image classification task on the CIFAR-10 dataset, one of the widely adopted and studied examples in the theory of neural networks (Smith et al., 2017; Ma et al., 2022; Jelassi and Li, 2022), and recently also in the literature of the Edge of Stability (EoS) (Cohen et al., 2021; Arora et al., 2022). We employ a VGG-11 model (Simonyan and Zisserman, 2014), trained with the SGD optimizer. The training spans 1000 epochs with a batch size of 128 and an initial learning rate of 0.1, to which we will refer as the *main path*. Starting from epoch 200 of the main path, we introduce additional *subpaths* of other 1000 epochs, during which the learning rate is reduced to 0.01. Due to the large number of subpaths, we only plot the learning curves of the main path (shown in Figure 1a) and present the behaviors of the training and testing losses of subpaths in the form of contour maps in Figure 1b and 1c.

As depicted in Figure 1, the timing of the learning rate decay has a small impact on the subsequent training phase (slices parallel to the $x$-axis) with the reduced learning rate. This becomes particularly clear when the training loss stabilizes after the initial $\sim$500 epochs, as evidenced by the seemingly chaotic pattern in Figure 1b. In contrast, the testing loss reveals a distinct correlation with the timing of learning rate decay. As illustrated in Figure 1c, delaying the reduction in learning rate results in a lower final testing loss. This observation is consistent with the findings of Wu et al. (2020a); Li et al. (2019b); Andriushchenko et al. (2022) that training with a larger learning rate

enhances generalization.

### 2.2 Visualization of Training and Testing Landscape

In order to analyze this phenomenon, we proceed by examining the loss landscapes of the neural networks. As illustrated in Figure 2, we visualize the training and testing loss landscapes for the model. We derive these loss landscapes based on the previous works of Lorch (2016); Li et al. (2018): (a) Collating and flattening the model's parameters at each epoch's conclusion for both the main path and subpaths, (b) Applying Principal Component Analysis (PCA) on these parameters to extract the two primary components, and (c) Calculating the training and test losses over a grid defined by these two principal components, centered around the mean when performing the PCA.

The primary two principal components account for 77.93% of the total variances, with each subsequent component explaining less than 5%. It is important to note that given the high dimensionality of neural networks, the full parameter space might exhibit more complex features not captured in our 2D representation. Furthermore, our visualization only aims to present the local landscape around the minimizers rather than a global one as in Li et al. (2018). Nevertheless, our findings still offer valuable insights into the structure of the loss landscapes and can be instrumental in explaining practical training behaviors, as we discuss subsequently.

### 2.3 Intuitive Cause of the Landscape Structure

As showcased in Figure 2a, the training loss landscape reveals a low-dimensional manifold of minima (with near-zero training loss). This coincides with the prevalent belief that neural networks, being overparameterized relative to the training sample count, manifest a zero-loss manifold that all training trajectories converge to and then oscillate around (Cooper, 2018, 2020; Li et al., 2021b). Recent studies (Wu et al., 2020a; Li et al., 2019b; Andriushchenko et al., 2022) demonstrate that the implicit bias of SGD with an initially large learning rate offers generalization advantages. This observation is exactly captured in Figure 2b. Contrasting the training loss landscape, the test loss landscape showcases an isolated minimum, with closed loss level sets encircling the minimum. While the training trajectory with a larger learning rate (denoted in orange) traverses the minima manifold of the training loss, its correlation with test loss reduction is tenuous. However, this traversal does lead to a lower final testing loss when the learning rate decays later (indicated in blue).

Based on these empirical findings, we hypothesize that despite the misalignment of the testing loss minimum and the minima manifold of the training loss, longer training with larger learning rates helps find the testing loss minimum. Then, with the decayed learning rate, the trajectory achieves better final generalization. The argument in Wu et al. (2018); Mulayoff et al. (2021); Nacson et al. (2022); Andriushchenko et al. (2023) for this phenomenon is that SGD with a larger learning rate identifies flatter minima of the training loss (characterized by the Hessian matrix), which are believed to generalize better (Keskar et al., 2016; Zhou et al., 2020). However, this argument is not sufficient to explain the observed discrepancy between the training and testing loss landscapes.

## 3 AN ILLUSTRATIVE MODEL

In this section, we consider a minimal nonlinear model that helps provide intuition for neural networks with more complicated architectures in the aforementioned phenomenon. We argue that the observed loss landscapes are caused by the nonlinearity of neural networks resulting from the composition of the layers.

### 3.1 Model Settings

Let $\boldsymbol{x} \in \mathbb{R}^d$ be the feature vector, $y \in \mathbb{R}$ be the label, and $\boldsymbol{w} \in \mathbb{R}^d$ be the model parameter. Our model is defined as

$$y = \|\boldsymbol{w}\|^\gamma \boldsymbol{w}^\top \boldsymbol{x} := \boldsymbol{\alpha}^\top \boldsymbol{x}, \tag{3.1}$$

where $\|\cdot\|$ denotes the $L^2$-norm, $\gamma \geq 0$ is a parameter reflecting the depth of the neural network and controlling the shrinkage, which we will explain afterward. $\boldsymbol{\alpha} = \|\boldsymbol{w}\|^\gamma \boldsymbol{w}$ denotes the effective linear predictor when interpreting Equation 3.1 as a reparametrization of the linear regression model. It is worth noting that the mapping between $\boldsymbol{w}$ and $\boldsymbol{\alpha}$ is invertible with $\|\boldsymbol{\alpha}\|^{-\frac{\gamma}{1+\gamma}} \boldsymbol{\alpha} = \boldsymbol{w}$. Thus, when no confusion arises, this bijection is always implicitly assumed, *i.e.* $\boldsymbol{\alpha} = \boldsymbol{\alpha}(\boldsymbol{w})$ and $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{\alpha})$. We adopt the quadratic loss defined as

$$\ell(\boldsymbol{x}, y; \boldsymbol{w}) = \frac{1}{2}(\|\boldsymbol{w}\|^\gamma \boldsymbol{w}^\top \boldsymbol{x} - y)^2.$$

Suppose we have a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ of $n$ samples, constructed with a ground truth weight $\boldsymbol{w}^*$, *i.e.* $y_i = \|\boldsymbol{w}^*\|^\gamma (\boldsymbol{w}^*)^\top \boldsymbol{x}_i := (\boldsymbol{\alpha}^*)^\top \boldsymbol{x}_i$ for $i \in [n]$. Define the data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, the training loss (or the empirical loss) can be expressed as

$$\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) = \frac{1}{2n} \left\| \boldsymbol{X}^\top (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \right\|^2.$$

We further assume the data are drawn from a standard Gaussian distribution, *i.e.* $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I})$, for $i \in [n]$, the testing loss (or the population loss) is thus

$$\begin{aligned}&\mathcal{L}(\boldsymbol{w}; \boldsymbol{w}^*) = \mathbb{E}_{\boldsymbol{x}_i \sim \mathcal{N}(0,\boldsymbol{I})}[\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*; \mathcal{D})] \\ =&\frac{1}{2n}\mathbb{E}_{\boldsymbol{X}}\left[(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top \boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\right] = \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|^2.\end{aligned}$$

Now that we consider the over-parametrized regime, we assume $d > n$ and $\boldsymbol{X}$ has full column rank. We will also denote the space spanned by the columns of $\boldsymbol{X}$ by $\boldsymbol{X}$ and the orthogonal complement of $\boldsymbol{X}$ by an orthogonal matrix $\boldsymbol{X}^\perp$. For any vector $\boldsymbol{x}$, we denote its normalized version as $\overline{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|$. For any matrix $\boldsymbol{Y}$, define the projection operator $\mathcal{P}_{\boldsymbol{Y}} = \boldsymbol{Y}(\boldsymbol{Y}^\top\boldsymbol{Y})^{-1}\boldsymbol{Y}^\top$ as the projection onto the column space of $\boldsymbol{Y}$. We define the set of global minima of the empirical loss, which forms the *minima manifold* $\mathcal{M}$ in the parameter space as

$$\mathcal{M} = \left\{\boldsymbol{w} \middle| \boldsymbol{\alpha} - \boldsymbol{\alpha}^* = \|\boldsymbol{w}\|^\gamma \boldsymbol{w} - \|\boldsymbol{w}^*\|^\gamma \boldsymbol{w}^* \in \boldsymbol{X}^\perp\right\}. \tag{3.2}$$

### 3.2 Motivation

One of the most studied models in the related literature is the *linear diagonal network* first proposed by Gunasekar et al. (2018), *i.e.* the reparametrization scheme

$$\boldsymbol{\alpha}' = \operatorname{diag}(\boldsymbol{w}_L)\operatorname{diag}(\boldsymbol{w}_{L-1})\cdots\operatorname{diag}(\boldsymbol{w}_2)\boldsymbol{w}_1. \tag{3.3}$$

Under this reparametrization, we denote the empirical and population losses by $\hat{\mathcal{L}}'(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$ and $\mathcal{L}'(\boldsymbol{w}; \boldsymbol{w}^*)$ respectively, where $\boldsymbol{w} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_L)$. The simplified version of this model $\boldsymbol{\alpha}' = \boldsymbol{w}^{\odot L}$, where $\odot$ denotes entry-wise multiplication, along with its many variations, have been analyzed in Vaskevicius et al. (2019); Woodworth et al. (2020); Pesme et al. (2021); HaoChen et al. (2021). This model, especially the two-layer version, has been widely adopted as the toy model for its tractability when taking gradients and allowing for explicit and even discrete time analysis (HaoChen et al., 2021). However, as we show in Figure 3b, this model has peculiar training and testing landscapes, which may not capture the real features of the loss landscapes of practical neural networks in Figure 2.

We consider adding $L^2$ regularization beside the linear diagonal network in Equation 3.3 and minimizing the total loss

$$\hat{\mathcal{L}}'(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D}) + \lambda \sum_{i=1}^{L} \|\boldsymbol{w}_i\|^2.$$

For each $i \geq 2$, fixing the rest of the parameters, $\boldsymbol{w}_i$ is encouraged to have identical entries by the regularization term. This does not affect the expressiveness of the model as long as the first layer $\boldsymbol{w}_1$ is not restricted. Moreover, the regularization also encourages $\boldsymbol{w}_i$, $i \in [n]$ to have identical $L^2$-norms and the linear diagonal network is thus reduced to our model in Equation 3.1, with $\gamma = L - 1$.

The additional weight $\|\boldsymbol{w}\|^\gamma$ in our model intuitively warps the parameter space from that of the linear regression model (Figure 3a) so that the loss landscape is no longer quadratic. While the loss landscapes of our model (Figure 3c) preserve the key characteristics of the linear regression model, such as almost quadratic testing loss with an isolated minimum, the level sets of the training loss exhibit a distinctive "shrinkage" as $\|\boldsymbol{w}\| \to \infty$. This is attributed to the nonlinearity introduced by the depth in neural networks.

Compared with the linear diagonal network (Figure 3b), the implicit regularization effect of our model appears more "isotropic". As a result, the loss landscape of our model is closer to that of the practical neural networks (Figure 2). To the best of our knowledge, our model has not been studied by the related literature in neural network theory.

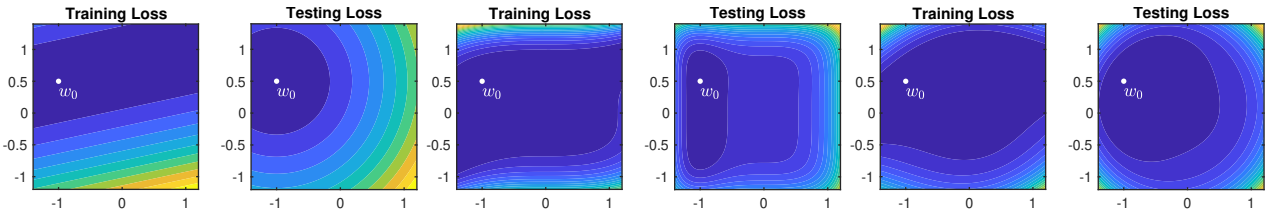
### 3.3 Main Results

In the following, we will analyze the training process of our model using SGD in detail. We first characterize this process into the following three phases and delve into the training behaviors exhibited within each:

I. **Initial phase with a large learning rate $\eta_L$:** In this phase, the actual gradient outweighs the noise in SGD, keeping the trajectory close to that of the gradient flow. The decrease in the training loss comes to saturation, and the trajectory approaches the minima manifold $\mathcal{M}$.

II. **Extended phase maintaining the large learning rate $\eta_L$:** In this phase, the trajectory actively navigates through the minima manifold $\mathcal{M}$ driven by the shape of the training loss landscape, during which the training loss only fluctuates, while the trajectory approaches the neighborhood of the minimum $L^2$-norm solution of the training loss in $\Theta(\eta_L^{-1})$ time.

III. **Final phase with a small learning rate $\eta_S$:** In this concluding phase, the trajectory realigns with the gradient flow, which rapidly penetrates into the minima manifold $\mathcal{M}$ and the final testing performance depends on the timing of the learning rate decay from the preceding phase.




(a) Linear regression model (b) Linear diagonal network (c) Our reparametrization model

Figure 3: Comparison of training and testing loss landscapes of the linear regression model $y = \boldsymbol{w}^\top \boldsymbol{x}$ (Wu et al., 2020a), the linear diagonal network model $y = (\boldsymbol{w}^{\odot 3})^\top \boldsymbol{x}$ (Gunasekar et al., 2018), and our reparametrization model $y = \|\boldsymbol{w}\|^2 \boldsymbol{w}^\top \boldsymbol{x}$. In this example, we choose $d = 2$, $n = 1$, $\boldsymbol{w}^* = (-1, 0.5)^\top$ and $\boldsymbol{X} = (0.15, -0.7)^\top$.

### 3.3.1 Phase I

Following previous works (Li et al., 2017, 2019a, 2021a; Mori et al., 2022), we represnet the initial phase with a large learning rate $\eta_L$ with the following stochastic differential equation (SDE):

$$\mathrm{d}\boldsymbol{w}(t) = -\nabla_{\boldsymbol{w}} \hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D})\mathrm{d}t + \sqrt{\eta_L}\mathrm{d}\boldsymbol{B}(t), \quad (3.4)$$

where $\boldsymbol{B}(t)$ is the standard Brownian motion. The dynamics $\boldsymbol{w}(t)$ reflect the discrete updates of the parameter $\boldsymbol{w}$ starting from an initialization denoted by $\boldsymbol{w}_0$. The time correspondence can be expressed as $\boldsymbol{w}(0) = \boldsymbol{w}_0$ and $\boldsymbol{w}(k\eta_L)$ approximating the parameter value post $k$ iterations. Within this framework, a large learning rate is symbolized by a large diffusion coefficient $\sqrt{\eta_L}$.

One should notice that Equation 3.4 represents the Langevin dynamics featuring a Gibbs-type stationary distribution $\exp(-\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})/\eta_L)$, which is concentrated around the minima manifold $\mathcal{M}$ of the empirical loss. Intuitively, during Phase I, the parameter $\boldsymbol{w}$ remains at a considerable distance from the minima manifold $\mathcal{M}$. Consequently, the dynamics $\boldsymbol{w}(t)$ are primarily driven by the gradient of the training loss until the parameter $\boldsymbol{w}$ approaches the minima manifold $\mathcal{M}$, and the noise within the stochastic gradient surpasses the gradient itself.

Instead of the hitting time analysis in the literature of stochastic gradient Langevin dynamics (Zhang et al., 2017; Chen et al., 2020), which relies on the assumption of the upper bound of $\mathrm{tr}\,\nabla^2 \hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$, we quantify the time span of Phase I through a mixing time analysis:

**Theorem 3.1.** *For any initilization $\boldsymbol{w}(0) = \boldsymbol{w}_0$, under the dynamics in Equation 3.4, the dynamics of the projection of $\boldsymbol{w}(t)$ onto the column space of $\boldsymbol{X}$, denoted by $\boldsymbol{w}_{\boldsymbol{X}}(t)$, have* exponential mixing property, i.e. *for any two initializations $\boldsymbol{w}_0$ and $\boldsymbol{w}_0'$, after time $\mathcal{O}(\log \delta^{-1})$, we have*

$$\|P_0^t(\boldsymbol{w}_0, \cdot) - P_0^t(\boldsymbol{w}_0', \cdot)\|_{TV} \le \delta, \quad (3.5)$$

*where $P_0^t(\boldsymbol{w}_0, \cdot)$ is the distribution of $\boldsymbol{w}_{\boldsymbol{X}}(t)$ starting from $\boldsymbol{w}_0$ at time $t$.*

Intuitively, $\delta$ controls the stability of the dynamics $\boldsymbol{w}_{\boldsymbol{X}}(t)$ in the column space of $\boldsymbol{X}$ (Del Moral and Villemonais, 2018), and thus that of the dynamics $\boldsymbol{w}(t)$. After a mixing time of $\mathcal{O}(\log \delta^{-1})$, $\boldsymbol{w}(t)$ forgets its initialization and is redistributed according to the Gibbs energy of the empirical loss $\exp(-\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})/\eta_L)$. As a result, the dynamics $\boldsymbol{w}(t)$ are close to the minima manifold $\mathcal{M}$ of the training loss towards the end of Phase I.

