# Three-stage Evolution and Fast Equilibrium for SGD with Non-degerate Critical Points

**Yi Wang** [1] **Zhiren Wang** [2]

## Abstract

We justify the fast equilibrium conjecture on stochastic gradient descent from (Li et al., 2020) under the assumptions that critical points are non-degenerate and the stochastic noise is a standard Gaussian. In this case, we prove an scaling invaraint SGD with constant effective learning rate consists of three stages: descent, diffusion and tunneling, and explicitly identify temporary equilibrium states that can be observed within practical training time. This interprets the gap between the mixing time in the fast equilibrium conjecture and the previously known upper bound. While our assumptions do not represent typical implementations of SGD of neural networks in practice, this is the first description of the three-stage mechanism in any case. The main finding in this mechanism is that a temporary equilibrium of local nature is quickly achieved after polynomial time (in term of the reciprocal of the intrinsic learning rate) and then stabilizes within observable time scales; and that the temporary equilibrium is in general different from the global Gibbs equilibrium, which will only appear after an exponentially long period beyond typical training limits. Our experiments support that this mechanism may extend to the general case.

## 1. Introduction

### 1.1. Background and motivation

Stochastic gradient descent (SGD) has been an indispensable tool in the training of neural networks and is known to work in both convex and non-convex settings. One theme that has recently attracted much attention is the description of the asymptoic behavior of the SGD, especially how learning rate schemes affect the distribution of random trajectories. A useful approach is to model the SGD

$$\boldsymbol{w}_{k+1} \leftarrow (1-\lambda)\boldsymbol{w}_k - \eta \nabla \mathcal{L}_{\mathcal{B}_k}(\boldsymbol{w}_k), \tag{1}$$

where $\mathcal{L}_{\mathcal{B}_k}$ is the loss function over a stochastic batch $\mathcal{B}_k$ and $\lambda$ and $\eta$ are respectively the weight decay and the learning rate, by stochastic differential equations (SDE). In this approach, the random noise in $\nabla \mathcal{L}_{\mathcal{B}_k}(\boldsymbol{w}_k)$ is regarded, by taking a continuous limit on time increments, as a Brownian motion with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{w}_k)$. A widely adopted approach is to approximate the SGD (1) by the value at $t = k\eta$ of the SDE:

$$\mathrm{d}\boldsymbol{W}_t = -\eta(\nabla \mathcal{L}(\boldsymbol{W}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\boldsymbol{W}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d) - \lambda_e \boldsymbol{W}_t \mathrm{d}t, \tag{2}$$

starting from $\boldsymbol{W}_0 = \boldsymbol{w}_0$, where $\lambda_e = \lambda\eta$, and $\boldsymbol{B}_t^d$ stands for the standard Brownian motion in $\mathbb{R}^d$ (see §2.1 for more details). The quantity $\lambda_e$ is called the **intrinsic learning rate** and is known to decide the limit behavior of the dynamics.

Modern neural networks often contain various normalization steps, such as batch normalization, weight normalization and layer normalization. Normalization makes the loss function $\mathcal{L}$ **scaling invariant** and typically non-smooth near the origin. A priori, it may take a long time for the SDE (2) to reach an equilibrium as $|\boldsymbol{W}_t|$ may remain too large or too small for a long period. Recently (Li et al., 2020) studied how $\lambda_e$ affects the distribution of the solutions through another quantity $\gamma_t^{-\frac{1}{2}} := |\boldsymbol{W}_t|^{-2}\eta$, called the **effective learning rate**. Under mild assumptions, they proved that $\gamma_t^{-\frac{1}{2}}$ stabilizes to the magnitude $O(\lambda_e^{\frac{1}{2}})$ after $O(\frac{1}{\lambda_e})$ time. They further conjectured:

**Conjecture 1.1. (Fast Equilibrium Conjecture)** (Li et al., 2020) Suppose $F(\boldsymbol{W}, \boldsymbol{x})$ is the output of the neural network with parameter $\boldsymbol{W}$ and input data $\boldsymbol{x}$, then the distribution of $F(\boldsymbol{W}_t, \boldsymbol{x})$ stabilizes in total variation distance for all $\boldsymbol{x}$ after $O(\frac{1}{\lambda_e})$ time to an equilibrium state. Moreover, this distribution is independent of the initial parameter $\boldsymbol{W}_0$.

Remark that the conjecture concerns the speed of convergence of a sequence of distributions towards an equilibrium state in terms of the total variation distance between probability measures. An interesting but different question in

[1]Department of Mathematics, Johns Hopkins University [2]Department of Mathematics, Pennsylvania State University. Correspondence to: Yi Wang <ywang@math.jhu.edu>, Zhiren Wang <zhirenw@psu.edu>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

similar settings, about convergence of a typical trajectory towards a local minima in terms of the gap in the loss function, has been studied in (Raginsky et al., 2017; Zhang et al., 2017; Xu et al., 2018; Huang & Becker, 2021).

While the conjecture is supported by numerical experiments, it currently lacks theoretical explanation. The relation between the convergence time of the SDE model and the learning rate $\lambda_e$ has been studied in (Bovier et al., 2004; Shi et al., 2020) and the best known upper bound for mixing time is $O(e^{C\lambda_e^{-1}})$ for networks without normalization. For systems with normalization, after adapting the stabilized value of effective learning rate $O(\lambda_e^{\frac{1}{2}})$ in (Li et al., 2020), in lieu of $\lambda_e$, this bound becomes $O(e^{C\lambda_e^{-\frac{1}{2}}})$, which is much larger than $O(\frac{1}{\lambda_e})$. This gap has not been theoretically explained and will be the main focus of this paper.

Thanks to scaling-invariance, the distribution of $F(\boldsymbol{W}_t, \boldsymbol{x})$ is determined by the distribution of $\frac{\boldsymbol{W}_t}{|\boldsymbol{W}_t|}$. We will study this later distribution on the unit sphere $\mathbb{S}^{d-1}$ and prove that it displays fast convergence to certain temporary equilibria. It turns out that, unlike in Conjecture 1.1, the temporary equilibrium of $\frac{\boldsymbol{W}_t}{|\boldsymbol{W}_t|}$ after such fast convergence does depend on the initial value $\boldsymbol{W}_0$. For possibile interpretations that reconcile such dependence with the initial parameter independence in Conjecture 1.1, see the discussions in §6.1.

We also note that several earlier works, such as (Mandt et al., 2017; Izmailov et al., 2018) investigated local mixing in the convex optimization case, when there is only one local minimum and the loss function is assumed to be quadratic. Compared to these works, the current paper deals with other obstructions that arise in a non-convex setting. Namely, we will mainly focus on the separation between stages, which is not an issue in the convex case since that case does not have a final stage during which trajectories move across basins.

### 1.2. Restriction of mathematical tools and our goal

It is worth noting that both the works (Bovier et al., 2004; Shi et al., 2020) assumed two hypothesis: (i) the noise $\boldsymbol{\Sigma}$ is the standard Gaussian (see Assumption 4.1), (ii) $\mathcal{L}$ is a Morse function, i.e. all critical points are non-degenerate and hence isolated (see Assumption 4.2). The reason lies in the limitation of mathematical tools: the only currently available mathematical theory that applies to the study of convergence of distributions towards equilibria around local minima is the seminal work of Barry Simon (Simon, 1983) on semiclassical analysis of low lying eigenvalues. However this theory is limited to the two assumption above. It is unclear whether similar conclusions can be achieved beyond these two assumptions without a major update to the underlying mathematical theory.

As no such updates exist to date, the goal of this paper is to work within the framework of Simon's theory and to explain the huge gap between the experimentally supported upper bound in Conjecture 1.1 and the previously known exponentially large upper bounds under these assumptions.

### 1.3. Our contributions

The main contributions of this paper are:

(1) We derive a spherical SDE model (Definition 3.2) of the SGD with constant effective learning rate. This model focuses on the normalized parameter $\frac{\boldsymbol{W}_t}{|\boldsymbol{W}_t|}$ and uses intrinsic differential operators of $\mathbb{S}^{d-1}$ (instead of those in $\mathbb{R}^d$). Since the output of the neural network depends only on $\frac{\boldsymbol{W}_t}{|\boldsymbol{W}_t|}$, this does not affect Conjecture 1.1. This spherical SDE is also the mathematical model of the Riemannian realization of batch normalization in (Cho & Lee, 2017).

(2) We introduce, for the first time to the best of our knowledge, the three-stage description of the SGD: **descent**, **diffusion** and **tunneling**. The descent stage sends, with high probability, a point near the bottom of attracting basin containing it. The diffusion stages stabilizes the distribution towards a temporary Gibbs equilibrium that is locally Gibbs within each individual attracting basin. The tunneling stage allows mass to slowly leak between basins to achieve the unique global Gibbs state. The three stages respectively take at most $O(\frac{1}{\lambda_e^{\frac{1}{2}}})$, $O(\frac{1}{\lambda_e})$ , $O(e^{\frac{C}{\lambda_e}})$ in time. It was previously known (Shi et al., 2020) that convergence towards the global Gibbs state is exponentially slow. But to our best knowledge our result for the first time identifies the distinction between the three stages, especially the fast-slow contrast between the diffusion and tunneling stages.

(3) Our proof is based on a completely new strategy. Instead of fully relying on Simon's semiclassical analysis, we complement this theory with a probabilistic argument (Lemma E.6) that characterizes microscopically the difficulty for an individual trajectory to escape from an attracting basin. This is the key to materialize the gap between small and large non-zero eigenvalues discovered in (Simon, 1983), which leads to the aforementioned separation between the diffusion and tunneling stages. Previous works (Bovier et al., 2004; Shi et al., 2020) only used the gap between 0 and the smallest non-zero eigenvalue.

(4) We derive an explicit formula of the temporary equilibrium achieved by the diffusion stage, which is the equilibrium observed in real word trainings: it is a linear combination of local Gibbs states $\mu_k$ in the attracting basins $U_k$. And the weight of each basin is approximately the same as the initial distribution of parameter among basins.

(5) Finally, we remark that our analysis also works in the setting of (Shi et al., 2020), where scaling invaraince is not assumed. Instead of the compact parametric space $\mathbb{S}^{d-1}$,

one assumes $\mathcal{L}$ is a loss function on $\mathbb{R}^d$ that grows fast as $|\boldsymbol{x}| \to \infty$. y (which is the case when the neural network has regularization such as weight decay). In fact, the main mathematical theories from (Simon, 1983; Freidlin & Wentzell, 2012), that our proof relies on, work in both the $\mathbb{S}^{d-1}$ and the fast growing functions in $\mathbb{R}^d$ settings.

### 1.4. Limitations

The main limitation of our study is the adoption of Assumptions 4.1 and Assumption 4.2. As mentioned above, the reason for these assumptions is their indispensability for the application of (Simon, 1983). On the one hand, the isotropic Gaussian is a popular assumption to use study SGD via SDE, and Morse functions are mathematically generic among $C^2$-functions. On the other hand, a serious limitation of these assumptions arises from the fact that most modern neural networks are overparametrized, which forces local minima to form connected regions, instead of being isolated like in the case of Morse functions, see (Garipov et al., 2018; Kuditipudi et al., 2019; Maddox et al., 2020; Benton et al., 2021; Cooper, 2021). The local dynamics near such regions has more recently been studied in (Li et al., 2021b).

Experimental evidences suggest that the temporary equilibria at constant effective learning rates should still be localized in the general setting. We hope the study of this phenomenon may shed light on better understanding the dynamics of SGD in schemes not covered by Simon's theory, and formulate it as Conjecture 6.3.

## 2. Preliminaries

### 2.1. SDE model

The SDE model of SGD has been extensively studied in recent years. Given a dataset $\mathcal{S} = \{\boldsymbol{x}_i\}_{i=1}^N$, at the $k$-th step of the SGD, a subset $\mathcal{B}_k \subset \mathcal{S}$ of fixed size $n$ is randomly drawn to train a neural network whose parameters are denoted by $\boldsymbol{w} \in \mathbb{R}^d$. For $\boldsymbol{x}$ from a mini-batch $\mathcal{B}$ and a parameter $\boldsymbol{w}$, the neural network outputs a loss function $\ell_{\mathcal{B}}(\boldsymbol{w}, \boldsymbol{x})$, which is assumed to be differentiable in $\boldsymbol{w}$. The loss function depends on $\mathcal{B}$ because of batch normalization steps inside the neural network. The loss functions over $\mathcal{B}$ is $\mathcal{L}_{\mathcal{B}}(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{x}\in\mathcal{B}}\ell_{\mathcal{B}}(\boldsymbol{w}, \boldsymbol{x})$. Also define the average loss function by $\mathcal{L}(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{x}\in\mathcal{S}}\ell_{\mathcal{B}}(\boldsymbol{w}, \boldsymbol{x}) = \mathbb{E}_{\substack{\mathcal{B}\subset\mathcal{S}\\|\mathcal{B}|=n}} \mathcal{L}_{\mathcal{B}}(\boldsymbol{w})$ (averaging all subsets $\mathcal{B}$ of size $n$ randomly drawn from $\mathcal{S}$). In the $k$-th step, the parameter $\boldsymbol{w}_k$ is updated by (1).

The function $\mathcal{L}$ now defines the gradient vector field $\nabla\mathcal{L}$ at every point $\boldsymbol{w} \in \mathbb{R}^d$. We also define a non-negative definite symmetric matrix $\boldsymbol{\Sigma}(\boldsymbol{w}) \in \mathrm{Mat}(d, d)$ at every $\boldsymbol{w} \in \mathbb{R}^d$ by

$$\boldsymbol{\Sigma}(\boldsymbol{w}) := \mathop{\mathbb{E}}_{\substack{\mathcal{B}\subset\mathcal{S}\\|\mathcal{B}|=n}} \Big( \big(\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w})\big) \big(\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w})\big)^{\top} \Big). \tag{3}$$

The matrix $\boldsymbol{\Sigma}(\boldsymbol{w})$ is called the gradient noise, and will take the role as a diffusion matrix.

The popular approach of approximating (1) by (2) has been studied in (Jastrzebski et al., 2017; Goyal et al., 2017; Smith & Le, 2018; Smith et al., 2018; Chaudhari & Soatto, 2018; Shi et al., 2020; Li et al., 2019, Li et al., 2020). For example, Li et al. (2019) proved that, for small $\eta$, starting from the same initial position the probability distributions of the random variables $\boldsymbol{W}_{k\eta}$ in (2) and $\boldsymbol{w}_k$ in (1) are close to each other for fixed $K$ and all $k \le \frac{K}{\eta}$. See also (Yaida, 2019; Smith et al., 2021; Li et al., 2021a) for discussions on the deficiencies of SDE view of SGD as well as conditions that guarantee validate this view.

Li et al. (2020, Theorem 5.1) proved that the SDE (2) is in turn equivalent to

$$\begin{aligned} \mathrm{d}\overline{\boldsymbol{W}}_t = &-\gamma_t^{-\frac{1}{2}} \Big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\Big) \\ &- \frac{1}{2}\gamma_t^{-1}\mathrm{Tr}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)\overline{\boldsymbol{W}}_t\mathrm{d}t \end{aligned}; \tag{4}$$

$$\frac{\mathrm{d}\gamma_t}{\mathrm{d}t} = -4\lambda_e\gamma_t + 2\mathrm{Tr}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t), \tag{5}$$

where $\overline{\boldsymbol{W}}_t = \frac{\boldsymbol{W}_t}{|\boldsymbol{W}_t|}$ and $\gamma_t = |\boldsymbol{W}_t|^4\eta^{-2}$. The quantity $\gamma_t^{-\frac{1}{2}}$ is called the *effective learning rate*. In particular, $\overline{\boldsymbol{W}}_t$ is a random process on the unit sphere $\mathbb{S}^{d-1}$. In addition, experimental results from (Li et al., 2020) suggest that both $\mathrm{Tr}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)$ and $\gamma_t$ stabilizes quickly near constant values, and the learning perfomance has little dependence on the initial parameter $\gamma_0$.

### 2.2. Normalization and scaling invariance within components

The neural network is assumed to be batch normalized, which guarantees that $\ell_{\mathcal{B}}(\boldsymbol{w}, \boldsymbol{x}) = \ell_{\mathcal{B}}(c\boldsymbol{w}, \boldsymbol{x})$ for all $c > 0$. In consequence (Li et al., 2020, Lemma B.1), $\boldsymbol{w}^{\top}\nabla\ell_{\mathcal{B}}(\boldsymbol{w}, \boldsymbol{x}) = 0$. The same orthogonality holds for $\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{w})$ and $\nabla\mathcal{L}(\boldsymbol{w})$, which are linear combinations of the $\nabla\ell_{\mathcal{B}}$'s. Thus,

$$\boldsymbol{w}^{\top}\nabla\mathcal{L}(\boldsymbol{w}) = 0, \boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{w} = 0. \tag{6}$$

For a vector $\boldsymbol{w}$, denote by $V_{\boldsymbol{w}}^{\perp}$ its $(d-1)$-dimensional orthogonal complement vector space. As $\boldsymbol{\Sigma}(\boldsymbol{w})$ is symmetric and has mutually orthogonal eigenspaces, $\boldsymbol{\Sigma}(\boldsymbol{w})$ preserves $V_{\boldsymbol{w}}^{\perp}$. Therefore $\boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}}$ annihilates $\boldsymbol{w}$ and preserves $V_{\boldsymbol{w}}^{\perp}$ as well. Recall $\boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}}$ is a uniquely defined positive semi-definite matrix (Horn & Johnson, 2013, Thm. 7.2.6).

## 3. The spherical model

By scaling invariance, the parameters $\boldsymbol{W}_t$ and $\overline{\boldsymbol{W}}_t$ yield the same outcome. So it suffices to understand the distribution of $\overline{\boldsymbol{W}}_t$ to consider Conjecture 1.1. We shall ignore (5) and focus on (4).

### 3.1. Description of model

Denote by $\overline{\nabla}$ the gradient on $\mathbb{S}^{d-1}$ with respect to the standard sphere metric . Then $\overline{\nabla}\mathcal{L}:\mathbb{S}^{d-1}\to T\mathbb{S}^{d-1}$ is a vector field on $\mathbb{S}^{d-1}$, whose value coincides with $\nabla\mathcal{L}$ by (6). For later use, we also write $\overline{\Delta}$ for the Laplacian on $\mathbb{S}^{d-1}$.

Write $\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})$ and $\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})^{\frac{1}{2}}$ for the restrictions of $\boldsymbol{\Sigma}(\boldsymbol{w})$ and $\boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}}$ to $\mathbb{S}^{d-1}$, viewed as tensor fields that send $\mathbb{R}^d$ to $T_{\boldsymbol{w}}\mathbb{S}^{d-1}=V_{\boldsymbol{w}}^{\perp}$ for $\boldsymbol{w}\in\mathbb{S}^{d-1}$. In particular, given the Brownian motion $\mathrm{d}\boldsymbol{B}_t^d$ on $\mathbb{R}^d$, $\overline{\boldsymbol{\Sigma}}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d$ is a random infinitesimal vector along the tangent space $T_{\boldsymbol{w}}\mathbb{S}^{d-1}$.

**Theorem 3.1.** *Starting from an initial value $\overline{\boldsymbol{W}}_0\in\mathbb{S}^{d-1}$, the SDE (4) is equivalent to the following SDE on $\mathbb{S}^{d-1}$:*

$$\mathrm{d}\overline{\boldsymbol{W}}_t=-\gamma_t^{-\frac{1}{2}}\left(\overline{\nabla}\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t+\overline{\boldsymbol{\Sigma}}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\right). \tag{7}$$

**Difference between the SDE's (4) and (7).** We now explain the meaning of Theorem 3.1, as it may at first glance seem unnatural to remove the last term $-\frac{1}{2}\gamma_t^{-1}\mathrm{Tr}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)\overline{\boldsymbol{W}}_t\mathrm{d}t$ from (4) without destroying the equality. The difference is as follows: (7) is a differential equation defined on the manifold $\mathbb{S}^{d-1}$ with respect to the intrinsic geometry of this manifold, where the terms $\overline{\nabla}\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t$ and $\overline{\boldsymbol{\Sigma}}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d$ are infinitesimal tangent vectors of $\mathbb{S}^{d-1}$. While the parameter $\overline{\boldsymbol{W}}_t$ flows along these vector fields according to (7), it by construction stays inside $\mathbb{S}^{d-1}$. Indeed, in the construction the diffusion process (7), the differential geometry of $\mathbb{S}^{d-1}$ is used (Hsu, 2002, §1.3) in addition to the values of vector fields. But (4) is a differential equation defined on the larger ambient manifold $\mathbb{R}^d$, and $\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t$ and $\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d$ are only vector fields on $\mathbb{R}^d$. A priori, (4) moves $\overline{\boldsymbol{W}}_t$ around in the entire $\mathbb{R}^d$ but not necessarily within $\mathbb{S}^{d-1}$. However, it turns out that given an initial position $\overline{\boldsymbol{W}}_0\in\mathbb{S}^{d-1}$, almost all random trajectories of (4) must remain in $\mathbb{S}^{d-1}$ since (4) arises from (2) with $\overline{\boldsymbol{W}}_t=\frac{\boldsymbol{W}_t}{|\boldsymbol{W}_t|}$. Theorem 3.1 then asserts the probability distributions of $\overline{\boldsymbol{W}}_t$ are the same for (4) and for (7).

In most of this paper, $\boldsymbol{\Sigma}(\boldsymbol{w})|_{V_{\boldsymbol{w}}^{\perp}}$ will be assumed to be the constant matrix $\sigma^2\mathrm{Id}|_{V_{\boldsymbol{w}}^{\perp}}$ for $\boldsymbol{w}\in\mathbb{S}^{d-1}$. In this case, (Li et al., 2020, Lemma 5.2) proved $\gamma_t=\gamma+(\gamma_0-\gamma)e^{-4\lambda_e t}$ converges exponentially fast to $\gamma=\frac{\sigma^2}{2\lambda_e}$. In light of this, we will focus on the SDE model assuming $\gamma_t$ is the constant $\gamma$:

**Definition 3.2.** The sphericial model of SGD with constant effective learning rate $\zeta:=\gamma^{-\frac{1}{2}}=\frac{\sqrt{2\lambda_e}}{\sigma}$ is the following SDE on $\mathbb{S}^{d-1}$:

$$\mathrm{d}\overline{\boldsymbol{W}}_t=-\zeta\left(\overline{\nabla}\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t+\overline{\boldsymbol{\Sigma}}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\right). \tag{8}$$

We will denote by $\mathcal{P}_{\overline{\boldsymbol{W}}_0=\boldsymbol{w}}(\overline{\boldsymbol{W}}_t)$ and $\mathcal{P}_{\overline{\boldsymbol{W}}_0\sim\nu}(\overline{\boldsymbol{W}}_t)$ the probability distributions of the random solution $\overline{\boldsymbol{W}}_t$ to the equation (8), respectively under initial conditions $\overline{\boldsymbol{W}}_0=\boldsymbol{w}$ and $\overline{\boldsymbol{W}}_0\sim\nu$, where $\boldsymbol{w}\in\mathbb{S}^{d-1}$ and $\nu$ is a probability measure on $\mathbb{S}^{d-1}$.

### 3.2. Relation to Riemannian implement of BN

Besides simulating the stablizing behavior of $\gamma_t$, the SDE (8) is also a model to the Riemannian approach to batch normalization introduced by (Cho & Lee, 2017). This approach aims to eliminate the ambiguity in scaling by perform SGD on the Riemannian manifold $\mathbb{S}^{d-1}$.

The basic algorithm from (Cho & Lee, 2017) can be simply stated as follows: the sequence of parameters $\boldsymbol{w}_k$ will always remain in $\mathbb{S}^{d-1}$ and be updated by

$$\boldsymbol{w}_{k+1}\leftarrow\exp_{\boldsymbol{w}_k}(-\zeta\nabla\mathcal{L}_{\mathcal{B}_k}(\boldsymbol{w}_k)), \tag{9}$$

where $\zeta$ is the learning rate and $\exp_{\boldsymbol{w}}(\boldsymbol{v}):=\boldsymbol{w}\cos|\boldsymbol{v}|+\boldsymbol{v}\frac{\sin|\boldsymbol{v}|}{|\boldsymbol{v}|}$ is the exponential map on $\mathbb{S}^{d-1}$ for $\boldsymbol{w}\in\mathbb{S}^{d-1}$ and $\boldsymbol{v}\in T_{\boldsymbol{w}}\mathbb{S}^{d-1}$. Note that $\nabla\mathcal{L}_{\mathcal{B}_k}(\boldsymbol{w}_k))\in T_{\boldsymbol{w}}\mathbb{S}^{d-1}$ as $\boldsymbol{w}^{\top}\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{w},\boldsymbol{x})=0$. An alternative to (9) is

$$\boldsymbol{w}_{k+1}\leftarrow\frac{\boldsymbol{w}_k-\zeta\nabla\mathcal{L}_{\mathcal{B}_k}(\boldsymbol{w}_k)}{|\boldsymbol{w}_k-\zeta\nabla\mathcal{L}_{\mathcal{B}_k}(\boldsymbol{w}_k)|}. \tag{10}$$

The updating methods (9) and (10) differ only by $O(\zeta^2)$ and hence are asymptotic to each other for small $\zeta$, as demonstrated by Theorem 3.3 below. However (10) has the advantage that it is computationally cheaper and the trajectories are guaranteed to stay in $\mathbb{S}^{d-1}$ despite of numerical errors.