The proof of Theorem 3.1 is technical and deferred to Appendix A. The main idea is to eliminate the dynamics of $\boldsymbol{w}_{\boldsymbol{X}^\perp}(t)$ by focusing on the SDE of $\boldsymbol{w}_{\boldsymbol{X}}(t)$:

$$\mathrm{d}\boldsymbol{w}_{\boldsymbol{X}}(t) = \boldsymbol{b}(t, \boldsymbol{w}_{\boldsymbol{X}}(t))\mathrm{d}t + \sqrt{\eta_L}\mathrm{d}\boldsymbol{B}_{\boldsymbol{X}}(t), \quad (3.6)$$

where the time-inhomogeneous drift coefficient $\boldsymbol{b}(t, \boldsymbol{w}_{\boldsymbol{X}}(t))$ is given by

$$\boldsymbol{b}(t, \boldsymbol{w}_{\boldsymbol{X}}(t)) = -\mathcal{P}_{\boldsymbol{X}} \nabla_{\boldsymbol{w}} \hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D}).$$

By choosing a Lyapunov function of the form $\exp(\alpha\|\boldsymbol{w}_{\boldsymbol{X}}\|)$, we are able to verify the Foster-Lyapunov condition (Meyn and Tweedie, 1993; Kulik, 2017) and the local Dobrushin contraction condition (Del Moral and Penev, 2017), which are sufficient for the exponential mixing property of the dynamics $\boldsymbol{w}_{\boldsymbol{X}}(t)$.

**Remark 3.2.** *We focus exclusively on the dynamics of $\boldsymbol{w}_{\boldsymbol{X}}(t)$ within the column space of $\boldsymbol{X}$ due to the following considerations: The level sets of the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$ are not compact. Consequently, the dynamics associated with the projection of $\boldsymbol{w}(t)$ onto $\boldsymbol{X}^\perp$, represented as $\boldsymbol{w}_{\boldsymbol{X}^\perp}(t)$, might not exhibit mixing properties. However, in Phase I, our primary anticipation is the stabilization of $\boldsymbol{w}_{\boldsymbol{X}}(t)$. This ensures that $\boldsymbol{w}(t)$ closely aligns with the minima manifold $\mathcal{M}$ of the training loss. A close examination of the dynamics of $\boldsymbol{w}_{\boldsymbol{X}^\perp}(t)$ is reserved for our analysis in Phase II.*

### 3.3.2 Phase II

During Phase II, the parameter trajectory $\boldsymbol{w}(t)$ already resides near the minima manifold $\mathcal{M}$ of the training loss and will be confined to a neighborhood of $\mathcal{M}$ (as depicted in the region between two dashed lines in Figure 4). Due to the large learning rate, rather than converging directly towards the minima manifold $\mathcal{M}$, the trajectory of $\boldsymbol{w}(t)$ will meander across the manifold, steered by the flatness of the training loss landscape.

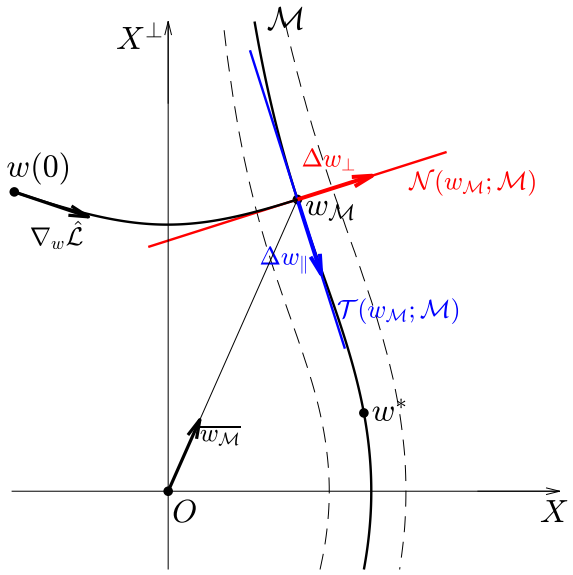



Figure 4: Illustration of the normal space $\mathcal{N}(\boldsymbol{w}_\mathcal{M};\mathcal{M})$ (highlighteed in red) and the tangent space $\mathcal{T}(\boldsymbol{w}_\mathcal{M};\mathcal{M})$ (highlighted in blue) of the manifold $\mathcal{M}$ around a point $\boldsymbol{w}_\mathcal{M}$ on $\mathcal{M}$. The gradient flow trajectory starting from $\boldsymbol{w}(0)$ is represented by a black curve. Due to the stochastic gradient, the dynamics of $\boldsymbol{w}(t)$ do not exactly follow the gradient flow but still enter the neighborhood of $\mathcal{M}$, depicted as the region between the two dashed lines in Phase I. During Phase II, we focus on the effective dynamics of $\boldsymbol{w}(t)$ along the minima manifold $\mathcal{M}$, denoted by $\boldsymbol{w}_\mathcal{M}(t)$. The time scale separation of the dynamics in the normal and tangent spaces allows a quasistatic approach for the analysis of $\boldsymbol{w}_\mathcal{M}(t)$.

Grounded in the insights from Li et al. (2021b); Ma et al. (2022), we sidestep the landscape of the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})$ where $\boldsymbol{w}$ is far from $\mathcal{M}$ and only consider the *effective dynamics* of $\boldsymbol{w}(t)$ along the minima manifold $\mathcal{M}$, denoted by $\boldsymbol{w}_\mathcal{M}(t)$. Specifically, we consider the quadratic expansion of $\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})$ around the minima manifold $\mathcal{M}$ of the training loss, and analyze the dynamics along the manifold by a *quasistatic* analysis: (a) Projecting the dynamics of $\boldsymbol{w}(t)$ onto the tangent space and the normal space of the manifold $\mathcal{M}$, (b) Determining the stationary distribution of the fast projected dynamics $\Delta\boldsymbol{w}_\perp(t)$ in the normal space, and (c) Identifying the effective dynamics in the tangent space by averaging the projected dynamics in the normal space.

**SGD with Label Noise.** In Phase II, we consider an SDE of the general form:

$$\mathrm{d}\boldsymbol{w}(t) = -\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D})\mathrm{d}t + \sqrt{\eta_L}\boldsymbol{\sigma}(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t), \tag{3.7}$$

where $\boldsymbol{\sigma}(\cdot) \in \mathbb{R}^{d\times d}$ is the diffusion tensor, which we will specify later. Fix the time $t$ and a point $\boldsymbol{w}_\mathcal{M}$ on $\mathcal{M}$ near $\boldsymbol{w}(t)$, we decompose $\boldsymbol{w}(t) - \boldsymbol{w}_\mathcal{M}$ into the tangent component $\Delta\boldsymbol{w}_\parallel(t)$ along the tangent space $\mathcal{T}(\boldsymbol{w}_\mathcal{M};\mathcal{M})$ and the normal component $\Delta\boldsymbol{w}_\perp(t)$ with respect to the normal space $\mathcal{N}(\boldsymbol{w}_\mathcal{M};\mathcal{M})$, *i.e.*

$$\boldsymbol{w}(t) - \boldsymbol{w}_\mathcal{M} = \Delta\boldsymbol{w}_\parallel(t) + \Delta\boldsymbol{w}_\perp(t). \tag{3.8}$$

By Proposition B.1 and $\boldsymbol{A}_\mathcal{M}$ defined therein, we are able to perform the projections with closed forms and rewrite Equation 3.7 as the following system of SDE:

$$\begin{cases} \mathrm{d}\Delta\boldsymbol{w}_\parallel(t) = -\mathcal{P}_{\boldsymbol{A}_\mathcal{M}^{-1}\boldsymbol{X}^\perp}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D})\mathrm{d}t \\ \qquad\qquad + \sqrt{\eta_L}\boldsymbol{\sigma}_\parallel(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t), \\ \mathrm{d}\Delta\boldsymbol{w}_\perp(t) = -\mathcal{P}_{\boldsymbol{A}_\mathcal{M}\boldsymbol{X}}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D})\mathrm{d}t \\ \qquad\qquad + \sqrt{\eta_L}\boldsymbol{\sigma}_\perp(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t), \end{cases} \tag{3.9}$$

where $\boldsymbol{\sigma}_\parallel(\cdot) = \mathcal{P}_{\boldsymbol{A}_\mathcal{M}^{-1}\boldsymbol{X}^\perp}\boldsymbol{\sigma}(\cdot)$ and $\boldsymbol{\sigma}_\perp(\cdot) = \mathcal{P}_{\boldsymbol{A}_\mathcal{M}\boldsymbol{X}}\boldsymbol{\sigma}(\cdot)$ are the diffusion tensors in the tangent and normal space respectively.

As studied in Blanc et al. (2020); Damian et al. (2021), the scenario where the labels $\{y_i\}_{i=1}^n$ of the dataset $\mathcal{D}$ contains binary noises, *i.e.* $y_i = (\boldsymbol{\alpha}^*)^\top\boldsymbol{x}_i + \xi_i$ where $\xi_i \sim \mathrm{Unif}\{-\sigma,\sigma\}$, corresponds to the diffusion tensor $\boldsymbol{\sigma}$ satisfying the following assumption:

$$\boldsymbol{\sigma}(\boldsymbol{w}(t))\boldsymbol{\sigma}(\boldsymbol{w}(t))^\top := \sigma^2\nabla^2\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D}), \tag{3.10}$$

causing $\boldsymbol{\sigma}_\parallel(\boldsymbol{w}(t))$ to vanish, for $\boldsymbol{w}(t)$ near $\mathcal{M}$.

**Stationary Distribution along the Normal Space.** Expand the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})$ around $\boldsymbol{w}_\mathcal{M}$ (*cf.* Equation B.3), we have

$$\begin{aligned} &\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}_\mathcal{M} + \Delta\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D}) \\ =&\frac{1}{n}\boldsymbol{A}_\mathcal{M}\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + o(\|\Delta\boldsymbol{w}\|), \end{aligned} \tag{3.11}$$

from which we see that the first order term of $\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})$ is in the normal space and thus the drift term of the dynamics of $\Delta\boldsymbol{w}_\perp(t)$ is of the first order of $\Delta\boldsymbol{w}$, while the drift term of the dynamics of $\Delta\boldsymbol{w}_\parallel(t)$ is at most of the second order of $\Delta\boldsymbol{w}$. This indicates that the dynamics of $\Delta\boldsymbol{w}_\parallel(t)$ are much slower than that of

$\Delta \boldsymbol{w}_\perp(t)$. Therefore, we consider rescaling the dynamics of $\Delta \boldsymbol{w}_\perp(t)$ by $\Delta \boldsymbol{w}_\perp(t+\tau)$ from time $t$ with a smaller time scale $\tau$. In fact, as we will see in Lemma 3.3, the scale of $t$ is $\Theta(\eta_L^{-1})$ greater than that of $\tau$.

By choosing the the parametrization $\Delta \boldsymbol{w}_\perp(t+\tau) = \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{\epsilon}(\tau)$, where $\boldsymbol{\epsilon}(\tau) \in \mathbb{R}^n$, and expanding around $\boldsymbol{w}_{\mathcal{M}}$, one can show that the rescaled dynamics of $\Delta \boldsymbol{w}_\perp(t+\tau)$ under the assumption (3.10) is simplified to the following Ornstein-Uhlenbeck (OU) process

$$\mathrm{d}\boldsymbol{\epsilon}(\tau) = -\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \boldsymbol{\epsilon}(\tau) \mathrm{d}\tau + \sqrt{\frac{\sigma^2 \eta_L}{n}} \mathrm{d}\boldsymbol{B}_\perp(\tau), \tag{3.12}$$

where $\boldsymbol{B}_\perp$ denotes an $n$-dimensional Brownian motion. Notice that the above OU process has exponentially mixing property, whereby we approximate the dynamics of $\Delta \boldsymbol{w}_\perp(t)$ at the large time scale as

$$\Delta \boldsymbol{w}_\perp(t) \sim \mathcal{N}\left( \boldsymbol{0}, \frac{\sigma^2 \eta_L}{2} \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \right), \tag{3.13}$$

*i.e.* assuming $\Delta \boldsymbol{w}_\perp(t)$ is at the stationary distribution of the OU process Equation 3.12 at time $t$.

**Effective Dynamics along the Tangent Space.** We then consider the effective dynamics of $\Delta \boldsymbol{w}_\parallel(t)$ along the tangent space $\mathcal{T}(\boldsymbol{w}_{\mathcal{M}}; \mathcal{M})$ by taking expectation of the stationary distribution (3.13) of the rescaled dynamics $\Delta \boldsymbol{w}_\perp(t+\tau)$ in the normal space $\mathcal{N}(\boldsymbol{w}_{\mathcal{M}}; \mathcal{M})$ for each time $t$, *i.e.*

$$\frac{\mathrm{d}\Delta \boldsymbol{w}_\parallel(t)}{\mathrm{d}t} = -\mathbb{E}_{\Delta \boldsymbol{w}_\perp(t)} \left[ \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^\perp} \nabla_{\boldsymbol{w}} \hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D}) \right]. \tag{3.14}$$

Summarizing the above analysis, we have the following characterization of the dynamics of $\boldsymbol{w}(t)$ in Phase II:

**Lemma 3.3.** *Under the dynamics in Equation 3.9 with the label noise assumption (Blanc et al., 2020; Damian et al., 2021), the effective dynamics of $\boldsymbol{w}(t)$ in Phase II are given by*

$$\frac{\mathrm{d}\boldsymbol{w}_{\mathcal{M}}(t)}{\mathrm{d}t} = -A \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1} C(\boldsymbol{w}_{\mathcal{M}}(t)) \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^\perp} \overline{\boldsymbol{w}_{\mathcal{M}}(t)}, \tag{3.15}$$

*where $A = \eta_L \sigma^2 \gamma / 2$ and*

$$C(\boldsymbol{w}_{\mathcal{M}}) = \frac{1}{n} \left( \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) - (\gamma+2) \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \boldsymbol{X} \boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}} \right). \tag{3.16}$$

The $\eta_L$ factor in $A$ reflects the time scale separation between the dynamics of $\boldsymbol{w}(t)$ and $\boldsymbol{w}_{\mathcal{M}}(t)$ as we claimed before. The effective dynamics (3.15) have a clear geometric interpretation: as shown in Figure 4, whenever $C(\boldsymbol{w}_{\mathcal{M}}(t)) \geq 0$, we have $\dot{\boldsymbol{w}}_{\mathcal{M}} \propto -\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^\perp} \overline{\boldsymbol{w}_{\mathcal{M}}(t)}$ pointing towards the column space of $\boldsymbol{X}$, which leads to the following result:

**Theorem 3.4.** *For any $\gamma \geq 0$, the effective dynamics (3.15) converge to the minimum $L^2$-norm solution of the population loss $\mathcal{L}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$, denoted by $\boldsymbol{w}^\dagger$, with probability $1 - \exp(-\Omega(nd))$. The norm of $\|\boldsymbol{w}_{\mathcal{M}}(t)\|$ satisfies:*

$$\frac{\mathrm{d}\|\boldsymbol{w}_{\mathcal{M}}(t)\|}{\mathrm{d}t} \leq -B \left( \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1} - \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-3} \|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2 \right), \tag{3.17}$$

*where $B = \frac{\sigma^2 \eta_L \gamma \underline{C} d}{2(1+\gamma)^2}$, $\underline{C} = \inf_{t \geq 0} C(\boldsymbol{w}_{\mathcal{M}}(t))$, and $\boldsymbol{\alpha}_{\boldsymbol{X}}^* = \mathcal{P}_{\boldsymbol{X}} \boldsymbol{\alpha}^*$. Moreover, for any $\gamma > 1/2$, the convergence in Equation 3.17 is exponentially fast,* i.e. *after time $\mathcal{O}\left( \frac{\log \delta^{-1}}{\sigma^2 \eta_L \underline{C} d} \right)$, we have $\|\boldsymbol{w}_{\mathcal{M}}(t)\| - \|\boldsymbol{w}^\dagger\| \leq \delta$.*

**Remark 3.5.** *We would like to make the following remarks regarding the above theorem:*

- *The high probability argument is due to the randomness in the data generation of $\boldsymbol{X}$. One should notice that $C(\boldsymbol{w}_{\mathcal{M}}) \geq 0$ holds if the numerical rank of the data matrix $r(\boldsymbol{X}) := \|\boldsymbol{X}\|_F / \|\boldsymbol{X}\| \geq \sqrt{\gamma+2}$. In high-dimensional spaces ($d \to \infty$), $r(\boldsymbol{X}) \gtrsim \sqrt{n}$, and thus $C(\boldsymbol{w}_{\mathcal{M}})$ can be lower bounded with high probability (*cf. *Lemma B.6).*
- *Mirroring the $L^2$-regularization or the ridge regression helps avoid overfitting in practice, the minimum $L^2$-norm solution $\boldsymbol{w}^\dagger$ of overparametrized models is both intuitively and provably the near-optimal solution under certain assumptions (Wu and Xu, 2020; Bartlett et al., 2020). Notably, our result here aligns with those in the literature of overparametrized ridgeless regrssion (Liang and Rakhlin, 2020; Hastie et al., 2022).*
- *As observed from Equation 3.17, the condition $\gamma \geq 0$ can be relaxed to $\gamma > -1$, by which we still have $\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1} \geq \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-3} \|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2$ and thus the convergence of the effective dynamics. This makes our model less restrictive compared with linear diagonal networks (Gunasekar et al., 2018; Woodworth et al., 2020). We would also like to point out that $\gamma = -1$ refers to the weight normalization (Wu et al., 2020b; Chou et al., 2023), where the convergence to the minimum $L^2$-norm solution is also obtained. Still, an additional parameter is needed to control the length of the parameter $\boldsymbol{\alpha}$.*

Further explanations of the above analysis and the proof of the above theorem are deferred to Appendix B.