**Theorem 3.3.** *If the vector $\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{w})-\nabla\mathcal{L}(\boldsymbol{w})$ is considered as a Gaussian noise $\boldsymbol{\xi}(\boldsymbol{w})$ with covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{w})$ defined by (3), then (8) is the continuous time limit of both (9) and (10).*

## 4. Three-stage evolution

We will from now on assume following conditions like in (Bovier et al., 2004; Shi et al., 2020) :

**Assumption 4.1.** *(Standard noise) On $\mathbb{S}^{d-1}$, the covariance matrix $\boldsymbol{\Sigma}$ coincides with a constant multiple of the Riemannian metric on $\mathbb{S}^{d-1}$. More precisely, there exists $\sigma>0$ such that for all $\boldsymbol{w}\in\mathbb{S}^{d-1}$ and $\boldsymbol{v}\in V_{\boldsymbol{w}}^{\perp}=T_{\boldsymbol{w}}\mathbb{S}^{d-1}$, $\boldsymbol{v}^{\top}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{v}=\sigma^2|\boldsymbol{v}|^2$.*

**Assumption 4.2.** *(Morse loss function) The restriction of $\mathcal{L}$ to $\mathbb{S}^{d-1}$ is a function in $C^2(\mathbb{S}^{d-1})$. Furthermore, it is a Morse function: every critical point $\boldsymbol{z}$ is non-degenerate, i.e. $\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z})\neq 0$.*

Note that the Morse function condition is generically true in the class $C^2(\mathbb{S}^{d-1})$.

### 4.1. Fokker-Planck equation and Gibbs density

Because the diffusion matrix $\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})^{\frac{1}{2}}(\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})^{\frac{1}{2}})^{\top} = \overline{\boldsymbol{\Sigma}}(\boldsymbol{w})$ coincides with $\sigma^2 \mathrm{Id}|_{T_{\boldsymbol{w}}\mathbb{S}^{d-1}}$ on $T_{\boldsymbol{w}}\mathbb{S}^{d-1}$, the marginal distribution of the stochastic process (8) is absolutely continuous on $\mathbb{S}^{d-1}$ for $t > 0$ and its density function $u_t(\boldsymbol{w})$ is a solution for $t \in (0, \infty)$ and $\boldsymbol{w} \in \mathbb{S}^{d-1}$ to the following Fokker-Planck equation by standard arguments on diffusion process (see §B for a proof).

**Proposition 4.3.** *The density function $u : (0, \infty) \times \mathbb{S}^{d-1} \to \mathbb{R}$ of the distribution of the stochastic process (8) satisfies*

$$\partial_t u = \zeta \overline{\nabla} \cdot (u \overline{\nabla} \mathcal{L}) + \frac{1}{2}\zeta^2 \sigma^2 \overline{\Delta} u. \tag{11}$$

Define

$$\beta := \frac{1}{2}\zeta\sigma^2 = \sqrt{\frac{\lambda_e}{2}}\sigma. \tag{12}$$

It is known that the Gibbs density

$$\mu^{(\beta)}(\boldsymbol{w}) := \frac{e^{-\frac{\mathcal{L}(\boldsymbol{w})}{\beta}}}{\int_{\mathbb{S}^{d-1}} e^{-\frac{\mathcal{L}}{\beta}} \mathbf{dm}} \tag{13}$$

is a stationary solution to (11).

### 4.2. Attracting basins

In view of Assumption 4.2, denote the set of critical points of the restriction of $\mathcal{L}$ on $\mathbb{S}^{d-1}$ by $Z \subset \mathbb{S}^{d-1}$, and the local minima among them by $\boldsymbol{z}_1, ..., \boldsymbol{z}_m \in Z$. Because the critical points are non-degenerate, they must be isolated. Hence $Z$ is finite.

**Definition 4.4.** The **attracting basin** $U_{\boldsymbol{z}}$ of a critical point $\boldsymbol{z} \in Z$ is the set of all points $\boldsymbol{w}_0 \in \mathbb{S}^{d-1}$ such that the unique solution $\overline{\boldsymbol{W}}_t$ to the ODE

$$\overline{\boldsymbol{W}}_t = -\zeta \overline{\nabla} \mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t \tag{14}$$

defined on $\mathbb{S}^{d-1}$, subject to the initial condition $\overline{\boldsymbol{W}}_0 = \boldsymbol{w}_0$, converges to $\boldsymbol{z}$ as $t \to \infty$. For simplicity, when $\boldsymbol{z} = \boldsymbol{z}_i$ is a local minima, we will write $U_i := U_{\boldsymbol{z}_i}$.

For $Q > 0$, let $U_{i,Q} := \{\boldsymbol{w} \in U_i : \mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{z}_i) < Q\}$, which is a neighborhood of $\boldsymbol{z}_i$ in $U_i$.

### 4.3. Main theorems

*In the analysis below, fix a parameter $\epsilon > 0$ as error tolerance, and let $\lambda_e$ be sufficiently small. All choices of parameters, as well as implicit constants in $O(\cdot)$ notations, are supposed to be dependent of the loss function $\mathcal{L}$.*

The dynamics of (8) consists of three different stages: **descent**, **diffusion** and **tunneling**.

**Stage 1: Descent.** In this stage, the trajectory, with probability close to 1, takes $O(\lambda_e^{-\frac{1}{2}})$ time, to descend to $U_{i,Q_1}$ in each $U_i$.

**Theorem 4.5.** *Under Assumptions 4.1 and 4.2, for all $\epsilon > 0$ and $Q_1 > 0$, there exist $C_{\mathrm{des}} > 0, \lambda_{\mathrm{des}} > 0$ and a set $\Lambda_\epsilon$ of volume $\mathbf{m}(\Lambda_\epsilon) > 1 - \epsilon$, such that for all $\lambda_e < \lambda_{\mathrm{des}}$, and all $\boldsymbol{w}_0 \in \Lambda_\epsilon$, the random solution $\overline{\boldsymbol{W}}_t$ to (8) starting at $\boldsymbol{w}_0$ satisfies $\mathbb{P}_{\overline{\boldsymbol{W}}_0 = \boldsymbol{w}_0}\left(\overline{\boldsymbol{W}}_{C_{\mathrm{des}}\lambda_e^{-\frac{1}{2}}} \in U_{k,Q_1}\right) > 1 - \epsilon$, where $U_k$ is the unique attracting basin that contains $\boldsymbol{w}_0$.*

**Stage 2: Diffusion.** The diffusion stage takes at most $O(\lambda_e^{-1})$ time in terms of $t$. During this period, the distribution converges to a **temporary equilbrium** given by the conditional measure of the Gibbs equilibrium inside each basin and remains stable for exponentially long time after that. The weights assigned to basins correspond to the initial distribution of mass among them. The next theorem is the central one in this paper.

**Theorem 4.6.** *Under Assumptions 4.1 and 4.2, for all $\epsilon > 0$, there exist constants $C_{\mathrm{dif}}, c_{\mathrm{dif}}, \lambda_{\mathrm{dif}} > 0$, and a set $\Lambda_\epsilon$ of volume $\mathbf{m}(\Lambda_\epsilon) > 1 - \epsilon$, such that:*

*With $\beta = \sqrt{\frac{\lambda_e}{2}}\sigma$, for all $\lambda_e < \lambda_{\mathrm{dif}}$, the random solution $\overline{\boldsymbol{W}}_t$ to (8) satisfies: for all $t \in [\frac{C_{\mathrm{dif}}}{\lambda_e}, e^{\frac{c_{\mathrm{dif}}}{\sqrt{\lambda_e}}}]$:*

*(i) For all initial positions $\boldsymbol{w}_0 \in \Lambda_\epsilon$,*

$$\mathrm{dist}_{\mathrm{TV}}\left(\mathcal{P}_{\overline{\boldsymbol{W}}_0 = \boldsymbol{w}_0}(\overline{\boldsymbol{W}}_t), \frac{\mu^{(\beta)}|_{U_k}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}\right) \le \epsilon,$$

*where $U_k$ is the unique attracting basin of the gradient flow of $\mathcal{L}$ that contains $\boldsymbol{w}_0$.*

*(ii) For all initial probability distribution $\nu_0$,*

$$\mathrm{dist}_{\mathrm{TV}}\left(\mathcal{P}_{\overline{\boldsymbol{W}}_0 \sim \nu_0}(\overline{\boldsymbol{W}}_t), \sum_{i=1}^m \nu_0(U_i)\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}\right)$$
$$\le \epsilon + \nu_0(\Lambda_\epsilon^c).$$

Here $\mathrm{dist}_{\mathrm{TV}}$ is the total variation distance $\mathrm{dist}_{\mathrm{TV}}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$ between measures, and $\Lambda_\epsilon^c$ is the complement of $\Lambda_\epsilon$.

Theorem 4.6.(ii) identifies a temporary equilibrium $\sum_{i=1}^m \nu_0(U_i)\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}$. The corollary below claims that in the case of non-convex optimization, this temporary equilibrium is often different from the eventual Gibbs equilibrium $\mu^{(\beta)}\mathbf{dm}$.

**Corollary 4.7.** *In the setting of Theorem 4.6, if in addition $m \ge 2$, then there exists a constant $\kappa_{\mathrm{dif}} > 0$ and a subset $\Omega_{\mathrm{dif}}$ with $\mathbf{m}(\Omega_{\mathrm{dif}}) > \kappa_{\mathrm{dif}}$, such that for all sufficiently small*

*$\lambda_e$, $\boldsymbol{w}_0 \in \Omega_{\rm dif}$ and $t \in [\frac{C_{\rm dif}}{\lambda_e}, e^{\frac{c_{\rm dif}}{\sqrt{\lambda_e}}}]$, the random solution $\overline{\boldsymbol{W}}_t$ to (8) satisfies*

$$\mathrm{dist}_{\rm TV}\Big(\mathcal{P}_{\overline{\boldsymbol{W}}_0=\boldsymbol{w}_0}(\overline{\boldsymbol{W}}_t), \mu^{(\beta)}\mathbf{dm}\Big) > \kappa_{\rm dif}.$$

*Moreover, if $\mathcal{L}$ attains its global minimum value on $\mathbb{S}^{d-1}$ at more than one local minima, then given any small $\epsilon > 0$, one can replace $\mathbf{m}(\Omega_{\rm dif}) > \kappa_{\rm dif}$ by $\mathbf{m}(\Omega_{\rm dif}) > 1-\epsilon$ for sufficiently small $\lambda_e$.*

The exceptional set $\Lambda_\epsilon^c$ in Theorems 4.5 and 4.6 is a small neighborhood of the set of points whose deterministic gradient descent trajectory converge to a stationary point of $\mathcal{L}$ that is not a local minimum. This later set is a submanifold of strictly lower dimension under Assumption 4.2, and should be viewed as the boundary between attracting basins of local minima. A careful analysis of the arguments in Appendix §D would control the radius of the neighbordhood in terms of $\epsilon$, the values of $\mathcal{L}$ and $\overline{\nabla}^2\mathcal{L}$ at the stationary points of $\mathcal{L}$ and the Lipschtiz norm of $\mathcal{L}$.

**Stage 3: Tunneling.** The final stage takes $O(e^{\frac{C}{\sqrt{\lambda_e}}})$ time. During this stage, mass leaks slowly between basins, and eventually equidistributes towards the Gibbs equilibrium $\mu^{(\beta)}\mathbf{dm}$. Because of the slow rate, this stage is not expected to be observed within typical training time.

The name "tunneling", following previous works, e.g. (Helffer & Sjöstrand, 1985; Hérau et al., 2011), comes from the quantum tunnel effect in solutions to the Schrödinger equation, which is related to our model through the Schrödinger operator $\mathcal{D}^{\beta,\#}f = \beta\overline{\Delta}f - (\frac{|\overline{\nabla}\mathcal{L}|^2}{4\beta} - \frac{\overline{\Delta}\mathcal{L}}{2})f$ (see §E.1).

**Theorem 4.8.** *Under Assumptions 4.1 and 4.2, and with $\beta = \sqrt{\frac{\lambda_e}{2}}\sigma$, there exist constants $C_{\rm tun}, \lambda_{\rm tun} > 0$ such that for all $\boldsymbol{w}_0 \in \mathbb{S}^{d-1}$, $\lambda_e < \lambda_{\rm tun}$ and $t \geq 0$, the random solution $\overline{\boldsymbol{W}}_t$ to (8) satisfies :*

$$\mathrm{dist}_{\rm TV}\Big(\mathcal{P}_{\overline{\boldsymbol{W}}_0=\boldsymbol{w}_0}(\overline{\boldsymbol{W}}_t), \mu^{(\beta)}\mathbf{dm}\Big) \leq O(e^{-e^{-\frac{C_{\rm tun}}{\sqrt{\lambda_e}}}t}).$$

We don't claim originality for Theorem 4.8 and only include it for a complete description of the stages. It was proved by (Shi et al., 2020) for fast growing loss functions $\mathcal{L}$ on $\mathbb{R}^d$, and their argument also works here (see §F).

Instead, our main contribution is Theorem 4.6, which identifies the temporary equilibrium and proves the time needed to reach it is $\leq O(\lambda_e^{-1})$, as well as Corollary 4.7, which asserts that typically the temporary and eventual equilibria are different, and the time needed to deviate from the temporary equilibrium towards the eventual one is at least $O(e^{\frac{c_{\rm dif}}{\sqrt{\lambda_e}}})$, and thus usually beyond practical observable windows. The contrast between Corollary 4.7 and Theorem 4.8 establishes the separation between the diffusion and tunneling stages.

## 5. Experiments

**Model with our assumptions**. We implemented the SDE (8) on $\mathbb{S}^{d-1}$ with $d = 100$, $\sigma = 0.1$, and a scaling-invariant loss function $\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{200}\sin(\boldsymbol{a}_i \cdot \frac{\boldsymbol{w}}{|\boldsymbol{w}|})^2$ where $\boldsymbol{a}_i$ are randomly chosen. For each $\zeta$, we ran 32 independent random instances of the discrete implementation

$$\boldsymbol{w}_{k+1} \leftarrow \frac{\boldsymbol{w}_k - \zeta\big(\nabla\mathcal{L}(\boldsymbol{w}_k) + \mathcal{N}_d(0,\sigma)\big)}{\big|\boldsymbol{w}_k - \zeta\big(\nabla\mathcal{L}(\boldsymbol{w}_k) + \mathcal{N}_d(0,\sigma)\big)\big|} \tag{15}$$

of (8). The same proof of Theorem 3.3 shows (8) is a continuous time limit of (15). All instances start at the same initial position and last $8 \times 10^5$ iteration steps.

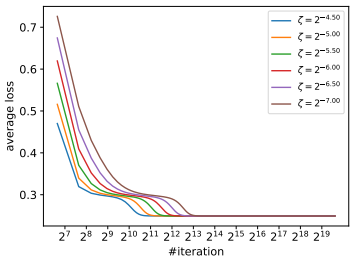


*Figure 1.* Train loss [Toy function experiment]

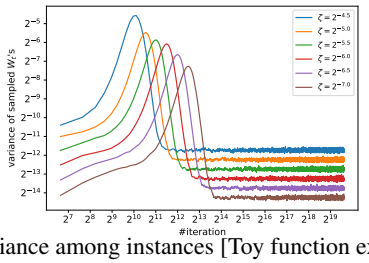


*Figure 2.* Variance among instances [Toy function experiment]

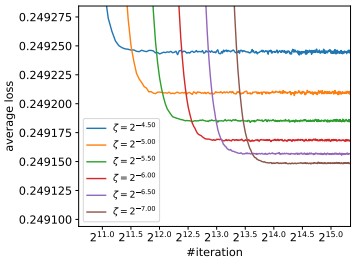


*Figure 3.* Train loss, zoom in view [Toy function experiment]

We choose two features to indicate the achievement of an equilibrium: the average loss among the instances, as well as the variance of the weights $\boldsymbol{W}_t$ of all the instances. Figures 1-4 show that when $\zeta$ is divided by $2^{0.5}$, the step at which these features become approximately constant is roughly multiplied by $2^{0.5}$. That is, equilibrium is observed in $O(\zeta^{-1})$ time. As $\zeta = \frac{\sqrt{2\lambda_e}}{\sigma}$, $O(\zeta^{-1}) = O(\lambda_e^{-\frac{1}{2}}) < O(\lambda_e^{-1})$, this is consistent with Theorem 4.6 and suggests there is room for further improvement (Question 6.2 below).

Experiments also support that the observed equilibrium is the one in Theorem 4.6. First of all, the observed equilib-

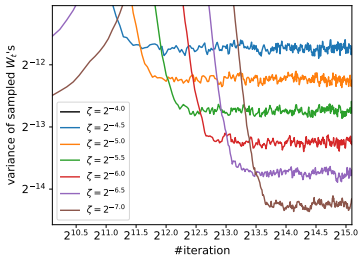


*Figure 4.* Variance, zoom in view [Toy function experiment]

rium is localized near a single local minimum point, as the variance among instances become very close to 0.

In addition, note $\mu_k^{(\beta)} := \frac{\mu^{(\beta)}|_{U_k}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}$ is concentrated near $\boldsymbol{z}_k$ for small $\beta$. As $\boldsymbol{z}_k$ is a non-degenerate local minimum, in suitable local linear coordinates $\boldsymbol{y} \in T_{\boldsymbol{z}_k}\mathbb{S}^{d-1}$ for $\boldsymbol{w}$ near $\boldsymbol{z}_k$, $\mathcal{L}(\boldsymbol{w}) = \mathcal{L}(\boldsymbol{z}_k) + \frac{1}{2}|\boldsymbol{y}|^2 + O(|\boldsymbol{y}|^3)$. Then $\mu_k^{(\beta)}$ is approximated by $\frac{e^{-\frac{1}{2\beta}|\boldsymbol{y}|^2}}{\text{constant}}\mathrm{d}\boldsymbol{y}$, which is the normal distribution $\mathcal{N}(0, \beta\mathrm{Id})$ in terms of $\boldsymbol{y}$. Since $\boldsymbol{w}$ is approximately an affine function of $\boldsymbol{y}$, $\mathrm{Var}_{\mu_k^{(\beta)}}(\boldsymbol{w})$ should be approximately proportional to $\beta = \frac{1}{2}\zeta\sigma^2$, and thus to $\zeta$ as well. In addition, $\mathbb{E}_{\mu_k^{(\beta)}}\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{z}_k) \approx \frac{1}{2}\mathbb{E}_{\mu_k^{(\beta)}}|\boldsymbol{y}|^2 = \frac{1}{2}\mathrm{Var}(\boldsymbol{y}) \approx \frac{1}{2}\mathrm{Tr}\beta\mathrm{Id} = \frac{d-1}{2}\beta$ is also proportional to both $\beta$ and $\zeta$.

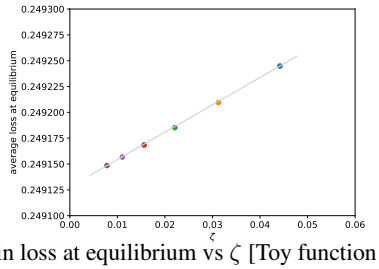


*Figure 5.* Train loss at equilibrium vs $\zeta$ [Toy function experiment]

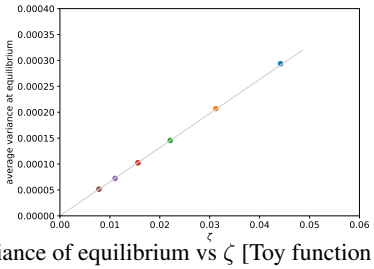


*Figure 6.* Variance of equilibrium vs $\zeta$ [Toy function experiment]

Figures 5 and 6 show that near the observed equilibrium, the training loss $\mathbb{E}\mathcal{L}(\boldsymbol{W}_t)$ is approximately linear in $\zeta$ and $\mathrm{Var}(\boldsymbol{W}_t)$ is approximately proportional to $\zeta$, which perfectly matches the predictions above.

**Underparametrized neural network.** We apply (10) to a tiny 4-layer CNN network of the same structure as above, but with on the MNIST dataset of only $d = 332$ parameters. There are two convolution layer followed by two linear layers. All layers are batch normalized to ensure scale-invariance. The number of parameters is chosen to be unpractically low for two reasons: (1) the underparametrization of the network guarantees that Assumption 4.2 still holds and local minima are isolated; (2) to allow faster training, as we need to run many independent instances and observe their distribution. However, Assumption 4.1 no longer holds in this case.

Due to the small size of the network, the achieved minimum of the cross entropy loss function is much worse than those in practical trainings. But our goal here is to empirically test the likelihood of whether our results extend to settings satisfying Assumption 4.2 but not Assumption 4.1.

For each learning rate $\zeta$, 16 independent random instances of (10) with the same initial position are performed, for 20,000 epochs. Each epoch has 54 iteration steps with a batch size of 1000.

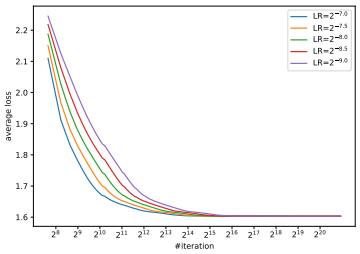


*Figure 7.* Train loss [Underparametrized NN]

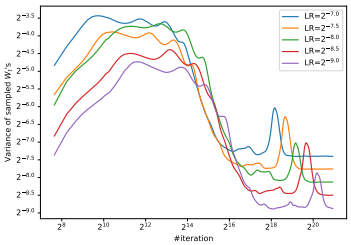


*Figure 8.* Variance among instances [Underparametrized NN]

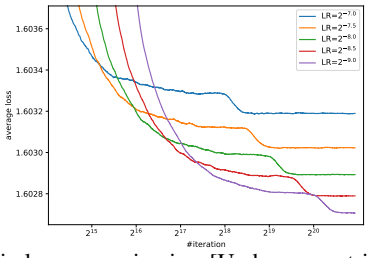


*Figure 9.* Train loss, zoom in view [Underparametrized NN]

As before, we use the average loss as well as the variance among parameters as indicators for reaching equilibrium. Figures 7-10 [1] again show that random trajectories with the same initial position stabilize to a local equilibrium

[1]Curves in Figures 7-10, as well as in Figure 14 later, are logarithmically smoothified by displaying at $n$ the average values from epoch $0.9n$ to epoch $n$.

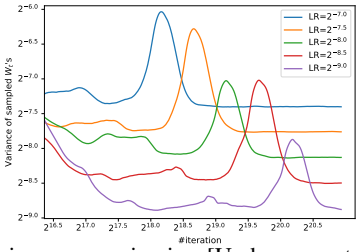


*Figure 10.* Variance, zoom in view [Underparametrized NN]

within $O(\zeta^{-1}) = O(\lambda_e^{-\frac{1}{2}})$ time. This local minimum is again near a single critical point as $\mathrm{Var}_{\mu_k^{(\beta)}}(\boldsymbol{w})$ is extremely small. As $\zeta$ decreases, the loss and variance decay in similar patterns as in the previous model. However, Figure 12 shows $\mathrm{Var}_{\mu_k^{(\beta)}}(\boldsymbol{w})$ is not exactly proportional to $\alpha$, but decreases to 0 faster than $\eta$, which suggests that with Assumption 4.2 but not Assumption 4.1, the local equilibrium in general doesn't take the form $\frac{\mu^{(\beta)}|_{U_k}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}$ than in the standard noise setting. See Conjecture 6.3 below.

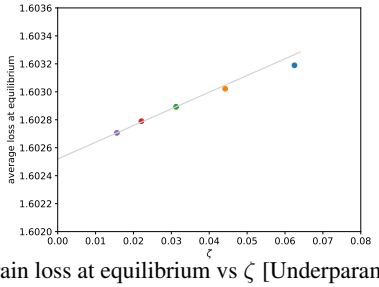


*Figure 11.* Train loss at equilibrium vs $\zeta$ [Underparametrized NN]

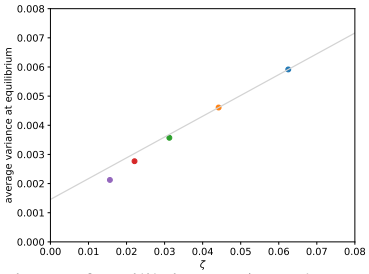


*Figure 12.* Variance of equilibrium vs $\zeta$ [Underparametrized NN]

**Overparametrized neural network.** Finally we apply (10) to a 4-layer CNN network of the same structure as above, but with $d \approx 100k$ parameters instead, on the MNIST. This general case is very different from the setting we studied as it is highly overparametrized and the local minima are likely submanifolds of large dimensions and sizes. Neither Assumptions 4.1 nor 4.2 is expected to hold. In particular, we no longer expect to see equilibria concentrated around single critical points.