### 3.3.3 Phase III

Suppose the trajectory of $\boldsymbol{w}(t)$ reaches the neighborhood of a point $\boldsymbol{w}_{\mathcal{M}}$ on the minima manifold $\mathcal{M}$ of the training loss. We then perform a local analysis

around $\boldsymbol{w}_{\mathcal{M}}$. When the step size $\eta_S$ is sufficiently small, the trajectory of performing SGD on the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$ can be approximated by the gradient flow (Smith et al., 2021), *i.e.*

$$\mathrm{d}\boldsymbol{w}(t) = -\nabla_{\boldsymbol{w}} \hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D})\mathrm{d}t. \tag{3.18}$$

Again, by approximating the local geometry by the quadratic expansion of $\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$ around $\boldsymbol{w}_{\mathcal{M}}$, we have the following exponential convergence:

**Theorem 3.6.** *Under the dynamics in Equation 3.18, the parameter $\boldsymbol{w}(t)$ converges to a nearby $\boldsymbol{w}_{\mathcal{M}} \in \mathcal{M}$ exponentially fast,* i.e. *after time $\mathcal{O}(\log \delta^{-1})$, we have $\|\boldsymbol{w}(t) - \boldsymbol{w}_{\mathcal{M}}\| \leq \delta$.*

The proof of this theorem is standard and will be deferred to Appendix C.

Recall that by Theorem 3.4, if Phase II is executed over a sufficiently long period, the effective dynamics $\boldsymbol{w}_{\mathcal{M}}(t)$ converge to the minimum $L^2$-norm solution $\boldsymbol{w}^\dagger$, which means the original dynamics $\boldsymbol{w}(t)$ approach the neighborhood of $\boldsymbol{w}^\dagger$, and then Phase III is able to recover the minimum $L^2$-norm solution $\boldsymbol{w}^\dagger$ of the training loss. In general, the final testing performance thus depends on the timing of the learning rate decay from Phase II to Phase III. This corroborates the empirically observed benefits of late learning rate decay.

# 4 DISCUSSIONS

In this paper, we delve into understanding the question of why late learning rate decay leads to better generalization. Our study is motivated by experimental observations and the visualization of the training and testing loss landscapes on an image classification task. We subsequently introduce an overparametrized model with a novel nonlinear reparametrization, which presents a "shrinking" training loss landscape as the $L^2$-norm of the parameter increases.

Upon establishing this model, we characterize the training process into three distinct phases. Our analysis emphasizes that during Phase II, which corresponds to the extended period before learning rate decay, the parameter approaches the minimum $L^2$-norm solution. We believe our model and results shed light on the training process of neural networks and provide a new perspective on the generalization of deep learning architectures. One of the limitations of our work is that our analysis is predominantly based on a continuous-time approach, and the discrete-time analysis for our model is left for future work.

**Acknowledgments**

We thank the anonymous reviewers for their helpful comments.

**References**

Agarwal, N., Goel, S., and Zhang, C. (2021). Acceleration via fractal learning rate schedules. In *International Conference on Machine Learning*, pages 87–99. PMLR.

Allen-Zhu, Z. and Li, Y. (2019). Can sgd learn recurrent neural networks with provable generalization? *Advances in Neural Information Processing Systems*, 32.

Andriushchenko, M., D'Angelo, F., Varre, A., and Flammarion, N. (2023). Why do we need weight decay in modern deep learning? *arXiv preprint arXiv:2310.04415*.

Andriushchenko, M., Varre, A., Pillaud-Vivien, L., and Flammarion, N. (2022). Sgd with large step sizes learns sparse features. *arXiv preprint arXiv:2210.05337*.

Arora, S., Li, Z., and Panigrahi, A. (2022). Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Beugnot, G., Mairal, J., and Rudi, A. (2022). On the benefits of large learning rates for kernel methods. In *Conference on Learning Theory*, pages 254–282. PMLR.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. (2020). Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR.

Chen, L. and Bruna, J. (2022). On gradient descent convergence beyond the edge of stability. *arXiv preprint arXiv:2206.04172*.

Chen, X., Du, S. S., and Tong, X. T. (2020). On stationary-point hitting time and ergodicity of stochastic gradient langevin dynamics. *Journal of Machine Learning Research*.

Chou, H.-H., Rauhut, H., and Ward, R. (2023). Robust implicit regularization via weight normalization. *arXiv preprint arXiv:2305.05448*.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. (2021). Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*.

Cooper, Y. (2018). The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*.

Cooper, Y. (2020). The critical locus of overparameterized neural networks. *arXiv preprint arXiv:2005.04210*.

Damian, A., Ma, T., and Lee, J. D. (2021). Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461.

Damian, A., Nichani, E., and Lee, J. D. (2022). Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*.

Del Moral, P. and Penev, S. (2017). *Stochastic Processes: From Applications to Theory*. CRC Press.

Del Moral, P. and Villemonais, D. (2018). Exponential mixing properties for time inhomogeneous diffusion processes with killing. *Bernoulli*.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Even, M., Pesme, S., Gunasekar, S., and Flammarion, N. (2023). (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *arXiv preprint arXiv:2302.08982*.

Grimmer, B. (2023). Provably faster gradient descent via long steps. *arXiv preprint arXiv:2307.06324*.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31.

HaoChen, J. Z., Wei, C., Lee, J., and Ma, T. (2021). Shape matters: Understanding the implicit bias of the noise covariance. In *Conference on Learning Theory*, pages 2315–2357. PMLR.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949.

Hochreiter, S. and Schmidhuber, J. (1997). Flat minima. *Neural computation*, 9(1):1–42.

Hoffer, E., Hubara, I., and Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. (2018). Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 392–402. Springer.

Jelassi, S. and Li, Y. (2022). Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pages 9965–10040. PMLR.

Katzenberger, G. S. (1990). *Solutions of a stochastic differential equation forced onto a manifold by a large drift*. The University of Wisconsin-Madison.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kleinberg, B., Li, Y., and Yuan, Y. (2018). An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pages 2698–2707. PMLR.

Kulik, A. (2017). *Ergodic Behavior of Markov Processes: With Applications to Limit Theorems*, volume 67. Walter de Gruyter GmbH & Co KG.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

Li, Q., Tai, C., and Weinan, E. (2017). Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR.

Li, Q., Tai, C., and Weinan, E. (2019a). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520.

Li, Y., Wei, C., and Ma, T. (2019b). Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32.

Li, Z., Lyu, K., and Arora, S. (2020). Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33:14544–14555.

Li, Z., Malladi, S., and Arora, S. (2021a). On the validity of modeling sgd with stochastic differential equations (sdes). *Advances in Neural Information Processing Systems*, 34:12712–12725.

Li, Z., Wang, T., and Arora, S. (2021b). What happens after sgd reaches zero loss?–a mathematical framework. *arXiv preprint arXiv:2110.06914*.

Li, Z., Wang, T., and Yu, D. (2022). Fast mixing of stochastic gradient descent with normalization and weight decay. *Advances in Neural Information Processing Systems*, 35:9233–9248.

Liang, T. and Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*.

Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., and Williams, M. (2022). Towards understanding grokking: An effective theory of representation learning. *Advances in Neural Information Processing Systems*, 35:34651–34663.

Lorch, E. (2016). Visualizing deep network training trajectories with pca. In *ICML Workshop on Visualization for Deep Learning*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lyu, K. and Li, J. (2019). Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*.

Ma, C., Kunin, D., Wu, L., and Ying, L. (2022). Beyond the quadratic approximation: the multiscale structure of neural network loss landscapes. *arXiv preprint arXiv:2204.11326*.

Mandt, S., Hoffman, M., and Blei, D. (2016). A variational analysis of stochastic gradient algorithms. In *International conference on machine learning*, pages 354–363. PMLR.

Meyn, S. P. and Tweedie, R. L. (1993). Stability of markovian processes iii: Foster–lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, 25(3):518–548.

Mori, T., Ziyin, L., Liu, K., and Ueda, M. (2022). Power-law escape rate of sgd. In *International Conference on Machine Learning*, pages 15959–15975. PMLR.

Mulayoff, R., Michaeli, T., and Soudry, D. (2021). The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34:17749–17761.

Nacson, M. S., Ravichandran, K., Srebro, N., and Soudry, D. (2022). Implicit bias of the step size in linear diagonal neural networks. In *International Conference on Machine Learning*, pages 16270–16295. PMLR.

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. (2021). Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR.

Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE.

Smith, S. L., Dherin, B., Barrett, D. G., and De, S. (2021). On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. (2017). Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*.

Smith, S. L. and Le, Q. V. (2017). A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*.

Vaskevicius, T., Kanade, V., and Rebeschini, P. (2019). Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, Y., Chen, M., Zhao, T., and Tao, M. (2021). Large learning rate tames homogeneity: Convergence and balancing effect. *arXiv preprint arXiv:2110.03677*.

Weinan, E., Li, T., and Vanden-Eijnden, E. (2021). *Applied stochastic analysis*, volume 199. American Mathematical Soc.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR.

Wu, D. and Xu, J. (2020). On the optimal weighted $\ell_2$ regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123.

Wu, J., Zou, D., Braverman, V., and Gu, Q. (2020a). Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. *arXiv preprint arXiv:2011.02538*.

Wu, L., Ma, C., et al. (2018). How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31.

Wu, L., Wang, M., and Su, W. (2022). The alignment property of sgd noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35:4680–4693.

Wu, X., Dobriban, E., Ren, T., Wu, S., Li, Z., Gunasekar, S., Ward, R., and Liu, Q. (2020b). Implicit regularization and convergence for weight normalization. *Advances in Neural Information Processing Systems*, 33:2835–2847.

Zhang, Y., Liang, P., and Charikar, M. (2017). A hitting time analysis of stochastic gradient langevin dynamics. In *Conference on Learning Theory*, pages 1980–2022. PMLR.

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. (2020). Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296.

Zhu, X., Wang, Z., Wang, X., Zhou, M., and Ge, R. (2022). Understanding edge-of-stability training dynamics with a minimalist example. *arXiv preprint arXiv:2210.03294*.

Žunkovič, B. and Ilievski, E. (2022). Grokking phase transitions in learning local rules with gradient descent. *arXiv preprint arXiv:2210.15435*.

# A MISSING PROOFS FOR PHASE I

In this section, we present the missing proofs of the results in Phase I in Section 3.3.1 of the main text.

Before we present the proofs, we first perform some common computations regarding the gradient of the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})$ and the population loss $\mathcal{L}(\boldsymbol{w};\boldsymbol{w}^*)$ for later reference:

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D}) &= \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*),\\ \nabla_{\boldsymbol{w}}\boldsymbol{\alpha} &= \|\boldsymbol{w}\|^\gamma\boldsymbol{I}+\gamma\|\boldsymbol{w}\|^{\gamma-1}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\boldsymbol{w}^\top = \|\boldsymbol{w}\|^\gamma\left(\boldsymbol{I}+\gamma\overline{\boldsymbol{w}}\;\overline{\boldsymbol{w}}^\top\right) := \boldsymbol{A}(\boldsymbol{w}),\\ \nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D}) &= (\nabla_{\boldsymbol{w}}\boldsymbol{\alpha})\left(\nabla_{\boldsymbol{\alpha}}\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})\right) = \frac{1}{n}\boldsymbol{A}(\boldsymbol{w})\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*),\end{aligned} \tag{A.1}$$

where the gradients are assumed to be column vectors and $\overline{\boldsymbol{w}} = \boldsymbol{w}/\|\boldsymbol{w}\|$ is the normalized parameter.

For readers' convenience, we restate the SDE in Equation 3.4 here:

$$\mathrm{d}\boldsymbol{w}(t) = -\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D})\mathrm{d}t + \sqrt{\eta_L}\mathrm{d}\boldsymbol{B}(t).$$

Let $\boldsymbol{w}_{\boldsymbol{X}}(t)$ be the projection of $\boldsymbol{w}(t)$ onto the column space of $\boldsymbol{X}$, and $\boldsymbol{w}_{\boldsymbol{X}^\perp}(t)$ be the projection of $\boldsymbol{w}(t)$ onto the orthogonal complement of the column space of $\boldsymbol{X}$. Then, the dynamics (3.4) can be decomposed into the following two SDEs:

$$\begin{cases}\mathrm{d}\boldsymbol{w}_{\boldsymbol{X}}(t) &= -\mathcal{P}_{\boldsymbol{X}}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D})\mathrm{d}t + \sqrt{\eta_L}\mathrm{d}\boldsymbol{B}_{\boldsymbol{X}}(t),\\ \mathrm{d}\boldsymbol{w}_{\boldsymbol{X}^\perp}(t) &= -\mathcal{P}_{\boldsymbol{X}^\perp}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D})\mathrm{d}t + \sqrt{\eta_L}\mathrm{d}\boldsymbol{B}_{\boldsymbol{X}^\perp}(t),\end{cases} \tag{A.2}$$

where $\boldsymbol{B}_{\boldsymbol{X}}(t)$ and $\boldsymbol{B}_{\boldsymbol{X}^\perp}(t)$ are the Brownian motions in the column space of $\boldsymbol{X}$ and its orthogonal complement $\boldsymbol{X}^\perp$, respectively.

One should notice that the two SDEs in (A.2) are coupled. Now that we are only interested in the dynamics of the projection $\boldsymbol{w}_{\boldsymbol{X}}(t)$, we can eliminate the dynamics of $\boldsymbol{w}_{\boldsymbol{X}^\perp}(t)$ by assuming a more general drift term in the SDE of $\boldsymbol{w}_{\boldsymbol{X}}(t)$ as presented in Equation 3.6:

$$\mathrm{d}\boldsymbol{w}_{\boldsymbol{X}}(t) = \boldsymbol{b}(t,\boldsymbol{w}_{\boldsymbol{X}}(t))\mathrm{d}t + \sqrt{\eta_L}\mathrm{d}\boldsymbol{B}_{\boldsymbol{X}}(t),$$

where the drift coefficient $\boldsymbol{b}(t,\boldsymbol{w}_{\boldsymbol{X}}(t))$ is a function of both time $t$ and the projection $\boldsymbol{w}_{\boldsymbol{X}}(t)$ such that

$$\boldsymbol{b}(t,\boldsymbol{w}_{\boldsymbol{X}}(t)) = -\mathcal{P}_{\boldsymbol{X}}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D}).$$

We denote the infinitestimal generator of the SDE (3.6) by $\mathcal{A}_s$, *i.e.*

$$\mathcal{A}_s = \boldsymbol{b}(s,\boldsymbol{w}_{\boldsymbol{X}}(s))\cdot\nabla + \frac{\eta_L}{2}\operatorname{tr}\nabla^2,$$

where the subscript $s$ indicates the dependence of the infinitesimal generator on time $s$.