In this setting it takes a huge number ($> 10k$) of epochs before statistics stabilize, and each epoch runs for longer time. Hence it was infeasible within our training budget to carry out similar experiments with many independent instances across several learning rates like in the two previous cases. However, in this case we want to know: Starting from a given initial position, is the fast equilibrium for (7) localized inside a single basin, and different from the global Gibbs equilibrium, like in Theorem 4.6? In other words, does the separation between the diffusion and tunneling stages extend to the overparametrized case?

Our experiment suggests that the answer is yes. At small effective learning rate, within observable time windows different initial positions lead to different equilibria that are far from each other. Since they cannot both be the unique Gibbs equilibrium, these equilibria are both local and temporary in nature. Starting from each of two randomly chosen initial positions $\boldsymbol{w}_0^1, \boldsymbol{w}_0^2 \in \mathbb{S}^{d-1}$, we ran 5 instances of the implementation (10) of (7) for 40000 epochs ($\approx 2M$ iterations) at $\zeta = 2^{-4.5}$. All instances use independent random seeds. For $1 \le i \le 5$ and $p = 1, 2$, let $\boldsymbol{w}_k^{i,p} \in \mathbb{S}^{d-1}$ denote the parameter at $k$-th step of the $i$-th instance starting at $\boldsymbol{w}_0^p$. We tracked the average distance squares between pairs of parameters that originated from the same initial position, as well as from different initial positions, or more precisely the quantities $\mathcal{V}_{11}(k) := \mathbb{E}_{i\neq j}|\boldsymbol{w}_k^{i,1} - \boldsymbol{w}_k^{j,1}|^2$; $\mathcal{V}_{22}(k) := \mathbb{E}_{i\neq j}|\boldsymbol{w}_k^{i,2} - \boldsymbol{w}_k^{j,2}|^2$; $\mathcal{V}_{12}(k) := \mathbb{E}_{i,j}|\boldsymbol{w}_k^{i,1} - \boldsymbol{w}_k^{j,2}|^2$. where the parameters are regarded as vectors in $\mathbb{R}^d$.

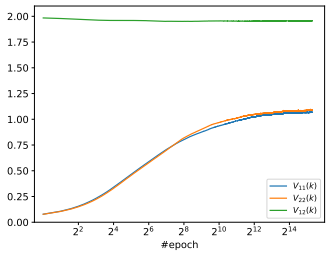


*Figure 13.* Different temporary equilibria [Overparametrized NN]

Figure 13 shows: (1) The equilibrium derived from an initial position is not concentrated near a single point, as $\mathcal{V}_{11}$ and $\mathcal{V}_{22}$ do not approach 0; (2) The two equilibria derived from $\boldsymbol{w}_0^1$ and $\boldsymbol{w}_0^2$ are far from each other, because $\mathcal{V}_{12}$ is much larger than $\mathcal{V}_{11}$ and $\mathcal{V}_{22}$. Indeed, $\mathcal{V}_{12} \approx 2$ for all time. Since the distance between two uniformly chosen random points on $\mathbb{S}^{d-1}$ is $\approx \sqrt{2}$ when $d$ is large, this suggests the two equilibria are independently located.

## 6. Discussions

### 6.1. Interpretations of initial parameter independence in Conjecture 1.1

Our Theorem 4.6 shows that, in the parameter space the equilibrium reached in $O(\frac{1}{\lambda_e})$ time is mostly concentrated in the attracting basin of the initial parameter. While this

is supported by real world neural network implementations (Figure 13), one needs to explain such deviation from the independence on the initial parameter in Conjecture 1.1 of (Li et al., 2020). Two possible interpretations are sketched below.

**Initial mass distribution.** Consider SGD with batch normalization starting from an inital distribution $\nu_0$ of parameters and let $\nu_t$ be the distribution at moment $t$. Theorem 4.6 shows that equilibrium can be reached in $t_1 \le O(\frac{1}{\lambda_e})$ steps after $\gamma_t^{-\frac{1}{2}}$ stabilizes. Moreover, the equilibrium is a linear combination $\sum_k p_k \mu_k^{(\beta)}$ of local equilibria $\mu_k^{(\beta)}$ in basins $U_k$. The allocation of weights $p_k$ is approximately the same as the initial allocation for this stage, or in other words the allocation at the end of the preparatory phase when $\gamma_t$ stabilizes. (Li et al., 2020) suggests this prepartory phase takes $t_0 \approx O(\frac{1}{\lambda_e})$ steps, during which $\gamma_t$ stabilizes exponentially fast (at least when $\mathrm{Tr}\boldsymbol{\Sigma}$ is constant). Experiments in (Li et al., 2020, Figure 5) also show that the initial effective learning rate $\gamma_0^{-\frac{1}{2}}$ is typically much larger than the limit value at $t_0$.

When $t$ is small, the effective learning rate $\gamma_t^{-\frac{1}{2}}$ is large for some time. Though this period is short, the large learning rate allows mixing in short time and reaches a global equilibrium (which we think of as $\mu^{(\beta_0)}$ with some very large $\beta_0$). This process fixes the allocation of mass $p_k$ among basins $U_k$. Once the effective learning rate becomes small, mechanisms similar to Theorem 4.6 prevent mass from leaking between basins. Finally, after $\gamma_t^{-\frac{1}{2}}$ stabilizes, the distribution is locally fine tuned to local Gibbs equilibria $\mu_k^{(\beta)}$ with small $\beta$ without changing the weights $p_k$.

In summary, in the fast equilibrium $\sum_k p_k \mu_k^{(\beta)}$, the components $\mu_k^{(\beta)}$'s are determined later at small effective learning rates, but the weights $p_k$'s are determined earlier at large effective learning rates.

**Similar landscapes among local minima.** We are grateful to an anonymous reviewer for suggesting to us that the use of an output function such as testing/training loss might be responsable for the independence in the initial parameter in Conjecture 1.1. On the one hand, for an arbitrary loss function $\mathcal{L}$, this is likely not the case because $\mathcal{L}$ may have two attracting basins with different local minimum values $\mathcal{L}(\boldsymbol{x}_1)$, $\mathcal{L}(\boldsymbol{x}_2)$. Different initial parameters in these basins will output the corresponding training losses. On the other hand, further experiments show that in realistic neural network implementations the use of an output function such as the loss $\mathcal{L}$ might explain the initial parameter independence. For instance, for the same initial parameters $\boldsymbol{w}_0^1$, $\boldsymbol{w}_0^2$ in Figure 13, though their stochastic trajectories converge in the diffusions stage to different temporary equilibria in the parameter space, Figure 14 shows that the loss function $\mathcal{L}$ has similar distribution over these temporary equilibria. More precisely, the quantities $\mathcal{VL}_{11}(k) := \mathbb{E}_{i\neq j}|\mathcal{L}(\boldsymbol{w}_k^{i,1}) - \mathcal{L}(\boldsymbol{w}_k^{j,1})|^2$, $\mathcal{VL}_{22}(k) := \mathbb{E}_{i\neq j}|\mathcal{L}(\boldsymbol{w}_k^{i,2}) - \mathcal{L}(\boldsymbol{w}_k^{j,2})|^2$ and $\mathcal{VL}_{12}(k) := \mathbb{E}_{i,j}|\mathcal{L}(\boldsymbol{w}_k^{i,1}) - \mathcal{L}(\boldsymbol{w}_k^{j,2})|^2$ are all distributed near 0 with the same pattern. This suggests the following possible interpretation of initial parameter independence: in an overparametrized scaling-invariant neural network, a majority portion of the parametric space $\mathbb{S}^{d-1}$ might be covered by attracting basins on which the landscapes of $\mathcal{L}$ are approximately the same.

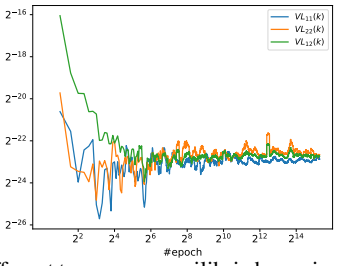


*Figure 14.* Different temporary equilibria have similar distributions of loss function [Overparametrized NN]

### 6.2. Open questions and future directions

The following questions might be of interest.

**Question 6.1.** *Analyze either one or both of the interpretations in §6.1.*

Though Theorem 4.6 shows the distribution stabilizes near temporary equilibria within $O(\frac{1}{\lambda_e})$ time, in our experiments, most features only require $O(\frac{1}{\sqrt{\lambda_e}})$ time to stabilize with constant $\gamma_t^{-\frac{1}{2}}$. Note that this wouldn't affect Conjecture 1.1 as the convergence of $\gamma_t^{-\frac{1}{2}}$ itself needs $O(\frac{1}{\lambda_e})$ time.

**Question 6.2.** *With Assumptions 4.1 and 4.2, can the bound $t \ge O(\frac{1}{\lambda_e})$ in Theorem 4.6 be improved to $t \ge O(\frac{1}{\sqrt{\lambda_e}})$?*

In light of the experiments in §5, we conjecture that

**Conjecture 6.3.** Without Assumptions 4.1 and 4.2, around each connected sets $Z_k \subset \mathbb{S}^{d-1}$ of local minima of $\mathcal{L}$, there are probability measures $\mu_k^{(\beta)}$ supported on attracting basins, such that $\mathbb{E}_{\boldsymbol{w}\sim\mu_k^{(\beta)}}\mathrm{dist}(\boldsymbol{w}, Z_k) \to 0$ as $\beta \to 0$; and Theorem 4.6 and Corollary 4.7 hold after replacing $\frac{\mu^{(\beta)}|_{U_k}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}$ with $\mu_k^{(\beta)}$.

## Acknowledgments

Y. W.'s research and the use of GPU units were supported by NSF grant DMS-1845033. Z.W.'s research was supported by NSF grant DMS-1753042. We thank the anonymous reviewers for their insightful comments and helpful suggestions.

## References


Benton, G., Maddox, W., Lotfi, S., and Wilson, A. G. G. Loss surface simplexes for mode connecting volumes and fast ensembling. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 769–779. PMLR, 18–24 Jul 2021.

Bovier, A., Eckhoff, M., Gayrard, V., and Klein, M. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 006(4):399–424, 2004.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations*, 2018.

Cho, M. and Lee, J. Riemannian approach to batch normalization. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3a0844cee4fcf57de0c71e9ad3035478-Paper.pdf.

Cooper, Y. Global minima of overparameterized neural networks. *SIAM Journal on Mathematics of Data Science*, 3(2):676–691, 2021.

Freidlin, M. I. and Wentzell, A. D. *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, third edition, 2012. ISBN 978-3-642-25846-6. doi: 10.1007/978-3-642-25847-3. URL https://doi.org/10.1007/978-3-642-25847-3. Translated from the 1979 Russian original by Joseph Szücs.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, pp. 8789–8798, 2018.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Helffer, B. and Sjöstrand, J. Effet tunnel pour l'opérateur de Schrödinger semi-classique. I. In *Proceedings of the conference on partial differential equations, Vol. 1, 2 (Saint Jean de Monts, 1985)*, pp. Exp. No. 13, 38. Soc. Math. France, Paris, 1985.

Hérau, F., Hitrik, M., and Sjöstrand, J. Tunnel effect and symmetries for Kramers-Fokker-Planck type operators. *J. Inst. Math. Jussieu*, 10(3):567–634, 2011. ISSN 1474-7480. doi: 10.1017/S1474748011000028. URL https://doi.org/10.1017/S1474748011000028.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013. ISBN 978-0-521-54823-6.

Hsu, E. P. *Stochastic analysis on manifolds*, volume 38 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002. ISBN 0-8218-0802-8. doi: 10.1090/gsm/038. URL https://doi.org/10.1090/gsm/038.

Huang, Z. and Becker, S. Stochastic gradient langevin dynamics with variance reduction. *arXiv:2102.06759*, 2021.

Hwang, C.-R. Laplace's method revisited: weak convergence of probability measures. *Ann. Probab.*, 8(6):1177–1182, 1980.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. Averaging weights leads to wider optima and better generalization. In Silva, R., Globerson, A., and Globerson, A. (eds.), *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018, pp. 876–885. Association For Uncertainty in Artificial Intelligence (AUAI), January 2018. 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018 ; Conference date: 06-08-2018 Through 10-08-2018.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Kolokoltsov, V. N. *Semiclassical analysis for diffusions and stochastic processes*, volume 1724 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000. ISBN 3-540-66972-8. doi: 10.1007/BFb0112488. URL https://doi.org/10.1007/BFb0112488.

Kuditipudi, R., Wang, X., Lee, H., Zhang, Y., Li, Z., Hu, W., Ge, R., and Arora, S. Explaining landscape connectivity of low-cost solutions for multilayer nets. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/46a4378f835dc8040c8057beb6a2da52-Paper.pdf.

Li, Q., Tai, C., and E, W. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019. URL http://jmlr.org/papers/v20/17-526.html.

Li, Z., Lyu, K., and Arora, S. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, abs/2010.02916, 2020.

Li, Z., Malladi, S., and Arora, S. On the validity of modeling SGD with stochastic differential equations (SDEs). In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021a. URL https://openreview.net/forum?id=goEdyJ_nVQI.

Li, Z., Wang, T., and Arora, S. What happens after sgd reaches zero loss? –a mathematical framework, 2021b.

Maddox, W. J., Benton, G., and Wilson, A. G. Rethinking parameter counting in deep models: Effective dimensionality revisited. *arXiv:2003.02139*, 2020.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

Michel, L. About small eigenvalues of the Witten Laplacian. *Pure Appl. Anal.*, 1(2):149–206, 2019. ISSN 2578-5885. doi: 10.2140/paa.2019.1.149. URL https://doi.org/10.2140/paa.2019.1.149.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1674–1703. PMLR, 07–10 Jul 2017. URL https://proceedings.mlr.press/v65/raginsky17a.html.

Shi, B., Su, W. J., and Jordan, M. I. On learning rates and schrödinger operators. *arXiv preprint arXiv:2004.06977*, 2020.

Simon, B. Semiclassical analysis of low lying eigenvalues. I. Nondegenerate minima: asymptotic expansions. *Ann. Inst. H. Poincaré Sect. A (N.S.)*, 38(3):295–308, 1983. ISSN 0246-0211.

Smith, S. L. and Le, Q. V. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018.

Smith, S. L., Kindermans, P.-J., and Le, Q. V. Don't decay the learning rate, increase the batch size. In *International Conference on Learning Representations*, 2018.

Smith, S. L., Dherin, B., Barrett, D., and De, S. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rq_Qr0c1Hyo.

Xu, P., Chen, J., Zou, D., and Gu, Q. Global convergence of langevin dynamics based algorithms for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3126?3137, 2018.

Yaida, S. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkNksoRctQ.

Zhang, Y., Liang, P., and Charikar, M. A hitting time analysis of stochastic gradient langevin dynamics. In Kale, S. and Shamir, O. (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1980–2022. PMLR, 07–10 Jul 2017.

## A. Proofs of results in §3

*Proof of Theorem 3.1.* First remark that given any initial position $\overline{\boldsymbol{W}}_0 \in \mathbb{S}^{d-1}$, the solution to (4) almost surely stays in $\mathbb{S}^{d-1}$ by (Li et al., 2020, Thm 5.1).

Write $\boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}} = (\boldsymbol{a}_1(\boldsymbol{w}), ..., \boldsymbol{a}_d(\boldsymbol{w}))$. Note that for $\boldsymbol{w} \in \mathbb{S}^{d-1}$, $\boldsymbol{a}_j(\boldsymbol{w}) \in T_{\boldsymbol{w}}\mathbb{S}^{d-1}$. Then the random solution $\overline{\boldsymbol{W}}_t$ to (7) satisfies the following property:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}f(\overline{\boldsymbol{W}}_t) = \mathbb{E}(\overline{\Psi}f(\overline{\boldsymbol{W}}_t)), \tag{16}$$

where $\overline{\Psi}$ is a differential operator on $\mathbb{S}^{d-1}$ defined by

$$\overline{\Psi}f(\boldsymbol{w}) = -\gamma_t^{-\frac{1}{2}}\overline{\nabla}\mathcal{L}(\boldsymbol{w})\cdot\overline{\nabla}f(\boldsymbol{w}) + \frac{1}{2}\gamma_t^{-1}\sum_{j=1}^{d}(\overline{\nabla}_{\boldsymbol{a}_j})^2 f(\boldsymbol{w}). \tag{17}$$

where $\overline{\nabla}_{\boldsymbol{a}_j}$ is the covariant derivative along the vector field $\boldsymbol{a}_j$ on $\mathbb{S}^{d-1}$ with respect to the spherical metric (see (Hsu, 2002, Theorem 1.3.4)). The operator $\overline{\Psi}$ is called the generator of (7). Similarly, the random solution $\overline{\boldsymbol{W}}_t$ to (4) satisfies : $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}f(\overline{\boldsymbol{W}}_t) = \mathbb{E}(\Psi f(\overline{\boldsymbol{W}}_t))$ where $\Psi$ is a differential operator on $\mathbb{R}^d$ defined by

$$\begin{aligned}\Psi f(\boldsymbol{w}) =& -\gamma_t^{-\frac{1}{2}}\nabla\mathcal{L}(\boldsymbol{w})\cdot\nabla f(\boldsymbol{w}) \\ &+ \frac{1}{2}\gamma_t^{-1}\sum_{j=1}^{d}\nabla^2_{\boldsymbol{a}_j} f(\boldsymbol{w}) \\ &- \frac{1}{2}\gamma_t^{-1}\mathrm{Tr}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{w}\cdot\nabla f(\boldsymbol{w}).\end{aligned} \tag{18}$$

We emphasize that besides the removal of the last term, $\overline{\Psi}$ differs from $\Psi$ in that differential operators and the Riemannian metric on $\mathbb{S}^{d-1}$ are used instead of those on $\mathbb{R}^{d-1}$.

We next analyze the difference between corresponding terms in (17) and (18) at $\boldsymbol{w} \in \mathbb{S}^{d-1}$.

Recall that $T_{\boldsymbol{w}}\mathbb{S}^{d-1}$ is the orthogonal complement $V_{\boldsymbol{w}}^{\perp}$ of $\boldsymbol{w}$, and the unit normal vector at $\boldsymbol{w}$ is given by $\boldsymbol{w}$ itself. Thus $\overline{\nabla}f(\boldsymbol{w})$ is the $V_{\boldsymbol{w}}^{\perp}$-component of $\nabla f(\boldsymbol{w})$ and $\overline{\nabla}f(\boldsymbol{w}) = \nabla f(\boldsymbol{w}) - (\nabla_{\boldsymbol{w}} f(\boldsymbol{w}))\boldsymbol{w}$. Similarly $\overline{\nabla}\mathcal{L} = \nabla\mathcal{L} - (\nabla_{\boldsymbol{w}}\mathcal{L})\boldsymbol{w} = \nabla\mathcal{L}$, where we used the fact that $\mathcal{L}$ is scaling-invariant and $\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}) = 0$. Moreover, by (6), $\nabla\mathcal{L}(\boldsymbol{w})\cdot(\overline{\nabla}f(\boldsymbol{w}) - \nabla f(\boldsymbol{w})) = (-\nabla_{\boldsymbol{w}} f(\boldsymbol{w}))(\nabla\mathcal{L}(\boldsymbol{w})\cdot\boldsymbol{w}) = 0$. Therefore, (17) and (18) share the same first term.

For the second term, we use the fact that

$$(\overline{\nabla}_{\boldsymbol{a}_i})^2 - (\nabla_{\boldsymbol{a}_i})^2 = -\Pi_{\boldsymbol{w}}(\boldsymbol{a}_i, \boldsymbol{a}_i)\nabla_{\boldsymbol{w}}$$

at $\boldsymbol{w} \in \mathbb{S}^{d-1}$, where $\Pi_{\boldsymbol{w}}$ is the second fundamental form of $\mathbb{S}^{d-1}$ at $\boldsymbol{w}$ and $\nabla_{\boldsymbol{w}}$ is the directional derivative along the outward normal direction. Since $\boldsymbol{w}\cdot\nabla f = \nabla_{\boldsymbol{w}} f$, after comparing, we know that there exists a function $G$ on $\mathbb{S}^{d-1}$ such that $\overline{\Psi}f(\boldsymbol{w}) - \Psi f(\boldsymbol{w}) = G(\boldsymbol{w})\nabla_{\boldsymbol{w}} f$.

Because trajectories of both (4) and (7) remain in $\mathbb{S}^{d-1}$, for $\boldsymbol{w} \in \mathbb{S}^{d-1}$ and two smooth functions $f, g$ that coincide on $\mathbb{S}^{d-1}$, $\Psi f(\boldsymbol{w}) = \Psi g(\boldsymbol{w})$ and $\overline{\Psi}f(\boldsymbol{w}) = \overline{\Psi}g(\boldsymbol{w})$. Hence $G(\boldsymbol{w})\nabla_{\boldsymbol{w}}(f-g)(\boldsymbol{w}) = 0$. Because $\nabla_{\boldsymbol{w}}(f-g)(\boldsymbol{w})$ can be chosen arbitrarily, we must have $G(\boldsymbol{w}) = 0$. Therefore, $\Psi f = \overline{\Psi}f$ on $\mathbb{S}^{d-1}$ for all functions $f$.

Therefore, restricted to $\mathbb{S}^{d-1}$, (4) and (7) have the same generator and are hence equivalent as diffusion processes. The proof is completed. □

*Proof of Theorem 3.3.* We first prove the theorem for the updating method (9). Under continuous limit, the surrogate for $\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{w}) = \nabla\mathcal{L}(\boldsymbol{w}) + \big(\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w})\big)$ is $\nabla\mathcal{L}(\boldsymbol{w})\mathrm{d}t + \mathrm{d}\boldsymbol{\xi}_t$ where $\boldsymbol{\xi}_t$ is a Gaussian diffusion of covariance $\Sigma(\boldsymbol{w})$, such a diffusion can be taken to be $\boldsymbol{\xi}_t = \int \boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d$.

Using the Taylor expansion

$$\begin{aligned}\exp_{\boldsymbol{w}}(\boldsymbol{v}) =& \boldsymbol{w} + \boldsymbol{w}(\cos|\boldsymbol{v}| - 1) + \boldsymbol{v}\frac{\sin|\boldsymbol{v}|}{|\boldsymbol{v}|} \\ =& \boldsymbol{w}(1 - \frac{1}{2}|\boldsymbol{v}|^2 + \frac{1}{4!}|\boldsymbol{v}|^4 - ...) \\ &+ \boldsymbol{v}(1 - \frac{1}{3!}|\boldsymbol{v}|^2 + ...),\end{aligned}$$

the surrogate dynamics for (9) is

$$\begin{aligned}&\mathrm{d}\overline{\boldsymbol{W}}_t \\ =&\overline{\boldsymbol{W}}_t\Big( -\frac{1}{2}\Big|\zeta\big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\big)\Big|^2 \\ &+ \frac{1}{4!}\Big|\zeta\big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\big)\Big|^4 - ...\Big) \\ &- \zeta\big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\big) \\ &\Big(1 - \frac{1}{3!}\Big|\zeta\big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\big)\Big|^2 + ...\Big).\end{aligned}$$

Itô's calculus gives

$$\begin{aligned}&\Big|\zeta\big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\big)\Big|^2 \\ =&\zeta^2(\mathrm{d}\boldsymbol{B}_t^d)^{\top}(\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}})^{\top}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d \\ =&\zeta^2 Tr\big((\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}})^{\top}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\big)\mathrm{d}t \\ =&\zeta^2 Tr\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)\mathrm{d}t.\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{d}\overline{\boldsymbol{W}}_t =&\overline{\boldsymbol{W}}_t\Big( -\frac{1}{2}\zeta^2\mathrm{Tr}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)\mathrm{d}t\Big) \\ &- \zeta\big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\big) \\ &(1 - \frac{1}{6}\zeta^2\mathrm{Tr}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)\mathrm{d}t) \\ =& -\zeta\big(\nabla\mathcal{L}(\overline{\boldsymbol{W}}_t)\mathrm{d}t + \boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t^d\big) \\ &- \frac{1}{2}\zeta^2\mathrm{Tr}\boldsymbol{\Sigma}(\overline{\boldsymbol{W}}_t)\overline{\boldsymbol{W}}_t\mathrm{d}t.\end{aligned}$$

Note that only the terms $\boldsymbol{w}(1-\frac{1}{2}|\boldsymbol{v}|^2)+\boldsymbol{v}$ came into the final expression. Since this is exactly the same equation as (4) with $\gamma_t^{-\frac{1}{2}}=\zeta$, the statement now follows from Theorem 3.1.

For the updating method (9), it suffices to use, for $\boldsymbol{w}\in\mathbb{S}^{d-1}$ and $\boldsymbol{v}\perp\boldsymbol{w}$, the Taylor expansion

$$\begin{aligned}\frac{\boldsymbol{w}+\boldsymbol{v}}{|\boldsymbol{w}+\boldsymbol{v}|}=&(\boldsymbol{w}+\boldsymbol{v})(1+|\boldsymbol{v}|^2)^{-\frac{1}{2}}\\ =&\boldsymbol{w}(1-\frac{1}{2}|\boldsymbol{v}|^2+\frac{3}{4}|\boldsymbol{v}|^4+...)\\ &+\boldsymbol{v}(1-\frac{1}{2}|\boldsymbol{v}|^2+\frac{3}{4}|\boldsymbol{v}|^4+...)\end{aligned}$$

Since we still have $\boldsymbol{w}(1-\frac{1}{2}|\boldsymbol{v}|^2)+\boldsymbol{v}$ as leading terms, the same argument as above applies and concludes the proof. □

## B. Formulation of the Fokker-Planck equation

In this section, we justify the Fokker-Planck equation (11), which is a standard fact.