In the following, we adapt the Foster-Lyapunov criteria (Meyn and Tweedie, 1993; Kulik, 2017) to analyze the mixing time of the dynamics of $\boldsymbol{w}_{\boldsymbol{X}}(t)$. Should no confusion arise, we will use $\boldsymbol{w}_{\boldsymbol{X}} \in \mathbb{R}^n$ to denote the projection of an arbitrary point $\boldsymbol{w} \in \mathbb{R}^d$ onto the column space of $\boldsymbol{X}$.

Define the Lyapunov function as

$$W(\boldsymbol{w}_{\boldsymbol{X}}) = \exp(\alpha\|\boldsymbol{w}_{\boldsymbol{X}}\|), \tag{A.3}$$

where $\alpha$ is a positive constant, we have the following lemma:

**Lemma A.1.** *For any time $s$, the infinitestimal generator $\mathcal{A}_s$ of the dynamics of $\boldsymbol{w}_{\boldsymbol{X}}(t)$ satisfies the following drift condition:*

$$\mathcal{A}_s W(\boldsymbol{w}_{\boldsymbol{X}}) \le -C_1 W(\boldsymbol{w}_{\boldsymbol{X}}) + C_2,$$

*where $C_1$ and $C_2$ are positive constants.*

*Proof.* Since the drift and diffusion tensors $\boldsymbol{b}(s, \boldsymbol{w}_{\boldsymbol{X}}(s))$ and $\eta_L$ are locally bounded and $V \in C^\infty(\mathbb{R}^n)$, we only have to prove

$$\limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} \mathcal{A}_s W(\boldsymbol{w}_{\boldsymbol{X}}) \le -C_1,$$

for some positive constant $C_1$, which is equivalent to

$$\limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} \frac{\mathcal{A}_s W(\boldsymbol{w}_{\boldsymbol{X}})}{W(\boldsymbol{w}_{\boldsymbol{X}})} < 0.$$

To compute $\mathcal{A}_s W(\boldsymbol{w}_{\boldsymbol{X}})$, we first calculate the gradient and the Hessian of $W(\boldsymbol{w}_{\boldsymbol{X}})$:

$$\begin{aligned}
\nabla W(\boldsymbol{w}_{\boldsymbol{X}}) &= \alpha W(\boldsymbol{w}_{\boldsymbol{X}}) \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \\
\nabla^2 W(\boldsymbol{w}_{\boldsymbol{X}}) &= \alpha^2 W(\boldsymbol{w}_{\boldsymbol{X}}) \overline{\boldsymbol{w}_{\boldsymbol{X}}}\ \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top + \alpha W(\boldsymbol{w}_{\boldsymbol{X}}) \frac{\boldsymbol{I} - \overline{\boldsymbol{w}_{\boldsymbol{X}}}\ \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top}{\|\boldsymbol{w}_{\boldsymbol{X}}\|},
\end{aligned}$$

where $\overline{\boldsymbol{w}_{\boldsymbol{X}}} = \boldsymbol{w}_{\boldsymbol{X}}/\|\boldsymbol{w}_{\boldsymbol{X}}\|$. Then we have

$$\begin{aligned}
\mathcal{A}_s W(\boldsymbol{w}_{\boldsymbol{X}}) &= \alpha W(\boldsymbol{w}_{\boldsymbol{X}}) \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{b}(s, \boldsymbol{w}_{\boldsymbol{X}}) \rangle + \frac{\eta_L}{2} \operatorname{tr}\left( \alpha^2 W(\boldsymbol{w}_{\boldsymbol{X}}) \overline{\boldsymbol{w}_{\boldsymbol{X}}}\ \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top + \alpha W(\boldsymbol{w}_{\boldsymbol{X}}) \frac{\boldsymbol{I} - \overline{\boldsymbol{w}_{\boldsymbol{X}}}\ \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top}{\|\boldsymbol{w}_{\boldsymbol{X}}\|} \right) \\
&= W(\boldsymbol{w}_{\boldsymbol{X}}) \left[ \alpha \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{b}(s, \boldsymbol{w}_{\boldsymbol{X}}) \rangle + \frac{\eta_L}{2} \left( \alpha^2 + \alpha \frac{n-1}{\|\boldsymbol{w}_{\boldsymbol{X}}\|} \right) \right],
\end{aligned} \tag{A.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Plugging (A.1) into (A.4), we have

$$\begin{aligned}
\langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{b}(s, \boldsymbol{w}_{\boldsymbol{X}}) \rangle &= -\frac{1}{n} \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \mathcal{P}_{\boldsymbol{X}} \nabla_{\boldsymbol{w}} \hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D}) \rangle \\
&= -\frac{1}{n} \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{A}(\boldsymbol{w}) \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle \\
&= -\frac{\|\boldsymbol{w}\|^\gamma}{n} \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle - \frac{\|\boldsymbol{w}\|^\gamma}{n} \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \overline{\boldsymbol{w}}\ \overline{\boldsymbol{w}}^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle \\
&= -\frac{\|\boldsymbol{w}\|^\gamma}{n} \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \rangle - \frac{\|\boldsymbol{w}\|^\gamma}{n} \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top \overline{\boldsymbol{w}}\ \overline{\boldsymbol{w}}^\top \boldsymbol{X}\boldsymbol{X}^\top (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*),
\end{aligned}$$

where the second equality is due to $\overline{\boldsymbol{w}_{\boldsymbol{X}}} \in \boldsymbol{X}$.

Notice that $\boldsymbol{\alpha} = \|\boldsymbol{w}\|^\gamma \boldsymbol{w}$ by definition, we have

$$\begin{aligned}
&\limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} \frac{\langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{b}(s, \boldsymbol{w}_{\boldsymbol{X}}) \rangle}{\|\boldsymbol{w}\|^{2\gamma} \|\boldsymbol{w}_{\boldsymbol{X}}\|} \\
&= \limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} -\frac{1}{n} \left\langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{X}\boldsymbol{X}^\top \left( \overline{\boldsymbol{w}_{\boldsymbol{X}}} - \frac{\boldsymbol{\alpha}^*}{\|\boldsymbol{w}\|^\gamma \|\boldsymbol{w}_{\boldsymbol{X}}\|} \right) \right\rangle - \frac{1}{n} \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top \overline{\boldsymbol{w}}\ \overline{\boldsymbol{w}}^\top \boldsymbol{X}\boldsymbol{X}^\top \left( \overline{\boldsymbol{w}} \frac{\|\boldsymbol{w}\|}{\|\boldsymbol{w}_{\boldsymbol{X}}\|} - \frac{\boldsymbol{\alpha}^*}{\|\boldsymbol{w}\|^\gamma \|\boldsymbol{w}_{\boldsymbol{X}}\|} \right) \\
&= \limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} -\frac{1}{n} \left\langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{X}\boldsymbol{X}^\top \overline{\boldsymbol{w}_{\boldsymbol{X}}} \right\rangle - \frac{1}{n} \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top \overline{\boldsymbol{w}}\ \overline{\boldsymbol{w}}^\top \boldsymbol{X}\boldsymbol{X}^\top \overline{\boldsymbol{w}} \frac{\|\boldsymbol{w}\|}{\|\boldsymbol{w}_{\boldsymbol{X}}\|} \\
&\le \limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} -\frac{1}{n} \overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top \boldsymbol{X}\boldsymbol{X}^\top \overline{\boldsymbol{w}_{\boldsymbol{X}}} < 0,
\end{aligned} \tag{A.5}$$

where the first equality is due to $\boldsymbol{X}^\top \boldsymbol{w} = \boldsymbol{X}^\top \boldsymbol{w}_{\boldsymbol{X}}$, the second equality is because $\|\boldsymbol{w}\| \ge \|\boldsymbol{w}_{\boldsymbol{X}}\|$ and hence $\|\boldsymbol{w}_{\boldsymbol{X}}\| \to \infty$ implies $\|\boldsymbol{w}\| \to \infty$, the second to last inequality is due to $\overline{\boldsymbol{w}_{\boldsymbol{X}}}^\top \overline{\boldsymbol{w}} \ge 0$ and $\overline{\boldsymbol{w}}^\top \boldsymbol{X}\boldsymbol{X}^\top \overline{\boldsymbol{w}}$, and the last inequality is due to $\overline{\boldsymbol{w}_{\boldsymbol{X}}} \in \boldsymbol{X}$ and $\boldsymbol{X}$ is full rank.

Suppose $\limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} \frac{\langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{b}(s, \boldsymbol{w}_{\boldsymbol{X}}) \rangle}{\|\boldsymbol{w}\|^{2\gamma} \|\boldsymbol{w}_{\boldsymbol{X}}\|} \le -C_3$ for some positive constant $C_3$, we have

$$\begin{aligned}
\limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} \frac{\mathcal{A}_s W(\boldsymbol{w}_{\boldsymbol{X}})}{W(\boldsymbol{w}_{\boldsymbol{X}})} &= \limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} \alpha \langle \overline{\boldsymbol{w}_{\boldsymbol{X}}}, \boldsymbol{b}(s, \boldsymbol{w}_{\boldsymbol{X}}) \rangle + \frac{\eta_L}{2} \left( \alpha^2 + \alpha \frac{n-1}{\|\boldsymbol{w}_{\boldsymbol{X}}\|} \right), \\
&\le -\limsup_{\|\boldsymbol{w}_{\boldsymbol{X}}\|\to\infty} \alpha C_3 \|\boldsymbol{w}\|^{2\gamma} \|\boldsymbol{w}_{\boldsymbol{X}}\| + \frac{\eta_L}{2} \left( \alpha^2 + \alpha \frac{n-1}{\|\boldsymbol{w}_{\boldsymbol{X}}\|} \right) < 0,
\end{aligned}$$

which proves our statement. □

In the following, we will use the notion of *Markov transitions*, referring to a transition kernel $T(\cdot, \mathrm{d}\boldsymbol{w}_{\boldsymbol{X}})$ on $\mathbb{R}^n$. We also define the following two integral operations induced by this transition kernel

$$Tf(\cdot) := \int_{\mathbb{R}^n} f(\boldsymbol{w}_{\boldsymbol{X}}) T(\cdot, \mathrm{d}\boldsymbol{w}_{\boldsymbol{X}}) \quad \text{and} \quad \mu T(\cdot) = \int_{\mathbb{R}^n} T(\boldsymbol{w}_{\boldsymbol{X}}, \cdot) \mathrm{d}\mu(\boldsymbol{w}_{\boldsymbol{X}}),$$

where $f$ is a bounded measurable function on $\mathbb{R}^n$ and $\mu$ is a probability measure on $\mathbb{R}^n$. As a special family of Markov transitions, we define the time-inhomogeneous Markov transition semigroup $P_s^{s+\tau}$ for $\tau \geq 0$ as:

$$P_s^{s+\tau} f(\boldsymbol{w}_{\boldsymbol{X}}) = \mathbb{E}[f(\boldsymbol{w}_{\boldsymbol{X}}(t)) \mid \boldsymbol{w}_{\boldsymbol{X}}(s) = \boldsymbol{w}_{\boldsymbol{X}}],$$

we immediately have the following corollary by Lemma A.1:

**Corollary A.2.** *For a fixed time interval $\tau \geq 0$, we have for any time $s$ the transition semigroup $P_s^{s+\tau}$ satisfies the following* Foster-Lyapunov *property w.r.t. $W$:*

$$P_s^{s+\tau} W(\boldsymbol{w}_{\boldsymbol{X}}) \leq C_4 W(\boldsymbol{w}_{\boldsymbol{X}}) + C_5, \tag{A.6}$$

*where $C_4$ and $C_5$ are positive constants.*

*Proof.* By Dynkin's formula (Weinan et al., 2021), we have

$$\begin{aligned}
&\mathbb{E}[e^{C_1\tau} W(\boldsymbol{w}_{\boldsymbol{X}}(s+\tau)) \mid \boldsymbol{w}_{\boldsymbol{X}}(s) = \boldsymbol{w}_{\boldsymbol{X}}] - W(\boldsymbol{w}_{\boldsymbol{X}}) \\
=&\mathbb{E}\left[\int_s^{s+\tau} \mathrm{d}\left(e^{C_1 u} W(\boldsymbol{w}_{\boldsymbol{X}}(u))\right) \middle| \boldsymbol{w}_{\boldsymbol{X}}(s) = \boldsymbol{w}_{\boldsymbol{X}}\right] \\
=&\mathbb{E}\left[\int_s^{s+\tau} e^{C_1 u} \mathcal{A}_u W(\boldsymbol{w}_{\boldsymbol{X}}(u)) + C_1 e^{C_1 u} W(\boldsymbol{w}_{\boldsymbol{X}}(u)) \mathrm{d}u \middle| \boldsymbol{w}_{\boldsymbol{X}}(s) = \boldsymbol{w}_{\boldsymbol{X}}\right] \\
\leq&\mathbb{E}\left[\int_s^{s+\tau} e^{C_1 u} \left(-C_1 W(\boldsymbol{w}_{\boldsymbol{X}}(u)) + C_2\right) + C_1 e^{C_1 u} W(\boldsymbol{w}_{\boldsymbol{X}}(u)) \mathrm{d}u \middle| \boldsymbol{w}_{\boldsymbol{X}}(s) = \boldsymbol{w}_{\boldsymbol{X}}\right] \\
=&\mathbb{E}\left[\int_s^{s+\tau} C_2 e^{C_1 u} \mathrm{d}u \middle| \boldsymbol{w}_{\boldsymbol{X}}(s) = \boldsymbol{w}_{\boldsymbol{X}}\right] \\
=&\frac{C_2}{C_1}\left(e^{C_1\tau} - 1\right),
\end{aligned}$$

and thus Equation A.6 follows by taking

$$C_4 = 1 - e^{-C_1\tau} \quad \text{and} \quad C_5 = \frac{C_2}{C_1}\left(1 - e^{-C_1\tau}\right),$$

which are both positive constants. □

In the further development of this section, we need the following measure of contraction of the total variation distance of probability measures induced by Markov transitions:

**Definition A.1** (Dobrushin ergodic coefficient)**.** *For any Markov transition $T$ and a Lyapunov function $W$ with $W \geq 1$, we define the Dobrushin ergodic coeeficient (Del Moral and Penev, 2017, Definition 8.2.11) as*

$$\beta(T) = \sup_{\boldsymbol{w}_{\boldsymbol{X}}, \boldsymbol{w}'_{\boldsymbol{X}} \in \mathbb{R}^n} \|T(\boldsymbol{w}_{\boldsymbol{X}}, \cdot) - T(\boldsymbol{w}'_{\boldsymbol{X}}, \cdot)\|_{\mathrm{TV}},$$

*where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance between two probability measures. We also define the $W$-Dobrushin ergodic coefficient (Del Moral and Penev, 2017, Definition 8.2.19) as*

$$\beta_W(T) = \sup_{\boldsymbol{w}_{\boldsymbol{X}}, \boldsymbol{w}'_{\boldsymbol{X}} \in \mathbb{R}^n} \frac{\|T(\boldsymbol{w}_{\boldsymbol{X}}, \cdot) - T(\boldsymbol{w}'_{\boldsymbol{X}}, \cdot)\|_W}{1 + W(\boldsymbol{w}_{\boldsymbol{X}}) + W(\boldsymbol{w}'_{\boldsymbol{X}})},$$

*where $\|\cdot\|_W$ is the $W$-norm defined as*

$$\|f\|_W = \sup_{\boldsymbol{w}_{\boldsymbol{X}} \in \mathbb{R}^n} \frac{|f(\boldsymbol{w}_{\boldsymbol{X}})|}{1/2 + W(\boldsymbol{w}_{\boldsymbol{X}})}.$$

The $W$-Dobrushin ergodic coefficient $\beta_W(T)$ has the following properties:

**Proposition A.3.** *For any Markov transitions $T$, $T_1$, and $T_2$, and any measures $\mu_1$ and $\mu_2$ on $R^n$, the $W$-Dobrushin ergodic coefficient $\beta_W(T)$ satisfies:*

$$\beta_W(T_1T_2) \le \beta_W(T_1)\beta_W(T_2), \tag{A.7}$$

*and*

$$\|\mu_1 T - \mu_2 T\|_V \le \beta_V(T)\|\mu_1 - \mu_2\|_V. \tag{A.8}$$

The following proposition is a direct consequence of the property of Gaussian noise in (3.6):

**Proposition A.4.** *For any time interval $\tau \ge 0$, the following Dobrushin local contraction condition is satisfied by the time-inhomogeneous Markov transition semigroup $P_s^{s+\tau}$ uniformly in time $s$:*

$$\beta(P_s^{s+\tau}; C) := \sup_{\boldsymbol{w_X}, \boldsymbol{w'_X} \in C} \|P_s^{s+\tau}(\boldsymbol{w_X}, \cdot) - P_s^{s+\tau}(\boldsymbol{w'_X}, \cdot)\|_{\mathrm{TV}} < 1,$$

*for any compact subset $C \subset \mathbb{R}^n$.*

*Proof.* For any compact subset $C \subset \mathbb{R}^n$, we choose a sufficiently large $R$ such that we have

$$\langle \overline{\boldsymbol{w_X}}, \boldsymbol{b}(s, \boldsymbol{w_X})\rangle < 0, \quad \forall s,\ \|\boldsymbol{w_X}\| > R$$

which is attainable because of Equation A.5.

Fix $\|\boldsymbol{w}(s)\| \le R$ and notice that the norm of $\boldsymbol{w}(s)$ satisfies the following SDE:

$$\mathrm{d}\|\boldsymbol{w_X}(s)\| = \left(\langle \overline{\boldsymbol{w_X}(s)}, \boldsymbol{b}(s, \boldsymbol{w_X}(s))\rangle + \frac{\eta_L}{2}\,\mathrm{tr}\left(\frac{\boldsymbol{I} - \overline{\boldsymbol{w_X}(s)}\ \overline{\boldsymbol{w_X}(s)}^\top}{\|\boldsymbol{w_X}(s)\|}\right)\right)\mathrm{d}s + \sqrt{\eta_L}\langle \overline{\boldsymbol{w_X}(s)}, \mathrm{d}\boldsymbol{B_X}(s)\rangle,$$

we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{w_X}(s+\tau)\|] =& \|\boldsymbol{w_X}(s)\| + \int_s^{s+\tau} \mathbb{E}\left[\langle \overline{\boldsymbol{w_X}(u)}, \boldsymbol{b}(u, \boldsymbol{w_X}(u))\rangle + \frac{\eta_L}{2}\,\mathrm{tr}\left(\frac{\boldsymbol{I} - \overline{\boldsymbol{w_X}(u)}\ \overline{\boldsymbol{w_X}(u)}^\top}{\|\boldsymbol{w_X}(u)\|}\right)\right]\mathrm{d}u,\\
\le& R + \int_s^{s+\tau} \mathbf{1}(\|\boldsymbol{w_X}(u)\| > R)\left[\langle \overline{\boldsymbol{w_X}(u)}, \boldsymbol{b}(u, \boldsymbol{w_X}(u))\rangle + \frac{\eta_L}{2}\frac{n-1}{\|\boldsymbol{w_X}(u)\|}\right]\mathrm{d}u,\\
\le& R + \frac{\tau\eta_L(n-1)}{2R},
\end{aligned}$$

and

$$\mathrm{Var}[\|\boldsymbol{w_X}(s+\tau)\|] = \int_s^{s+\tau}\left(\sqrt{\eta_L}\langle \overline{\boldsymbol{w_X}(s)}, \mathrm{d}\boldsymbol{B_X}(s)\rangle\right)^2 = \eta_L\tau.$$

Therefore, for $\epsilon > 0$, it holds that

$$\|\boldsymbol{w_X}(s+\tau)\| \le R + \frac{\tau\eta_L(n-1)}{2R},$$

with positive probability for any $s$ and $\boldsymbol{w_X}(s) \in C$.

Therefore, by the isotropic property of the Gaussian noise, we have for any Borel set $A \subset \mathbb{R}^n$, there exists $\epsilon_A > 0$ such that

$$\mathbb{P}(\boldsymbol{w_X}(s+\tau) \in A) \ge \epsilon_A \nu(A),$$

where $\nu$ is the Lebesgue measure on $\mathbb{R}^n$, which implied the local Dobrushin contraction condition (Del Moral and Penev, 2017, Proposition 8.2.18). □

With the above definition, we have the following lemma:

**Lemma A.5.** *For any time interval $\tau \ge 0$, there exists a positive function $V$ such that $\beta_V(P_s^{s+\tau}) < 1$ holds uniformly for any time $s$,* i.e.

$$\beta_V(P_s^{s+\tau}) \le e^{-\kappa\tau}, \quad \forall s. \tag{A.9}$$

*Proof.* As shown in Lemma A.4 and Corollary A.2, for any time interval $\tau \geq 0$, the time-inhomogeneous Markov transition semigroup $P_s^{s+\tau}$ satisfies both the Dobrushin local contraction condition and the Foster-Lyapunov condition w.r.t. the Lyapunov function $W$ (A.3), uniformly in time $s$. Therefore, by Del Moral and Penev (2017, Theorem 8.2.21), there exists a positive function $V$ such that $\beta_V(T) < 1$ uniformly for any $s$. Then we take

$$\kappa = -\frac{1}{\tau} \log \sup_s \beta_V(P_s^{s+\tau}),$$

Equation A.9 is thus satisfied.

Without loss of generality, we assume $N = T/\tau \in \mathbb{Z}$. Notice that for any measures $\mu_1$ and $\mu_2$ on $\mathbb{R}^n$, we have

$$\begin{aligned}
\|\mu_1 P_0^t - \mu_2 P_0^t\|_{\mathrm{TV}} \leq & \beta_V(P_0^t)\|\mu_1 - \mu_2\|_{\mathrm{TV}}, \\
\leq & \prod_{i=0}^{N-1} \beta_V(P_{i\tau}^{(i+1)\tau})\|\mu_1 - \mu_2\|_{\mathrm{TV}}, \\
\leq & e^{-\kappa t}\|\mu_1 - \mu_2\|_{\mathrm{TV}} \to 0, \quad \text{as } t \to \infty,
\end{aligned}$$

which implies Equation A.9. □

We are now ready to prove Theorem 3.1 in the main text:

*Proof of Theorem 3.1.* By the definition of the $V$-Dobrushin ergodic coefficient, for any $f$ satisfying $|f(\boldsymbol{w}_{\boldsymbol{X}})| \leq 1/2 + V(\boldsymbol{w}_{\boldsymbol{X}})$, we have

$$|P_0^t f(\boldsymbol{w}_{\boldsymbol{X}}) - P_0^t f(\boldsymbol{w}'_{\boldsymbol{X}})| \leq \beta_V(P_0^t)\left(1 + V(\boldsymbol{w}_{\boldsymbol{X}}) + V(\boldsymbol{w}'_{\boldsymbol{X}})\right) \leq e^{-\kappa t}\left(1 + V(\boldsymbol{w}_{\boldsymbol{X}}) + V(\boldsymbol{w}'_{\boldsymbol{X}})\right),$$

for any $t \geq 0$ and $\boldsymbol{w}_{\boldsymbol{X}}, \boldsymbol{w}'_{\boldsymbol{X}} \in \mathbb{R}^n$.

Then, by the definition of the total variation distance, we have

$$\begin{aligned}
\|P_0^t(\boldsymbol{w}_{\boldsymbol{X}}, \cdot) - P_0^t f(\boldsymbol{w}'_{\boldsymbol{X}})\|_{\mathrm{TV}} = & \sup_{|f(\boldsymbol{w}_{\boldsymbol{X}})| \leq 1/2} \int_{\mathbb{R}^n} \left(P_0^t f(\boldsymbol{w}_{\boldsymbol{X}}) - P_0^t f(\boldsymbol{w}'_{\boldsymbol{X}})\right) \mu(\mathrm{d}\boldsymbol{w}_{\boldsymbol{X}}) \\
\leq & \sup_{|f(\boldsymbol{w}_{\boldsymbol{X}})| \leq 1/2 + V(\boldsymbol{w}_{\boldsymbol{X}})} \int_{\mathbb{R}^n} |P_0^t f(\boldsymbol{w}_{\boldsymbol{X}}) - P_0^t f(\boldsymbol{w}'_{\boldsymbol{X}})| \mu(\mathrm{d}\boldsymbol{w}_{\boldsymbol{X}}) \\
\leq & e^{-\kappa t}\left(1 + V(\boldsymbol{w}_{\boldsymbol{X}}) + V(\boldsymbol{w}'_{\boldsymbol{X}})\right),
\end{aligned}$$

and the statement follows by taking $\boldsymbol{w}_{\boldsymbol{X}} = \mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}_0$ and $\boldsymbol{w}'_{\boldsymbol{X}} = \mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}'_0$. □

# B MISSING PROOFS FOR PHASE II

In this section, we provide the missing proofs for the results in Phase II in Section 3.3.2 of the main text. For convenience, we will assume the time $t$ is reset to 0 after Phase I and thus the initial condition $\boldsymbol{w}(0)$ of the SDE (3.7) is already near the minima manifold $\mathcal{M}$, as shown in Theorem 3.1

Before we dive into the discussion of the effective dynamics of $\boldsymbol{w}(t)$ along the minima manifold $\mathcal{M}$, we first introduce the following lemma that characterizes the space around $\boldsymbol{w}_{\mathcal{M}} \in \mathcal{M}$:

**Proposition B.1.** *For any $\boldsymbol{w}_{\mathcal{M}} \in \mathcal{M}$, the normal space of the manifold $\mathcal{M}$ around $\boldsymbol{w}_{\mathcal{M}}$ is given by*

$$\mathcal{N}(\boldsymbol{w}_{\mathcal{M}}; \mathcal{M}) = \boldsymbol{w}_{\mathcal{M}} + \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X}, \tag{B.1}$$

*and the tangent space of the manifold $\mathcal{M}$ around $\boldsymbol{w}_{\|}$ is given by*

$$\mathcal{T}(\boldsymbol{w}_{\mathcal{M}}; \mathcal{M}) = \boldsymbol{w}_{\mathcal{M}} + \boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^{\perp}, \tag{B.2}$$

*where $\boldsymbol{A}_{\mathcal{M}} = \boldsymbol{A}(\boldsymbol{w}_{\mathcal{M}})$.*

*Proof.* For simplicity, we will also adopt the notation $\boldsymbol{\alpha}_{\mathcal{M}} = \boldsymbol{\alpha}(\boldsymbol{w}_{\mathcal{M}})$.

Consider the following expansion of the gradient $\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$ around $\boldsymbol{w}_{\mathcal{M}}$ up to the first order:

$$\begin{aligned}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D}) &= \frac{1}{n}\left(\boldsymbol{A}_{\mathcal{M}} + O(\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}}\|)\right)\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \\ &= \frac{1}{n}\left(\boldsymbol{A}_{\mathcal{M}} + O(\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}}\|)\right)\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathcal{M}}) \\ &= \frac{1}{n}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathcal{M}}) + o(\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}}\|)\end{aligned} \tag{B.3}$$

where the second to last equality is due to $\boldsymbol{\alpha}_{\mathcal{M}} - \boldsymbol{\alpha}^* \in \boldsymbol{X}^\perp$ and therefore $\boldsymbol{X}^\top(\boldsymbol{\alpha}_\infty - \boldsymbol{\alpha}^*) = \boldsymbol{0}$, and the last equality is because

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}_\infty\| \asymp \|\nabla_{\boldsymbol{w}}\boldsymbol{\alpha}(\boldsymbol{w}_{\mathcal{M}})^\top(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}})\| \asymp \|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}}\|,$$

where $\asymp$ denotes the equivalence up to a constant factor.

Let $\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathcal{M}} = n\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^n$, we have

$$\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D}) = \boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{\epsilon} + o(\|\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}}\|) = \boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|),$$

and Equation B.1 follows by taking $\|\boldsymbol{\epsilon}\| \to 0$. Then, it is straightforward to see Equation B.2 holds by noticing $\boldsymbol{A}_{\mathcal{M}}$ is invertible. □

### B.1 Proof of Lemma 3.3

For readers' convenience, we restate the form of the SDE that we are considering in Phase II:

$$\mathrm{d}\boldsymbol{w}(t) = -\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D}) + \sqrt{\eta_L}\boldsymbol{\sigma}(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t).$$

*Proof of Equation 3.9.* The dynamics in the tangent space $\mathcal{T}(\boldsymbol{w}_{\mathcal{M}}; \mathcal{M})$ are given by

$$\begin{aligned}\mathrm{d}\Delta\boldsymbol{w}_{\parallel}(t) &= -\mathcal{P}_{\mathcal{T}(\boldsymbol{w}_{\mathcal{M}};\mathcal{M})}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D})\mathrm{d}t + \mathcal{P}_{\mathcal{T}(\boldsymbol{w}_{\mathcal{M}};\mathcal{M})}\sqrt{\eta_L}\boldsymbol{\sigma}(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t), \\ &= -\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^\perp}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D})\mathrm{d}t + \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}}\sqrt{\eta_L}\boldsymbol{\sigma}(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t), \\ &= -\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^\perp}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D})\mathrm{d}t + \sqrt{\eta_L}\boldsymbol{\sigma}_{\parallel}(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t),\end{aligned}$$

and the case for the dynamics in the normal space $\mathcal{N}(\boldsymbol{w}_{\mathcal{M}}; \mathcal{M})$ is similar. □

In order to analyze the empirical loss landscape $\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$ locally around $\boldsymbol{w}_{\mathcal{M}}$, we denote $\Delta\boldsymbol{w} = \Delta\boldsymbol{w}_{\parallel} + \Delta\boldsymbol{w}_{\perp} = \boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}}$ and perform the following calculation on the expansion of $\boldsymbol{A}(\boldsymbol{w})$ around $\boldsymbol{w}_{\mathcal{M}}$ up to the first order:

$$\begin{aligned}\nabla\boldsymbol{A}[\Delta\boldsymbol{w}](\boldsymbol{w}_{\mathcal{M}}) = \gamma\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^\top\Delta\boldsymbol{w}\right)\boldsymbol{I} + \gamma(\gamma-2)\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^\top\Delta\boldsymbol{w}\right)\overline{\boldsymbol{w}_{\mathcal{M}}}\ \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \\ + \gamma\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\left(\Delta\boldsymbol{w}\overline{\boldsymbol{w}_{\mathcal{M}}}^\top + \overline{\boldsymbol{w}_{\mathcal{M}}}\left(\Delta\boldsymbol{w}\right)^\top\right) + o(\|\Delta\boldsymbol{w}\|),\end{aligned}$$

where we adopt the notation of the directional derivative:

$$\nabla\boldsymbol{f}[\boldsymbol{v}](\boldsymbol{x}) = \lim_{\epsilon\to 0}\frac{\boldsymbol{f}(\boldsymbol{x} + \epsilon\boldsymbol{v}) - f(\boldsymbol{x})}{\epsilon}, \tag{B.4}$$

for any $\boldsymbol{v}$ of the same dimension as $\boldsymbol{x}$.

Also, by the definition of the matrix $\boldsymbol{A}$ in Equation A.1, we have the expansion of $\alpha(\boldsymbol{w})$ around $\boldsymbol{w}_{\mathcal{M}}$ up to the second order:

$$\boldsymbol{\alpha} - \boldsymbol{\alpha}_{\mathcal{M}} = \boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w} + \frac{1}{2}\nabla^2_{\boldsymbol{w}}\boldsymbol{\alpha}[\Delta\boldsymbol{w}, \Delta\boldsymbol{w}](\boldsymbol{w}_{\mathcal{M}}) + o(\|\Delta\boldsymbol{w}\|^2),$$

where the notation $\nabla^2_{\boldsymbol{w}}\boldsymbol{\alpha}[\Delta\boldsymbol{w}, \Delta\boldsymbol{w}](\boldsymbol{w}_{\mathcal{M}})$ is defined analogously to the directional derivative in Equation B.4.