*Proof of Proposition 4.3.* As in the proof of Theorem 3.1, write $\boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}}=(\boldsymbol{a}_1(\boldsymbol{w}),...,\boldsymbol{a}_d(\boldsymbol{w}))$. Because for $\boldsymbol{w}\in\mathbb{S}^{d-1}$

$$\begin{aligned}&\big(\sum_{j=1}^d\boldsymbol{a}_j(\boldsymbol{w})^\top\boldsymbol{a}_j(\boldsymbol{w})\big)\big|_{T_{\boldsymbol{w}}\mathbb{S}^{d-1}}\\ =&\big((\boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}})^\top\boldsymbol{\Sigma}(\boldsymbol{w})^{\frac{1}{2}}\big)\big|_{T_{\boldsymbol{w}}\mathbb{S}^{d-1}}\\ =&\boldsymbol{\Sigma}(\boldsymbol{w})|_{T_{\boldsymbol{w}}\mathbb{S}^{d-1}}\\ =&\sigma^2\mathrm{Id}|_{T_{\boldsymbol{w}}\mathbb{S}^{d-1}},\end{aligned}$$

the operator $\sum_{j=1}^d\overline{\nabla}_{a_j}^2$ equals $\sigma^2\overline{\Delta}$. (Recall $\overline{\Delta}$ denotes the Laplacian on $\mathbb{S}^{d-1}$.)

As we are working with the effective learning rate $\gamma_t^{-\frac{1}{2}}$ equal to a constant $\zeta$. The random solution $\overline{\boldsymbol{W}}_t$ to (8) satisfies (16) with

$$\begin{aligned}\overline{\Psi}f(\boldsymbol{w})=&-\zeta\overline{\nabla}\mathcal{L}(\boldsymbol{w})\cdot\overline{\nabla}f(\boldsymbol{w})+\frac{1}{2}\zeta^2\sum_{j=1}^d(\overline{\nabla}_{\boldsymbol{a}_j})^2f(\boldsymbol{w})\\ =&-\zeta\overline{\nabla}\mathcal{L}(\boldsymbol{w})\cdot\overline{\nabla}f(\boldsymbol{w})+\frac{1}{2}\zeta^2\sigma^2\overline{\Delta}f(\boldsymbol{w}).\end{aligned}\tag{19}$$

Then for all smooth test functions $f$, (16) can be reformulated as

$$\begin{aligned}&\partial_t\int_{\mathbb{S}^{d-1}}u(t,\boldsymbol{w})f(\boldsymbol{w})\\ =&\int_{\mathbb{S}^{d-1}}u(t,\boldsymbol{w})\big(-\zeta\overline{\nabla}\mathcal{L}(\boldsymbol{w})\cdot\overline{\nabla}f(\boldsymbol{w})+\frac{1}{2}\zeta^2\sigma^2\overline{\Delta}f(\boldsymbol{w})\big).\end{aligned}$$

After integration by parts, this is equivalent to the desired Fokker-Planck equation (11). □

## C. Linear time change and notations

We first perform a time change. Let $\beta$ and $\mu^{(\beta)}$ be as in (12), (13). Write $T=\zeta t$ and accordingly

$$\widetilde{\boldsymbol{W}}_T^{(\beta)}:=\overline{\boldsymbol{W}}_{\frac{T}{\zeta}};\ \ \tilde{u}(T,\boldsymbol{w})=u(\frac{T}{\zeta},\boldsymbol{w}).\tag{20}$$

After this time change, (8) and (11) respectively become

$$\mathrm{d}\widetilde{\boldsymbol{W}}_T^{(\beta)}=-\overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(\beta)})\mathrm{d}T-\beta^{\frac{1}{2}}\overline{\boldsymbol{\Sigma}}(\widetilde{\boldsymbol{W}}_T^{(\beta)})^{\frac{1}{2}}\mathrm{d}\boldsymbol{B}_T^d,\tag{21}$$

and

$$\partial_T\tilde{u}=\mathcal{D}^{(\beta)}\tilde{u},\tag{22}$$

where

$$\mathcal{D}^{(\beta)}\tilde{u}:=\overline{\nabla}\cdot(\tilde{u}\overline{\nabla}\mathcal{L})+\beta\overline{\Delta}\tilde{u}.\tag{23}$$

Like in the notations for $\overline{\boldsymbol{W}}_t$, We will denote by $\mathcal{P}_{\widetilde{\boldsymbol{W}}_0^{(\beta)}=\boldsymbol{w}}(\widetilde{\boldsymbol{W}}_T^{(\beta)})$ and $\mathcal{P}_{\widetilde{\boldsymbol{W}}_0^{(\beta)}\sim\nu}(\widetilde{\boldsymbol{W}}_T^{(\beta)})$ the probability distribution of the random solution $\widetilde{\boldsymbol{W}}_t$ to (21), respectively under initial conditions $\widetilde{\boldsymbol{W}}_0^{(\beta)}=\boldsymbol{w}$ and $\widetilde{\boldsymbol{W}}_0^{(\beta)}\sim\nu$ where $\nu$ is a measure (which we allow to be a non-probability). In addition we define an operator $\widetilde{\mathcal{F}}_T^{(\beta)}$ between measures by:

$$\widetilde{\mathcal{F}}_T^{(\beta)}\nu:=\mathcal{P}_{\widetilde{\boldsymbol{W}}_0^{(\beta)}\sim\nu}(\widetilde{\boldsymbol{W}}_T^{(\beta)}),\tag{24}$$

in other words $\widetilde{\mathcal{F}}_T^{(\beta)}$ is the pushforward by the SDE (21) for time $T$.

Remark that $\widetilde{\mathcal{F}}_T^{(\beta)}$ is a linear operator between positive measures and preserves total mass. Moreover, $\widetilde{\mathcal{F}}_T^{(\beta)}$ forms a semigroup parametered by $T$: $\widetilde{\mathcal{F}}_T^{(\beta)}\circ\widetilde{\mathcal{F}}_{T'}^{(\beta)}=\widetilde{\mathcal{F}}_{T+T'}^{(\beta)}$.

For convenience, we will occasionally denote by $\widetilde{\boldsymbol{W}}_T^{(\beta)}(\boldsymbol{w})$ fo the value of $\widetilde{\boldsymbol{W}}_T^{(\beta)}$ subject to initial condition $\widetilde{\boldsymbol{W}}_0^{(\beta)}=\boldsymbol{w}$. Note that this value is random.

## D. Proof of the descent stage

This section contains the proof of Theorem 4.5. After the time change, Theorem 4.5 is equivalent to:

**Theorem D.1.** *Under Assumptions 4.1 and 4.2, for all $\epsilon>0$ and $Q_1>0$, there exists $C_{\mathrm{des}},\beta_{\mathrm{des}}>0$ and a set $\Lambda_\epsilon$ of volume $\mathbf{m}(\Lambda_\epsilon)>1-\epsilon$, such that for all $\beta<\beta_{\mathrm{des}}$ and all $\boldsymbol{w}_0\in\Lambda_\epsilon$, the solutions to (21) starting at $\boldsymbol{w}_0$ satisfy $\mathbb{P}_{\widetilde{\boldsymbol{W}}_0^{(\beta)}=\boldsymbol{w}_0}\Big(\widetilde{\boldsymbol{W}}_{T_{\mathrm{des}}}^{(\beta)}\in U_{k,Q_1}\Big)>1-\epsilon$, where $U_k$ is the unique attracting basin in $\mathbb{S}^{d-1}$ that contains $\boldsymbol{w}_0$.*

We will first focus on the gradient flow without an diffusion term and then prove the diffusion component is a small peturbation in the begining of the dynamics.

### D.1. Study of the gradient flow

Note that by allowing to endow $\beta$ with value 0, again via the time change $T = \zeta t$, the gradient flow (14) is equivalent to the flow $\widetilde{\boldsymbol{W}}_T^{(0)}$ defined by (21) with the same initial condition. Hence we will think of the attracting basins $U_{\boldsymbol{z}}$ as attracting basins for the flow $\widetilde{\boldsymbol{W}}_T^{(0)}$. We emphasize that $\widetilde{\boldsymbol{W}}_T^{(0)}$ is deterministic and has only one possible trajectory starting from a given initial position.

**Lemma D.2.** *Under Assumption 4.2,*

1. $\mathbb{S}^{d-1}$ *is the disjoint union of* $\{U_{\boldsymbol{z}}\}_{\boldsymbol{z}\in Z}$*;*
2. $U_i$ *is open for each* $i = 1, ..., m$*, and the union* $\bigcup_{i=1}^m U_i$ *is dense in* $\mathbb{S}^{d-1}$*;*
3. *If* $\boldsymbol{z} \in Z\backslash\{\boldsymbol{z}_1, ..., \boldsymbol{z}_m\}$*, then* $U_{\boldsymbol{z}}$ *is a submanifold whose dimension is strictly less than* $d - 1$*.*

This is an elementary fact and the proof is omitted.

**Lemma D.3.** *For all* $Q' > 0$ *and* $\epsilon > 0$*, there exists* $\theta = \theta(Q', \epsilon) > 0$*, such that the volume of the set*

$$\begin{aligned}\Lambda_\theta^1 := \bigcup_{i=1}^m \Big\{ &\boldsymbol{w}_0 \in U_i : \forall T \geq 0, \\ &\text{either } \mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}(\boldsymbol{w}_0)) - \mathcal{L}(\boldsymbol{z}_i) < Q' \\ &\text{or } |\overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}(\boldsymbol{w}_0))|^2 > \theta. \Big\}\end{aligned}$$

*satisfies* $\mathbf{m}(\Lambda_\theta^1) > 1 - \epsilon$*.*

*Proof.* Since the set $\Lambda_\theta^1$ is decreasing in $\theta$, it suffices to show that $\mathbf{m}(\bigcup_{\theta>0} \Lambda_\theta^1) = 1$. We claim that $\bigcup_{\theta>0} \Lambda_\theta^1 \supseteq \bigcup_{i=1}^m U_i$. The lemma would then follow from Lemma D.2.

Suppose $\widetilde{\boldsymbol{W}}_0^{(0)} = \boldsymbol{w}_0 \in U_i$ and $T \geq 0$, then $\widetilde{\boldsymbol{W}}_T^{(0)} \in U_i$ as well. It suffices to show there exists $\theta$ that depends on $\boldsymbol{w}_0$ but not on $T$, such that if $\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) - \mathcal{L}(\boldsymbol{z}_i) \geq Q'$ then $|\overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)})|^2 > \theta$.

Indeed, as $\lim_{T\to\infty}(\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) - \mathcal{L}(\boldsymbol{z}_i)) = 0$, $\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) - \mathcal{L}(\boldsymbol{z}_i) \geq Q'$ only happens on a fixed interval $[0, T_0]$.

Moreover, $\overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) > 0$ for all $T \geq 0$, because otherwise $\widetilde{\boldsymbol{W}}_T^{(0)} \in Z$ and hence $\boldsymbol{w} = \widetilde{\boldsymbol{W}}_T^{(0)}$ is a critical point, which must be $\boldsymbol{z}_i$ since $\boldsymbol{w} \in U_i$. In this case $\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) - \mathcal{L}(\boldsymbol{z}_i) \geq Q'$ does not hold.