Then we have the following expansion of the gradient $\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})$ around $\boldsymbol{w}_{\mathcal{M}}$ up to the second order:

$$\begin{aligned}
&\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D}) = \nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{M}}+\Delta\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})\\
=&\frac{1}{n}\big(\boldsymbol{A}(\boldsymbol{w}_{\mathcal{M}})+\nabla\boldsymbol{A}[\Delta\boldsymbol{w}](\boldsymbol{w}_{\mathcal{M}})+o(\|\Delta\boldsymbol{w}\|)\big)\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\alpha}-\boldsymbol{\alpha}_{\mathcal{M}})\\
=&\frac{1}{n}\Bigg(\boldsymbol{A}_{\mathcal{M}}+\gamma\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^\top\Delta\boldsymbol{w}\right)\boldsymbol{I}+\gamma(\gamma-2)\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^\top\Delta\boldsymbol{w}\right)\overline{\boldsymbol{w}_{\mathcal{M}}}\;\overline{\boldsymbol{w}_{\mathcal{M}}}^\top\\
&+\gamma\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\left(\Delta\boldsymbol{w}\overline{\boldsymbol{w}_{\mathcal{M}}}^\top+\overline{\boldsymbol{w}_{\mathcal{M}}}\,(\Delta\boldsymbol{w})^\top\right)+o(\|\Delta\boldsymbol{w}\|)\Bigg)\boldsymbol{X}\boldsymbol{X}^\top\\
&\left(\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}+\frac{1}{2}\nabla^2_{\boldsymbol{w}}\boldsymbol{\alpha}[\Delta\boldsymbol{w},\Delta\boldsymbol{w}](\boldsymbol{w}_{\mathcal{M}})+o(\|\Delta\boldsymbol{w}\|^2)\right)\\
=&\frac{1}{n}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}\\
&+\frac{1}{n}\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\Bigg(\gamma\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}(\Delta\boldsymbol{w})^\top\overline{\boldsymbol{w}_{\mathcal{M}}}+\gamma(\gamma-2)\overline{\boldsymbol{w}_{\mathcal{M}}}\;\overline{\boldsymbol{w}_{\mathcal{M}}}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}(\Delta\boldsymbol{w})^\top\overline{\boldsymbol{w}_{\mathcal{M}}}\\
&+\gamma\left(\Delta\boldsymbol{w}(\Delta\boldsymbol{w})^\top\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top\overline{\boldsymbol{w}_{\mathcal{M}}}+\overline{\boldsymbol{w}_{\mathcal{M}}}(\Delta\boldsymbol{w})^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}\right)\Bigg)\\
&+\frac{1}{2n}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top\nabla^2_{\boldsymbol{w}}\boldsymbol{\alpha}[\Delta\boldsymbol{w},\Delta\boldsymbol{w}](\boldsymbol{w}_{\mathcal{M}})+o(\|\Delta\boldsymbol{w}\|^2).
\end{aligned}\tag{B.5}$$

As a direct corollary of the above expansion, we obtain the Hessian matrix of the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w};\boldsymbol{w}^*,\mathcal{D})$ around $\boldsymbol{w}_{\mathcal{M}}$:

$$\hat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{M}};\boldsymbol{w}^*,\mathcal{D}) = \frac{1}{n}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}.$$

Following the reasoning in the main text, we adopt the label noise model (Blanc et al., 2020; Damian et al., 2021) and the following assumption on the diffusion tensor as in (3.10):

$$\boldsymbol{\sigma}(\boldsymbol{w}(t))\boldsymbol{\sigma}(\boldsymbol{w}(t))^\top := \sigma^2\nabla^2\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D}) \approx \sigma^2\nabla^2\hat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{M}};\boldsymbol{w}^*,\mathcal{D}) = \frac{\sigma^2}{n}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}},\tag{B.6}$$

where the approximation is due to the fact that $\boldsymbol{w}(t)$ is close to $\boldsymbol{w}_{\mathcal{M}}$. A natural and simple solution to Equation B.6 is to take

$$\boldsymbol{\sigma}(\boldsymbol{w}(t)) := \frac{\sigma}{\sqrt{n}}\begin{bmatrix}\boldsymbol{0} & \boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\end{bmatrix},$$

where $\boldsymbol{0}$ is a zero matrix of size $d\times(d-n)$, by which the diffusion term of the SDE (3.7) is simplified as

$$\boldsymbol{\sigma}(\boldsymbol{w}(t))\mathrm{d}\boldsymbol{B}(t) = \frac{\sigma}{\sqrt{n}}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\mathrm{d}\boldsymbol{B}_\perp(t),\tag{B.7}$$

where $\mathrm{d}\boldsymbol{B}_\perp(t)$ is a $n$-dimensional Brownian motion. One can verify that with this diffusion tensor, we have

$$\boldsymbol{\sigma}_\parallel(t) = \boldsymbol{0},\quad \boldsymbol{\sigma}_\perp(t) = \frac{\sigma}{\sqrt{n}}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X},\tag{B.8}$$

which intuitively means that the label noise only affects the normal space of the manifold $\mathcal{M}$. This is in accordance with the discussions in Li et al. (2021b); Ma et al. (2022).

*Proof of Equation 3.12 and Equation 3.13.* Plugging the diffusion tensor (B.8) and the first order expansion of the empirical loss at the small scale $\hat{\mathcal{L}}(\boldsymbol{w}(t+\tau);\boldsymbol{w}^*,\mathcal{D})$ into the dynamics of $\Delta\boldsymbol{w}_\perp(t)$ in (3.9), we have

$$\begin{aligned}
\mathrm{d}\Delta\boldsymbol{w}_\perp(t+\tau) &= -\frac{1}{n}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}(t+\tau)\mathrm{d}\tau+\sqrt{\eta_L}\boldsymbol{\sigma}_\perp(\boldsymbol{w}(t+\tau))\mathrm{d}\boldsymbol{B}(t),\\
&= -\frac{1}{n}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}_\perp(t+\tau)\mathrm{d}\tau+\sqrt{\frac{\sigma^2\eta_L}{n}}\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}\mathrm{d}\boldsymbol{B}_\perp(\tau),
\end{aligned}$$

where the second equality is due to $\boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}}\Delta\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{A}_{\mathcal{M}}\Delta\boldsymbol{w}_\perp$.

Due to the local representation of the normal space in Proposition B.1, we simplify the above dynamics by taking the parametrization $\Delta \boldsymbol{w}_\perp = \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^n$, and obtain

$$\boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \mathrm{d}\boldsymbol{\epsilon}(\tau) = -\frac{1}{n} \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \boldsymbol{\epsilon}(\tau) \mathrm{d}\tau + \sqrt{\frac{\sigma^2 \eta_L}{n}} \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \mathrm{d}\boldsymbol{B}_\perp(\tau), \tag{B.9}$$

which is an Ornstein-Uhlenbeck process. This parametrization is invertible since $\boldsymbol{A}_{\mathcal{M}}$ is invertible and $\boldsymbol{X}$ is full rank.

Rewrite Equation B.9 as

$$\mathrm{d}\boldsymbol{\epsilon}(\tau) = -\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \boldsymbol{\epsilon}(\tau) \mathrm{d}\tau + \sqrt{\frac{\sigma^2 \eta_L}{n}} \mathrm{d}\boldsymbol{B}_\perp(\tau),$$

we obtain the OU process as in Equation 3.12.

As a classical result, the stationary distribution of the above OU process is given by $\boldsymbol{\epsilon}(\tau) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ satisfies

$$\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \boldsymbol{\Sigma} + \boldsymbol{\Sigma} \frac{1}{n} (\boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X})^\top = \frac{\sigma^2 \eta_L}{n} \boldsymbol{I}, \quad \textit{i.e.} \quad \boldsymbol{\Sigma} = \frac{\sigma^2 \eta_L}{2} \left( \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \right)^{-1},$$

and the claim in Equation 3.13 follows. □

Now we are ready to present the proof of Lemma 3.3 in the main text:

*Proof of Lemma 3.3.* From the stationary distribution (3.13), we have

$$\mathbb{E}_{\Delta \boldsymbol{w}_\perp(t)}[\Delta \boldsymbol{w}_\perp] = \boldsymbol{0}, \quad \text{and} \quad \mathbb{E}_{\Delta \boldsymbol{w}_\perp(t)}[\Delta \boldsymbol{w}_\perp \Delta \boldsymbol{w}_\perp^\top] = \frac{\sigma^2 \eta_L}{2} \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \left( \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}, \tag{B.10}$$

where $\mathbb{E}_{\Delta \boldsymbol{w}_\perp(t)}$ denotes taking expectation w.r.t. the stationary distribution of $\Delta \boldsymbol{w}_\perp(t)$ (3.13).

Plugging Equation B.10 into Equation B.5, we have

$$\begin{aligned}
&\mathbb{E}_{\Delta \boldsymbol{w}_\perp(t)}[\nabla_{\boldsymbol{w}} \hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})] \\
=&\frac{\sigma^2 \eta_L}{2n} \|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1} \Bigg( \gamma \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \overline{\boldsymbol{w}_{\mathcal{M}}} + \gamma(\gamma-2) \overline{\boldsymbol{w}_{\mathcal{M}}}\ \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \overline{\boldsymbol{w}_{\mathcal{M}}} \\
&+ \gamma \left( \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}} + \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X}) \overline{\boldsymbol{w}_{\mathcal{M}}} \right) \Bigg) \\
&+ \frac{\sigma^2 \eta_L}{4n} \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{X}^\top \mathbb{E}_{\Delta \boldsymbol{w}_\perp(t)} \left[ \nabla_{\boldsymbol{w}}^2 \boldsymbol{\alpha}[\Delta \boldsymbol{w}_\perp, \Delta \boldsymbol{w}_\perp](\boldsymbol{w}_{\mathcal{M}}) \right] + o(\|\Delta \boldsymbol{w}_\perp(t)\|^2).
\end{aligned} \tag{B.11}$$

The effective force along the tangent space up to the second order is thus obtained by projecting Equation B.11 onto the tangent space, causing all terms in the column space of $\boldsymbol{A}_{\mathcal{M}} \boldsymbol{X}$ vanishing:

$$\begin{aligned}
&- \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^\perp} \mathbb{E}_{\Delta \boldsymbol{w}_\perp(t)}[\nabla_{\boldsymbol{w}} \hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})] \\
=&- \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^\perp} \frac{\sigma^2 \eta_L \gamma}{2n} \|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1} \left( \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \overline{\boldsymbol{w}_{\mathcal{M}}} + (\gamma-2) \overline{\boldsymbol{w}_{\mathcal{M}}}\ \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \overline{\boldsymbol{w}_{\mathcal{M}}} + \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X}) \overline{\boldsymbol{w}_{\mathcal{M}}} \right) \\
=&- \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^\perp} \frac{\sigma^2 \eta_L \gamma}{2n} \|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1} \Bigg( (1+\gamma) \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \boldsymbol{X} \boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}} + (\gamma-2)(1+\gamma) \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \left( \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \boldsymbol{X} \boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}} \right) \overline{\boldsymbol{w}_{\mathcal{M}}} \\
&+ \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \left( \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) + \gamma \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \boldsymbol{X} \boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}} \right) \overline{\boldsymbol{w}_{\mathcal{M}}} \Bigg) \\
=&- \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1} \boldsymbol{X}^\perp} \frac{\sigma^2 \eta_L \gamma}{2n} \|\boldsymbol{w}_{\mathcal{M}}\|^{2\gamma-1} \left( (1+\gamma) \boldsymbol{X} \boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}} + \mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) \overline{\boldsymbol{w}_{\mathcal{M}}} + (\gamma^2-2) \left( \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \boldsymbol{X} \boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}} \right) \overline{\boldsymbol{w}_{\mathcal{M}}} \right),
\end{aligned} \tag{B.12}$$

where the second to last equality is due to

$$\boldsymbol{A}_{\mathcal{M}} \overline{\boldsymbol{w}_{\mathcal{M}}} = \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \left( \boldsymbol{I} + \gamma \overline{\boldsymbol{w}_{\mathcal{M}}}\ \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \right) \overline{\boldsymbol{w}_{\mathcal{M}}} = (1+\gamma) \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \overline{\boldsymbol{w}_{\mathcal{M}}}.$$

By the Sherman-Morrison formula, we have

$$\boldsymbol{A}_{\mathcal{M}}^{-1} = \|\boldsymbol{w}_{\mathcal{M}}\|^{-\gamma}\left(\boldsymbol{I} - \frac{\gamma}{1+\gamma}\overline{\boldsymbol{w}_{\mathcal{M}}}\,\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\right), \tag{B.13}$$

and thus

$$\begin{aligned}\boldsymbol{A}_{\mathcal{M}}^{-1}\overline{\boldsymbol{w}_{\mathcal{M}}} &= \|\boldsymbol{w}_{\mathcal{M}}\|^{-\gamma}\left(\boldsymbol{I} - \frac{\gamma}{1+\gamma}\overline{\boldsymbol{w}_{\mathcal{M}}}\,\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\right)\overline{\boldsymbol{w}_{\mathcal{M}}} = \frac{1}{1+\gamma}\|\boldsymbol{w}_{\mathcal{M}}\|^{-\gamma}\overline{\boldsymbol{w}_{\mathcal{M}}}\\ \boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}} &= \|\boldsymbol{w}_{\mathcal{M}}\|^{-\gamma}\left(\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}} - \frac{\gamma}{1+\gamma}(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}})\overline{\boldsymbol{w}_{\mathcal{M}}}\right),\end{aligned}$$

and recall

$$\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}} = \boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}((\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp})^{\top}\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp})^{-1}(\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp})^{\top} = \boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}((\boldsymbol{X}^{\perp})^{\top}\boldsymbol{A}_{\mathcal{M}}^{-2}\boldsymbol{X}^{\perp})^{-1}(\boldsymbol{X}^{\perp})^{\top}\boldsymbol{A}_{\mathcal{M}}^{-1},$$

by which Equation B.12 can be further simplified as

$$\begin{aligned}&-\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}}\frac{\sigma^2\eta_L\gamma}{2n}\|\boldsymbol{w}_{\mathcal{M}}\|^{2\gamma-1}\left((1+\gamma)\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}} + \operatorname{tr}(\boldsymbol{X}^{\top}\boldsymbol{X})\overline{\boldsymbol{w}_{\mathcal{M}}} + (\gamma^2-2)\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\\ =&-\frac{\sigma^2\eta_L\gamma}{2n}\|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-1}\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}\left((\boldsymbol{X}^{\perp})^{\top}\boldsymbol{A}_{\mathcal{M}}^{-2}\boldsymbol{X}^{\perp}\right)(\boldsymbol{X}^{\perp})^{\top}\left((1+\gamma)\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}} - \gamma\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\overline{\boldsymbol{w}_{\mathcal{M}}}\right.\\ &\left.+\frac{1}{1+\gamma}\operatorname{tr}(\boldsymbol{X}^{\top}\boldsymbol{X})\overline{\boldsymbol{w}_{\mathcal{M}}} + \frac{\gamma^2-2}{1+\gamma}\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\\ =&-\frac{\sigma^2\eta_L\gamma}{2n}\|\boldsymbol{w}_{\mathcal{M}}\|^{2\gamma-1}\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}\left((\boldsymbol{X}^{\perp})^{\top}\boldsymbol{A}_{\mathcal{M}}^{-2}\boldsymbol{X}^{\perp}\right)(\boldsymbol{X}^{\perp})^{\top}\left(-\gamma(1+\gamma)\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\boldsymbol{A}_{\mathcal{M}}^{-1}\overline{\boldsymbol{w}_{\mathcal{M}}}\right.\\ &\left.+\operatorname{tr}(\boldsymbol{X}^{\top}\boldsymbol{X})\boldsymbol{A}_{\mathcal{M}}^{-1}\overline{\boldsymbol{w}_{\mathcal{M}}} + (\gamma^2-2)\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\boldsymbol{A}_{\mathcal{M}}^{-1}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\\ =&-\frac{\sigma^2\eta_L\gamma}{2n}\|\boldsymbol{w}_{\mathcal{M}}\|^{2\gamma-1}\left(\operatorname{tr}(\boldsymbol{X}^{\top}\boldsymbol{X}) - (\gamma+2)\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\right)\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}}\overline{\boldsymbol{w}_{\mathcal{M}}},\end{aligned}$$

where the second to last equality is due to $(\boldsymbol{X}^{\perp})^{\top}\boldsymbol{X} = \boldsymbol{0}$.

The dynamics of $\Delta\boldsymbol{w}_{\|}(t)$ are thus given by

$$\begin{aligned}\mathrm{d}\Delta\boldsymbol{w}_{\|}(t) =& -\mathbb{E}_{\Delta\boldsymbol{w}_{\perp}(t)}\left[\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}}\nabla_{\boldsymbol{w}}\hat{\mathcal{L}}(\boldsymbol{w}(t);\boldsymbol{w}^*,\mathcal{D})\right]\mathrm{d}t\\ =&-\frac{\sigma^2\eta_L\gamma}{2n}\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1}\left(\operatorname{tr}(\boldsymbol{X}^{\top}\boldsymbol{X}) - (\gamma+2)\left(\overline{\boldsymbol{w}_{\mathcal{M}}}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}}\right)\right)\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}}\overline{\boldsymbol{w}_{\mathcal{M}}}\mathrm{d}t.\end{aligned} \tag{B.14}$$

Since the above analysis is based on the assumption that $\boldsymbol{w}(t)$ is close to $\boldsymbol{w}_{\mathcal{M}}$, we conclude that the effective dynamics of $\boldsymbol{w}_{\mathcal{M}}(t)$ up to the second order are given by

$$\mathrm{d}\boldsymbol{w}_{\mathcal{M}}(t) \approx \mathrm{d}\Delta\boldsymbol{w}_{\|}(t) = -\frac{\sigma^2\eta_L\gamma}{2n}\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1}\left(\operatorname{tr}(\boldsymbol{X}^{\top}\boldsymbol{X}) - (\gamma+2)\left(\overline{\boldsymbol{w}_{\mathcal{M}}(t)}^{\top}\boldsymbol{X}\boldsymbol{X}^{\top}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\right)\right)\mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^{\perp}}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\mathrm{d}t,$$

and the statement in Lemma 3.3 follows. □

**Remark B.2.** *As an alternative approach, the effective dynamics are in accordance with Li et al. (2021b, Corollary 5.2) by taking $c = \sigma^2$, where the authors applied the results in Katzenberger (1990) to analyze the dynamics of the gradient flow on the manifold $\mathcal{M}$.*

### B.2 Proof of Theorem 3.4

We restate the definition of the minima manifold $\mathcal{M}$ (3.2) as follows:

$$\mathcal{M} = \left\{\boldsymbol{w}\middle|\boldsymbol{\alpha} - \boldsymbol{\alpha}^* = \|\boldsymbol{w}\|^{\gamma}\boldsymbol{w} - \|\boldsymbol{w}^*\|^{\gamma}\boldsymbol{w}^* \in \boldsymbol{X}^{\perp}\right\}.$$

In the following, we will also use $\boldsymbol{\alpha}_{\boldsymbol{X}}^*$ to denote the projection of $\boldsymbol{\alpha}^*$ onto the column space of $\boldsymbol{X}$, *i.e.* $\boldsymbol{\alpha}_{\boldsymbol{X}}^* = \mathcal{P}_{\boldsymbol{X}}\boldsymbol{\alpha}^*$, and use $\boldsymbol{\alpha}_{\boldsymbol{X}^{\perp}}^*$ to denote the projection of $\boldsymbol{w}^*$ and $\boldsymbol{\alpha}^*$ onto the null space of $\boldsymbol{X}$, *i.e.* $\boldsymbol{\alpha}_{\boldsymbol{X}^{\perp}}^* = \mathcal{P}_{\boldsymbol{X}^{\perp}}\boldsymbol{\alpha}^*$.

We first closely examine the geometry of the minima manifold $\mathcal{M}$ and summarize in the following proposition:

**Proposition B.3.** *For every $\boldsymbol{w}_{\mathcal{M}} \in \mathcal{M}$, it has the following representation:*

$$\boldsymbol{w}_{\mathcal{M}} = \lambda \boldsymbol{\alpha}^*_{\boldsymbol{X}} + \sqrt{\lambda^{-2/\gamma} - \lambda^2 \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2} \overline{\boldsymbol{r}_{\boldsymbol{X}^\perp}},$$

*where $0 < \lambda = \|\boldsymbol{w}_{\mathcal{M}}\|^{-\gamma} \leq \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{-\frac{\gamma}{1+\gamma}}$ and $\overline{\boldsymbol{r}_{\boldsymbol{X}^\perp}}$ is an arbitrary unit vector in $\boldsymbol{X}^\perp$. Then we have*

$$\mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}_{\mathcal{M}} = \lambda \boldsymbol{\alpha}^*_{\boldsymbol{X}}, \quad \textit{and} \quad \mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}} = \sqrt{\lambda^{-2/\gamma} - \lambda^2 \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2} \overline{\boldsymbol{r}_{\boldsymbol{X}^\perp}}.$$

*Proof.* Following the definition of the minima manifold $\mathcal{M}$, $\boldsymbol{w}_{\mathcal{M}} \in \mathcal{M}$ is equivalent to the following equation:

$$\boldsymbol{0} = \boldsymbol{X}^\top \left( \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \boldsymbol{w}_{\mathcal{M}} - \boldsymbol{\alpha}^* \right) = \boldsymbol{X}^\top \left( \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}_{\mathcal{M}} - \boldsymbol{\alpha}^*_{\boldsymbol{X}} \right),$$

where the second equality is due to $\boldsymbol{X}^\top \mathcal{P}_{\boldsymbol{X}^\perp} = \boldsymbol{0}$. Therefore, notice that $\|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}_{\mathcal{M}} - \boldsymbol{\alpha}^*_{\boldsymbol{X}} \in \boldsymbol{X}$, we have

$$\mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}_{\mathcal{M}} = \|\boldsymbol{w}_{\mathcal{M}}\|^{-\gamma} \boldsymbol{\alpha}^*_{\boldsymbol{X}} := \lambda \alpha^*_{\boldsymbol{X}},$$

where $\lambda = \|\boldsymbol{w}_{\mathcal{M}}\|^{-\gamma} > 0$.

Moreover, since

$$\|\boldsymbol{w}_{\mathcal{M}}\|^2 = \|\mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}_{\mathcal{M}}\|^2 + \|\mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}}\|^2 = \lambda^2 \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2 + \|\mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}}\|^2,$$

we have

$$\|\mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}}\|^2 = \|\boldsymbol{w}_{\mathcal{M}}\|^2 - \lambda^2 \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2 = \lambda^{-2/\gamma} - \lambda^2 \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2,$$

and thus $\mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}}$ can be represented as

$$\mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}} = \sqrt{\lambda^{-2/\gamma} - \lambda^2 \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2} \overline{\boldsymbol{r}_{\boldsymbol{X}^\perp}},$$

where $\overline{\boldsymbol{r}_{\boldsymbol{X}^\perp}}$ can take any unit vector in $\boldsymbol{X}^\perp$. The statement in the lemma follows. □

Then we can prove the following lemma, which is the key to the proof of Theorem 3.4:

**Lemma B.4.** *The optimization problem that finds the minimum $L^2$-norm solution of the empirical loss $\hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D})$,* i.e.

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}_{\mathcal{M}}\|, \quad \textit{s.t.} \quad \hat{\mathcal{L}}(\boldsymbol{w}; \boldsymbol{w}^*, \mathcal{D}) = 0,$$

*or equivalently*

$$\min_{\boldsymbol{w}_{\mathcal{M}}} \frac{1}{2} \|\boldsymbol{w}_{\mathcal{M}}\|^2, \quad \textit{s.t.} \quad \boldsymbol{X}^\top \left( \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \boldsymbol{w}_{\mathcal{M}} - \boldsymbol{\alpha}^* \right) = \boldsymbol{0}, \tag{B.15}$$

*has a unique local minimum, and thus the global minimum, which is given by*

$$\boldsymbol{w}^\dagger = \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{-\frac{\gamma}{1+\gamma}} \boldsymbol{\alpha}^*_{\boldsymbol{X}}, \quad \textit{withoptimalvalue} \quad \|\boldsymbol{w}^\dagger\| = \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{\frac{1}{1+\gamma}}$$

*One should notice that $\boldsymbol{w}^\dagger$ is in the column space of $\boldsymbol{X}$,* i.e. $\boldsymbol{w}^\dagger = \mathcal{P}_{\boldsymbol{X}} \boldsymbol{w}^\dagger$.

*Proof.* Apply the KKT conditions to the above constrainted optimization problem (B.15), we have for $\boldsymbol{\mu} \in \mathbb{R}^n$ such that

$$\begin{aligned}
\boldsymbol{0} &= \nabla_{\boldsymbol{w}} \left[ \frac{1}{2} \|\boldsymbol{w}_{\mathcal{M}}\|^2 + \boldsymbol{\mu}^\top \boldsymbol{X}^\top \left( \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \boldsymbol{w}_{\mathcal{M}} - \boldsymbol{\alpha}^* \right) \right] \\
&= \boldsymbol{w}_{\mathcal{M}} + \nabla_{\boldsymbol{w}} \boldsymbol{\alpha}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{\mu}, \\
&= \boldsymbol{w}_{\mathcal{M}} + \left( \|\boldsymbol{w}_{\mathcal{M}}\|^\gamma \boldsymbol{I} + \gamma \|\boldsymbol{w}_{\mathcal{M}}\|^{\gamma-2} \boldsymbol{w}_{\mathcal{M}} \boldsymbol{w}_{\mathcal{M}}^\top \right) \boldsymbol{X} \boldsymbol{\mu}.
\end{aligned}$$

Then consider the projection of the above equation onto the column space of $\boldsymbol{X}$ and its null space, and plug in Proposition B.3, we have

$$\begin{cases}
\lambda \boldsymbol{\alpha}^*_{\boldsymbol{X}} + \lambda^{-1} \boldsymbol{X} \boldsymbol{\mu} + \gamma \lambda^{-1+2/\gamma} \lambda \boldsymbol{\alpha}^*_{\boldsymbol{X}} \left( \lambda \boldsymbol{\alpha}^*_{\boldsymbol{X}} \right)^\top \boldsymbol{X} \boldsymbol{\mu} &= \boldsymbol{0}, \\
\mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}} + \gamma \lambda^{-1+2/\gamma} \mathcal{P}_{\boldsymbol{X}^\perp} \boldsymbol{w}_{\mathcal{M}} \left( \lambda \boldsymbol{\alpha}^*_{\boldsymbol{X}} \right)^\top \boldsymbol{X} \boldsymbol{\mu} &= \boldsymbol{0}.
\end{cases} \tag{B.16}$$

From the first equation of Equation B.16, we see $\boldsymbol{X}\boldsymbol{\mu} \propto \boldsymbol{\alpha}_{\boldsymbol{X}}^*$, whereby we set $\boldsymbol{X}\boldsymbol{\mu} = \kappa\boldsymbol{\alpha}_{\boldsymbol{X}}^*$, and obtain

$$\lambda\boldsymbol{\alpha}_{\boldsymbol{X}}^* + \lambda^{-1}\kappa\boldsymbol{\alpha}_{\boldsymbol{X}}^* + \gamma\lambda^{1+2/\gamma}\kappa\left(\boldsymbol{\alpha}_{\boldsymbol{X}}^*\right)^\top \boldsymbol{\alpha}_{\boldsymbol{X}}^*\boldsymbol{\alpha}_{\boldsymbol{X}}^* = \boldsymbol{0},$$

which solves $\kappa$ as

$$\kappa = -\frac{\lambda^2}{1+\lambda^{2+2/\gamma}\gamma\|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2}.$$

Then the second equation of Equation B.16 becomes

$$\left(1+\gamma\lambda^{2/\gamma}\kappa\|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2\right)\mathcal{P}_{\boldsymbol{X}^\perp}\boldsymbol{w}_{\mathcal{M}} = \boldsymbol{0}.$$

However, since

$$1+\gamma\lambda^{2/\gamma}\kappa\|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2 = 1 - \frac{\lambda^{2+2/\gamma}\gamma\|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2}{1+\lambda^{2+2/\gamma}\gamma\|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2} = \frac{1}{1+\lambda^{2+2/\gamma}\gamma\|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^2} > 0,$$

we must have $\mathcal{P}_{\boldsymbol{X}^\perp}\boldsymbol{w}_{\mathcal{M}} = \boldsymbol{0}$, which means $\boldsymbol{w}_{\mathcal{M}} = \lambda\boldsymbol{\alpha}_{\boldsymbol{X}}^*$. By Proposition B.3, we have

$$\lambda^{-1/\gamma} = \|\boldsymbol{w}_{\mathcal{M}}\| = \lambda\|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|,$$

and thus

$$\lambda = \|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^{-\frac{\gamma}{1+\gamma}}.$$

It is straightforward to verify that the linear independence constraint qualification (LICQ) is always satisfied, and thus $\boldsymbol{w}^\dagger = \|\boldsymbol{\alpha}_{\boldsymbol{X}}^*\|^{-\frac{\gamma}{1+\gamma}}\boldsymbol{\alpha}_{\boldsymbol{X}}^*$ is the unique local minimum of the optimization problem (B.15), and therefore the global minimum. □

The following lemma is drawn from the random matrix theory:

**Lemma B.5** (Vershynin (2018, Theorem 4.4.5))**.** *Let $\boldsymbol{A}$ by an $m \times n$ random matrix with i.i.d. entries drawn from $\mathcal{N}(0,1)$. Then for any $t > 0$, we have*

$$\|\boldsymbol{A}\| \le \mathcal{O}(\sqrt{m} + \sqrt{n} + t),$$

*with probability at least $1 - 2\exp(-t^2)$.*

Using the above lemma, we can prove the following lemma:

**Lemma B.6.** *Let $\boldsymbol{X}$ be the randomly generated data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, with $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{id}) \sim \mathcal{N}(0, \boldsymbol{I})$ for $i \in [n]$. For sufficiently large $d$ and $n$, we have*

$$C(\boldsymbol{w}_{\mathcal{M}}) = \frac{1}{n}\left(\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}) - (\gamma+2)\|\boldsymbol{X}^\top\overline{\boldsymbol{w}_{\mathcal{M}}}\|^2\right) \ge \epsilon d,$$

*for all $\boldsymbol{w}_{\mathcal{M}} \in \mathcal{M}$, with probability at least $1 - \exp(-\Omega(nd))$, where $\epsilon$ is an arbitrary small positive constant.*

*Proof.* By Lemma B.5, we have

$$\|\boldsymbol{X}^\top\overline{\boldsymbol{w}_{\mathcal{M}}}\|^2 \le \|\boldsymbol{X}^\top\|^2\|\overline{\boldsymbol{w}_{\mathcal{M}}}\|^2 = \|\boldsymbol{X}\|^2 \le \mathcal{O}((\sqrt{n}+\sqrt{d}+t)^2),$$

with probability at least $1 - 2\exp(-t^2)$. We take $t = \sqrt{d}$ and obtain

$$\|\boldsymbol{X}^\top\overline{\boldsymbol{w}_{\mathcal{M}}}\|^2 \le \mathcal{O}(d), \quad \text{with probability at least} \quad 1 - 2\exp(-d^2).$$

By the central limit theorem,

$$\sqrt{nd}\left[\frac{\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X})}{nd} - 1\right] = \sqrt{nd}\left[\frac{\sum_{i=1}^n\sum_{j=1}^d x_{ij}^2}{nd} - 1\right] \xrightarrow{D} \mathcal{N}(0,2),$$

and thus for any $K > 0$, $\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) \geq nd/2$ with probability at least

$$1 - \mathcal{P}(\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) \leq nd/2) \to 1 - \Phi\left(\frac{nd/2 - nd}{\sqrt{2nd}}\right) \geq 1 - \Phi(-\Omega(\sqrt{nd})) \geq 1 - \exp(-\Omega(nd)).$$

Therefore, with a probability of at least

$$1 - \exp(-\Omega(nd)) - 2\exp(-d^2) \geq 1 - \exp(-\Omega(nd)),$$

we have

$$C(\boldsymbol{w}_{\mathcal{M}}) = \frac{1}{n}\left(\mathrm{tr}(\boldsymbol{X}^\top \boldsymbol{X}) - (\gamma+2)\|\boldsymbol{X}^\top \overline{\boldsymbol{w}_{\mathcal{M}}}\|^2\right) \geq \frac{1}{n}\left(\frac{nd}{2} - (\gamma+2)\mathcal{O}(d)\right) \geq \Omega(d),$$

The statement in the lemma follows. □

We are now ready to prove Theorem 3.4:

*Proof of Theorem 3.4.* By Lemma B.6, we have with probability at least $1 - \exp(-\Omega(nd))$, the data matrix $\boldsymbol{X}$ and the ground truth $\boldsymbol{w}^*$ permits the existence of following constant

$$\underline{C} = \inf_{t \geq 0} \frac{C(\boldsymbol{w}_{\mathcal{M}}(t))}{d} > 0,$$

The following ODE of $\|\boldsymbol{w}_{\mathcal{M}}(t)\|$ is directly computed from Equation 3.15 from Lemma 3.3:

$$\begin{aligned}\frac{\mathrm{d}\|\boldsymbol{w}_{\mathcal{M}}(t)\|}{\mathrm{d}t} &= -\frac{\sigma^2\eta_L\gamma}{2}\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1}C(\boldsymbol{w}_{\mathcal{M}}(t))\langle\overline{\boldsymbol{w}}_{\mathcal{M}}(t), \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^\perp}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\rangle \\ &\leq -\frac{\sigma^2\eta_L\gamma\underline{C}d}{2}\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1}\langle\overline{\boldsymbol{w}}_{\mathcal{M}}(t), \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^\perp}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\rangle.\end{aligned} \tag{B.17}$$

Define $\boldsymbol{v}(t) = (\boldsymbol{X}^\perp)^\top\overline{\boldsymbol{w}_{\mathcal{M}}(t)}$, by Equation B.13, we have

$$\begin{aligned}((\boldsymbol{X}^\perp)^\top\boldsymbol{A}_{\mathcal{M}}^{-2}\boldsymbol{X}^\perp)^{-1} &= \left(\left(\boldsymbol{X}^\perp\right)^\top\|\boldsymbol{w}_{\boldsymbol{X}}(t)\|^{-2\gamma}\left(\boldsymbol{I} - \frac{\gamma}{1+\gamma}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\,\overline{\boldsymbol{w}_{\mathcal{M}}(t)}^\top\right)^2\boldsymbol{X}^\perp\right)^{-1} \\ &= \|\boldsymbol{w}_{\boldsymbol{X}}(t)\|^{2\gamma}\left(\left(\boldsymbol{X}^\perp\right)^\top\left(\boldsymbol{I} - \frac{\gamma^2+2\gamma}{(1+\gamma)^2}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\,\overline{\boldsymbol{w}_{\mathcal{M}}(t)}^\top\right)\boldsymbol{X}^\perp\right)^{-1} \\ &= \|\boldsymbol{w}_{\boldsymbol{X}}(t)\|^{2\gamma}\left(\boldsymbol{I} - \frac{\gamma^2+2\gamma}{(1+\gamma)^2}\boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right)^{-1} \\ &= \|\boldsymbol{w}_{\boldsymbol{X}}(t)\|^{2\gamma}\left(\boldsymbol{I} + \frac{\frac{\gamma^2+2\gamma}{(1+\gamma)^2}\boldsymbol{v}(t)\boldsymbol{v}(t)^\top}{1+\frac{\gamma^2+2\gamma}{(1+\gamma)^2}\|\boldsymbol{v}(t)\|^2}\right),\end{aligned}$$

where the last equality is due to the Sherman-Morrison formula.

Then we compute as follows:

$$\begin{aligned}\langle\overline{\boldsymbol{w}}_{\mathcal{M}}(t), \mathcal{P}_{\boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^\perp}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\rangle &= \langle\overline{\boldsymbol{w}_{\mathcal{M}}(t)}, \boldsymbol{A}_{\mathcal{M}}^{-1}\boldsymbol{X}^\perp((\boldsymbol{X}^\perp)^\top\boldsymbol{A}_{\mathcal{M}}^{-2}\boldsymbol{X}^\perp)^{-1}(\boldsymbol{X}^\perp)^\top\boldsymbol{A}_{\mathcal{M}}^{-1}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\rangle \\ &= \frac{\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-2\gamma}}{(1+\gamma)^2}\langle\overline{\boldsymbol{w}_{\mathcal{M}}(t)}, \boldsymbol{X}^\perp((\boldsymbol{X}^\perp)^\top\boldsymbol{A}_{\mathcal{M}}^{-2}\boldsymbol{X}^\perp)^{-1}(\boldsymbol{X}^\perp)^\top\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\rangle, \\ &= \frac{\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-2\gamma}}{(1+\gamma)^2}\langle\boldsymbol{v}(t), \|\boldsymbol{w}_{\boldsymbol{X}}(t)\|^{2\gamma}\left(\boldsymbol{I} + \frac{\frac{\gamma^2+2\gamma}{(1+\gamma)^2}\boldsymbol{v}(t)\boldsymbol{v}(t)^\top}{1+\frac{\gamma^2+2\gamma}{(1+\gamma)^2}\|\boldsymbol{v}(t)\|^2}\right)\boldsymbol{v}(t)\rangle, \\ &= \frac{1}{(1+\gamma)^2}\|\boldsymbol{v}(t)\|^2\left(1 + \frac{\frac{\gamma^2+2\gamma}{(1+\gamma)^2}\|\boldsymbol{v}(t)\|^2}{1+\frac{\gamma^2+2\gamma}{(1+\gamma)^2}\|\boldsymbol{v}(t)\|^2}\right) \geq \frac{1}{(1+\gamma)^2}\|\boldsymbol{v}(t)\|^2,\end{aligned}$$

where the second equality is by Equation B.13.

Plugging into Equation B.17, we have

$$\frac{\mathrm{d}\|\boldsymbol{w}_{\mathcal{M}}(t)\|}{\mathrm{d}t} \leq -\frac{\sigma^2\eta_L\gamma\underline{C}d}{2(1+\gamma)^2}\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1}\|\boldsymbol{v}(t)\|^2. \tag{B.18}$$

We consider the representation of $\boldsymbol{w}_{\mathcal{M}}(t)$ in Proposition B.3 as $\lambda(t) = \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-\gamma}$, by which we have

$$\begin{aligned}
\|\boldsymbol{v}(t)\|^2 &= \overline{\boldsymbol{w}_{\mathcal{M}}(t)}^\top \boldsymbol{X}^\perp(\boldsymbol{X}^\perp)^\top\overline{\boldsymbol{w}_{\mathcal{M}}(t)} = \overline{\boldsymbol{w}_{\mathcal{M}}(t)}^\top \boldsymbol{X}^\perp(\boldsymbol{X}^\perp)^\top\boldsymbol{X}^\perp(\boldsymbol{X}^\perp)^\top\overline{\boldsymbol{w}_{\mathcal{M}}(t)} \\
&= \|\mathcal{P}_{\boldsymbol{X}^\perp}\overline{\boldsymbol{w}_{\mathcal{M}}(t)}\|^2 = \frac{\|\mathcal{P}_{\boldsymbol{X}^\perp}\boldsymbol{w}_{\mathcal{M}}(t)\|^2}{\|\boldsymbol{w}_{\mathcal{M}}(t)\|^2} \\
&= \frac{\lambda(t)^{-2/\gamma} - \lambda(t)^2\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2}{\lambda(t)^{-2/\gamma}} = 1 - \lambda(t)^{2+2/\gamma}\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2 \\
&= 1 - \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-2-2\gamma}\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2.
\end{aligned}$$

Plugging into Equation B.18, we have

$$\frac{\mathrm{d}\|\boldsymbol{w}_{\mathcal{M}}(t)\|}{\mathrm{d}t} \leq -\frac{\sigma^2\eta_L\gamma\underline{C}d}{2(1+\gamma)^2}\left(\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1} - \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-3}\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2\right).$$

For any $\gamma > -1$, since $\boldsymbol{w}^\dagger$ is the unique minimizer, we have $\|\boldsymbol{w}_{\mathcal{M}}(t)\| \geq \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{\frac{1}{1+\gamma}}$, and thus

$$\|\boldsymbol{w}_{\mathcal{M}}(t)\|^{2\gamma-1} \geq \|\boldsymbol{w}_{\mathcal{M}}(t)\|^{-3}\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2,$$

where the equality holds if and only if $\boldsymbol{w}_{\mathcal{M}}(t) = \boldsymbol{w}^\dagger$. Consequently, $\|\boldsymbol{w}_{\mathcal{M}}(t)\|$ converges to $\|\boldsymbol{w}^\dagger\|$.

For any $\gamma > 1/2$, by taking $\delta(t) = \|\boldsymbol{w}(t)\| - \|\boldsymbol{w}^\dagger\| = \|\boldsymbol{w}(t)\| - \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{\frac{1}{1+\gamma}}$, we have

$$\begin{aligned}
\frac{\mathrm{d}\delta(t)}{\mathrm{d}t} &\leq -\frac{\sigma^2\eta_L\gamma\underline{C}d}{2(1+\gamma)^2}\left(\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{\frac{2\gamma-1}{1+\gamma}} + (2\gamma-1)\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{\frac{2\gamma-2}{1+\gamma}}\delta(t) - \|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^{\frac{-3}{1+\gamma}}\|\boldsymbol{\alpha}^*_{\boldsymbol{X}}\|^2\right) \\
&\leq -\sigma^2\eta_L\underline{C}d\frac{\gamma(2\gamma-1)}{2(\gamma+1)^2}\delta(t),
\end{aligned}$$

solving which yields

$$\delta(t) \leq \delta(0)\exp\left(-\sigma^2\eta_L\underline{C}d\frac{\gamma(2\gamma-1)}{2(\gamma+1)^2}t\right). \tag{B.19}$$

And the statement in the theorem follows. □

**Remark B.7.** *Besides Remark 3.5 for Theorem 3.4 in the main text, one should also notice that although the convergence of $\|\boldsymbol{w}_{\mathcal{M}}(t)\|$ to $\|\boldsymbol{w}^\dagger\|$ depends on the parameter $\gamma$, representing the depth of the neural network, in Equation B.19, its convergence rate has an upper bound as $\gamma \to \infty$.*

# C MISSING PROOF FOR PHASE III

In this section, we prove Theorem 3.6. For convenience, we will assume the time $t$ is reset to 0 at the beginning of Phase III, and we suppose $\boldsymbol{w}(0)$ is near a point $\boldsymbol{w}_{\mathcal{M}}$ on the minima manifold $\mathcal{M}$. Specifically, we consider the projection of the dynamics $\boldsymbol{w}(t)$ onto the tangent space $\mathcal{T}(\boldsymbol{w}_{\mathcal{M}};\mathcal{M})$ and the normal space $\mathcal{N}(\boldsymbol{w}_{\mathcal{M}};\mathcal{M})$ of the minima manifold $\mathcal{M}$ around $\boldsymbol{w}_{\mathcal{M}}$, *i.e.*

$$\boldsymbol{w}(t) = \boldsymbol{w}_{\mathcal{M}} + \Delta\boldsymbol{w}_{\|}(t) + \Delta\boldsymbol{w}_{\perp}(t),$$

As argued in Section 3.3.2, the dynamics of $\Delta\boldsymbol{w}_{\perp}(t)$ are faster than that of $\Delta\boldsymbol{w}_{\|}(t)$ by a factor of $\Theta(\eta_L)$. Thus, we assume $\Delta\boldsymbol{w}_{\|}(t) \equiv 0$ during Phase III and focus on the dynamics of $\Delta\boldsymbol{w}_{\perp}(t)$ (or equivalently rescale the time $t$ with $\tau$ of a smaller scale as we did for Phase II). Also, it is straightforward to see that $\Delta\boldsymbol{w}_{\perp}(t) \in \boldsymbol{A}_{\mathcal{M}}\boldsymbol{X}$ by Proposition B.1.

*Proof of Theorem 3.6.* We approximate the landscape $\hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D})$ by its second-order Taylor expansion around $\boldsymbol{w}_{\mathcal{M}}$:

$$\hat{\mathcal{L}}(\boldsymbol{w}(t); \boldsymbol{w}^*, \mathcal{D}) \approx \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}})^\top \nabla^2 \hat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{M}}; \boldsymbol{w}^*, \mathcal{D})(\boldsymbol{w} - \boldsymbol{w}_{\mathcal{M}}),$$

due to

$$\hat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{M}}; \boldsymbol{w}^*, \mathcal{D}) = 0 \quad \text{and} \quad \nabla \hat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{M}}; \boldsymbol{w}^*, \mathcal{D}) = \boldsymbol{0}.$$

And the dynamics of $\Delta \boldsymbol{w}_\perp(t)$ are thus approximated by

$$\frac{\mathrm{d}\Delta \boldsymbol{w}_\perp(t)}{\mathrm{d}t} = -\nabla^2 \hat{\mathcal{L}}(\boldsymbol{w}_{\mathcal{M}}; \boldsymbol{w}^*, \mathcal{D}) \Delta \boldsymbol{w}_\perp(t) = -\frac{1}{n} \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}} \Delta \boldsymbol{w}_\perp(t).$$

Since $\Delta \boldsymbol{w}_\perp(t) \in \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X}$, let $\Delta \boldsymbol{w}_\perp(t) = \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \boldsymbol{\epsilon}(t)$, we have

$$\boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \frac{\mathrm{d}\boldsymbol{\epsilon}(t)}{\mathrm{d}t} = -\frac{1}{n} \boldsymbol{A}_{\mathcal{M}} \boldsymbol{X} \left( \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \right) \boldsymbol{\epsilon}(t), \quad \textit{i.e.} \quad \frac{\mathrm{d}\boldsymbol{\epsilon}(t)}{\mathrm{d}t} = -\frac{1}{n} \left( \boldsymbol{X}^\top \boldsymbol{A}_{\mathcal{M}}^2 \boldsymbol{X} \right) \boldsymbol{\epsilon}(t).$$

Then we have

$$\frac{\mathrm{d}\|\boldsymbol{\epsilon}(t)\|}{\mathrm{d}t} = -\frac{1}{n} \langle \overline{\boldsymbol{\epsilon}(t)}, (\boldsymbol{X}^\top (\boldsymbol{A}_{\mathcal{M}})^2 \boldsymbol{X}) \boldsymbol{\epsilon}(t) \rangle \leq -\frac{\lambda_{\min}(\boldsymbol{X}^\top (\boldsymbol{A}_{\mathcal{M}})^2 \boldsymbol{X})}{n} \|\boldsymbol{\epsilon}(t)\|,$$

where $\lambda_{\min}(\boldsymbol{X}^\top (\boldsymbol{A}_{\mathcal{M}})^2 \boldsymbol{X})$ is the smallest eigenvalue of $\boldsymbol{X}^\top (\boldsymbol{A}_{\mathcal{M}})^2 \boldsymbol{X}$, which is guaranteed to be positive, given $\boldsymbol{X}$ is full rank and

$$\boldsymbol{X}^\top (\boldsymbol{A}_{\mathcal{M}})^2 \boldsymbol{X} = \|\boldsymbol{w}_{\mathcal{M}}\|^{2\gamma} \boldsymbol{X}^\top \left( \boldsymbol{I} + \gamma \overline{\boldsymbol{w}_{\mathcal{M}}} \, \overline{\boldsymbol{w}_{\mathcal{M}}}^\top \right)^2 \boldsymbol{X}$$

is positive definite. Thus, we have the exponential convergence

$$\|\boldsymbol{\epsilon}(t)\| \leq \|\boldsymbol{\epsilon}(0)\| \exp\left( -\frac{\lambda_{\min}(\boldsymbol{X}^\top (\boldsymbol{A}_{\mathcal{M}})^2 \boldsymbol{X})}{n} t \right),$$

and thus

$$\|\Delta \boldsymbol{w}_\perp(t)\| \leq \|\boldsymbol{A}_{\mathcal{M}} \boldsymbol{X}\| \|\boldsymbol{\epsilon}(t)\| \leq \|\boldsymbol{A}_{\mathcal{M}}\| \|\boldsymbol{X}\| \|\boldsymbol{\epsilon}(0)\| \exp\left( -\frac{\lambda_{\min}(\boldsymbol{X}^\top (\boldsymbol{A}^\dagger)^2 \boldsymbol{X})}{n} t \right).$$

And the statement in the theorem follows. □