Therefore, as $\overline{\nabla}\mathcal{L}$ is continuous on the compact interval $[0, T_0]$, it suffices to choose $\theta = \sup_{T\in[0,T_0]} |\overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)})|^2 > 0$. □

**Corollary D.4.** *Let* $\theta = \theta(Q', \epsilon)$ *be as in the lemma above. If* $\boldsymbol{w}_0 \in \Lambda_\theta^1$ *then* $\boldsymbol{w}_0 \in U_i$ *for some* $1 \leq i \leq m$*. Moreover, if* $\widetilde{\boldsymbol{W}}_0^{(0)} = \boldsymbol{w}_0$*, then for* $T = \frac{\max\mathcal{L} - \min\mathcal{L}}{\theta}$*,* $\widetilde{\boldsymbol{W}}_T^{(0)} \in U_i$ *and* $\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) - \mathcal{L}(\boldsymbol{z}_i) < Q'$*.*

*Proof.* By definition $\boldsymbol{w} \in U_i$ for some $i$. Assume that $\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) - \mathcal{L}(\boldsymbol{z}_i) \geq Q'$, then $\mathcal{L}(\widetilde{\boldsymbol{W}}_s^{(0)}) - \mathcal{L}(\boldsymbol{z}_i) \geq Q$ for all $s \in [0, T]$ as the gradient flow decreases $\mathcal{L}$. Thus $|\overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_s^{(0)})|^2 \geq \theta$ and

$$\begin{aligned}&\mathcal{L}(\boldsymbol{w}_0) - \mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) \\ =& -\int_0^T \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}(\widetilde{\boldsymbol{W}}_s^{(0)}) \\ =& -\int_0^T \overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_s^{(0)})^\top \cdot (-\overline{\nabla}\mathcal{L}(\widetilde{\boldsymbol{W}}_s^{(0)}))\mathrm{d}s \\ =& \int_0^T |\overline{\nabla}\mathcal{L}(g_s w)|^2 \mathrm{d}s \geq \theta T \geq \max\mathcal{L} - \min\mathcal{L},\end{aligned}$$

which cannot be true as $\mathcal{L}(\widetilde{\boldsymbol{W}}_T^{(0)}) > \mathcal{L}(\boldsymbol{z}_i) \geq \min\mathcal{L}$. This completes the proof. □

### D.2. Perturbative estimate

As the following lemma shows, the SDE (21) is a perturbation of the gradient flow on short time scales, during which the diffusion effect is weak and dominated by the speed to the gradient flow.

**Lemma D.5.** *Assume* $\widetilde{\boldsymbol{W}}_T^{(\beta)}$ *and* $\widetilde{\boldsymbol{W}}_T^{(0)}$ *start from the same initial position* $\boldsymbol{w}_0$ *at* $T = 0$*, then*

$$\mathbb{E}\Big(\mathrm{dist}(\widetilde{\boldsymbol{W}}_T^{(\beta)}(\boldsymbol{w}_0), \widetilde{\boldsymbol{W}}_T^{(0)}(\boldsymbol{w}_0))\Big)^2 = O_T(\beta)$$

*for a fixed* $T > 0$ *and sufficiently small* $\beta$*.*

*Proof.* By (Freidlin & Wentzell, 2012, p32), there exists a Lipschitz type constant $C = C(\mathcal{L})$, such that $\mathbb{E}\Big(\mathrm{dist}(\widetilde{\boldsymbol{W}}_T^{(\beta)}(\boldsymbol{w}_0), \widetilde{\boldsymbol{W}}_T^{(0)}(\boldsymbol{w}_0))\Big)^2 = O\Big(\beta C^2 e^{2CT}\int_0^T e^{(2C+\beta C^2)s}\mathrm{d}s\Big)$. For small $\beta$, the right hand side is $O_T(\beta)$. □

**Corollary D.6.** *In the setting as above, for a fixed* $T$*, there exists a subset* $\Lambda_{T,\epsilon}^2$ *with* $\mathbf{m}(\Lambda_{T,\epsilon}^2) > 1 - \epsilon$*, such that for sufficiently small* $\beta$ *and all* $\boldsymbol{w}_0 \in \Lambda_{T,\epsilon}^2$*, the solutions to (21) starting at* $\boldsymbol{w}$ *satisfy*

$$\begin{aligned}&\mathbb{P}\Big(\widetilde{\boldsymbol{W}}_T^{(\beta)}(\boldsymbol{w}_0) \text{ and } \boldsymbol{w} \text{ belong to the same } U_i, \\ &\quad \text{and } \mathrm{dist}(\widetilde{\boldsymbol{W}}_T^{(\beta)}(\boldsymbol{w}_0), \widetilde{\boldsymbol{W}}_T^{(0)}(\boldsymbol{w}_0)) < O_T(\beta^{\frac{1}{2}})\Big) \\ >& 1 - \epsilon.\end{aligned}$$

*Proof.* Lemma D.5 implies, by Chebyshev inequality, that $\mathbb{P}\Big(\mathrm{dist}(\widetilde{\boldsymbol{W}}_T^{(\beta)}(\boldsymbol{w}_0), \widetilde{\boldsymbol{W}}_T^{(0)}(\boldsymbol{w}_0)) < O_T(\beta^{\frac{1}{2}})\Big) > 1 - \epsilon$ for all initial positions $\boldsymbol{w}$. Remember that $\widetilde{\boldsymbol{W}}_T^{(0)}(\boldsymbol{w}_0)$ and $\boldsymbol{w}_0$ belong to the same attracting basin. For $\widetilde{\boldsymbol{W}}_T^{(\beta)}$ and $\boldsymbol{w}$ to be

in the same $U_i$, we can take

$$\widetilde{\Lambda}^2_{T,\beta} := \bigcup_{i=1}^m \{\boldsymbol{w}_0 \in U_i : B_{O_T(\beta^{\frac{1}{2}})}(\widetilde{\boldsymbol{W}}^{(0)}_T(\boldsymbol{w}_0)) \subseteq U_i\}.$$

By Lemma D.2, if $\boldsymbol{w}_0 \notin \widetilde{\Lambda}^2_{T,\beta}$, then the neighborhood $B_{O_T(\beta^{\frac{1}{2}})}(\widetilde{\boldsymbol{W}}^{(0)}_T(\boldsymbol{w}_0))$ meets the boundary of $U_i$, which is contained in $\bigcup_{\boldsymbol{z}\in Z\setminus\{\boldsymbol{z}_1,\dots,\boldsymbol{z}_i\}} U_{\boldsymbol{z}}$, a union of finitely many proper submanifolds. In other words, $\widetilde{\boldsymbol{W}}^{(0)}_T(\boldsymbol{w}_0)$ is in the $O_T(\beta^{\frac{1}{2}})$-neighborhood $B_{O_T(\beta^{\frac{1}{2}})}\Big(\bigcup_{\boldsymbol{z}\in Z\setminus\{\boldsymbol{z}_1,\dots,\boldsymbol{z}_i\}} U_{\boldsymbol{z}}\Big)$ of this union. It follows that $(\widetilde{\Lambda}^2_{T,\beta})^c \subseteq \widetilde{\boldsymbol{W}}^{(0)}_{-T}\Big(B_{O_T(\beta^{\frac{1}{2}})}\Big(\bigcup_{\boldsymbol{z}\in Z\setminus\{\boldsymbol{z}_1,\dots,\boldsymbol{z}_i\}} U_{\boldsymbol{z}}\Big)\Big)$, where $\widetilde{\boldsymbol{W}}^{(0)}_{-T}$ is the time-reversed gradient flow. Because

$$\begin{aligned}
&\bigcap_{\beta>0} \widetilde{\boldsymbol{W}}^{(0)}_{-T}\Big(B_{O_T(\beta^{\frac{1}{2}})}\Big(\bigcup_{\boldsymbol{z}\in Z\setminus\{\boldsymbol{z}_1,\dots,\boldsymbol{z}_i\}} U_{\boldsymbol{z}}\Big)\Big)\\
=&\widetilde{\boldsymbol{W}}^{(0)}_{-T}\Big(\bigcup_{\boldsymbol{z}\in Z\setminus\{\boldsymbol{z}_1,\dots,\boldsymbol{z}_i\}} U_{\boldsymbol{z}}\Big)\\
=&\bigcup_{\boldsymbol{z}\in Z\setminus\{\boldsymbol{z}_1,\dots,\boldsymbol{z}_i\}} U_{\boldsymbol{z}},
\end{aligned}$$

we conclude $\lim_{\beta\to 0}\mathbf{m}(\widetilde{\Lambda}^2_{T,\beta}) = 1$. Since the sets $\widetilde{\Lambda}^2_{T,\beta}$ are decreasing in $\beta$, one may fix a sufficiently small $\beta_0 = \beta_0(T,\epsilon)$ and set $\Lambda^2_{T,\epsilon} = \widetilde{\Lambda}^2_{T,\beta_0}$, such that $\mathbf{m}(\Lambda^2_\epsilon) > 1-\epsilon$. □

### D.3. Proof of Theorem 4.5

*Proof of Theorem 4.5.* It suffices to prove Theorem D.1.

Take $T_{\text{des}} = \frac{\max\mathcal{L}-\min\mathcal{L}}{\theta(\frac{Q_1}{2},\epsilon)}$ and $\Lambda_\epsilon = \Lambda^1_\epsilon \cap \Lambda^2_{T_{\text{des}},\epsilon}$. It follows from Corollaries D.4 and D.6 that

$$\begin{aligned}
&\mathbb{P}_{\widetilde{\boldsymbol{W}}^{(\beta)}_0=\boldsymbol{w}_0}\Big(\widetilde{\boldsymbol{W}}^{(\beta)}_{T_{\text{des}}} \in U_k \text{ , and }\\
&\mathcal{L}(\widetilde{\boldsymbol{W}}^{(\beta)}_{T_{\text{des}}}) - \mathcal{L}(\boldsymbol{z}_k) < \frac{Q_1}{2} + \max|\overline{\nabla}\mathcal{L}|\cdot O_\epsilon(\beta^{\frac{1}{2}})\Big)\\
>&1-\epsilon.
\end{aligned}$$

For sufficiently small $\beta$, $\frac{Q_1}{2} + \max|\overline{\nabla}\mathcal{L}|\cdot O_\epsilon(\beta^{\frac{1}{2}}) < Q_1$. Moreover, $\mathbf{m}(\Lambda_\epsilon) > 1-2\epsilon$. To deduce the proposition, it suffices to rewrite $2\epsilon$ as $\epsilon$. □

## E. Proof of the diffusion stage

This section contains the proof of Theorem 4.6. Using the time change $T = \zeta t$, Theorem 4.6 follows from:

**Theorem E.1.** *Under Assumptions 4.1 and 4.2, for all $\epsilon > 0$, there exist constants $R_{\text{dif}}, r_{\text{dif}}, \lambda_{\text{dif}} > 0$, and a set $\Lambda_\epsilon$ of volume $\mathbf{m}(\Lambda_\epsilon) > 1-\epsilon$, such that:*

*With $\beta = \sqrt{\frac{\lambda_e}{2}}\sigma$, for all $\lambda_e < \lambda_{\text{dif}}$, the following is true for all $T \in [\frac{R_{\text{dif}}}{\beta}, e^{\frac{r_{\text{dif}}}{\beta}}]$:*

*(i) For all initial positions $\boldsymbol{w}_0 \in \Lambda_\epsilon$, $\text{dist}_{\text{TV}}\Big(\mathcal{P}_{\widetilde{\boldsymbol{W}}^{(\beta)}_0=\boldsymbol{w}_0}(\widetilde{\boldsymbol{W}}^{(\beta)}_T), \frac{\mu^{(\beta)}|_{U_k}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}\Big) \le \epsilon$, where $U_k$ is the unique attracting basin of the gradient flow of $\mathcal{L}$ that contains $\boldsymbol{w}_0$.*

*(ii) More generally, for all initial probability distribution $\nu_0$,*

$$\begin{aligned}
&\text{dist}_{\text{TV}}\Big(\mathcal{P}_{\widetilde{\boldsymbol{W}}^{(\beta)}_0\sim\nu_0}(\widetilde{\boldsymbol{W}}^{(\beta)}_T), \sum_{i=1}^m \nu_0(U_i)\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}\Big)\\
\le&\epsilon + \nu_0(\Lambda^c_\epsilon).
\end{aligned}$$

### E.1. Relevant Hilbert spaces and operators

Consider the adjoint operator $(\mathcal{D}^{(\beta)})^* = -\overline{\nabla}\mathcal{L}\cdot\overline{\nabla} + \beta\overline{\Delta}$. It is known that $(\mathcal{D}^{(\beta)})^*$ is self-adjoint on the Hilbert space $L^2(\mu^{(\beta)}) := L^2(\mathbb{S}^{d-1}, \mu^{(\beta)}\mathbf{dm})$ (Kolokoltsov, 2000, §8.5). More precisely, for smooth functions $f, g$,

$$\int f((\mathcal{D}^{(\beta)})^* g)\mu^{(\beta)}\mathbf{dm} = \int((\mathcal{D}^{(\beta)})^* f)g\mu^{(\beta)}\mathbf{dm}. \quad (25)$$

This is equivalent to

$$\int \mathcal{D}^{(\beta)}(f\mu^{(\beta)})g\mathbf{dm} = \int f\mathcal{D}^{(\beta)}(g\mu^{(\beta)})\mathbf{dm},$$

or

$$\begin{aligned}
&\int \mathcal{D}^{(\beta)}(f\mu^{(\beta)})g\mu^{(\beta)}\cdot\frac{1}{\mu^{(\beta)}}\mathbf{dm}\\
=&\int f\mu^{(\beta)}\mathcal{D}^{(\beta)}(g\mu^{(\beta)})\cdot\frac{1}{\mu^{(\beta)}}\mathbf{dm}.
\end{aligned}$$

This shows $\mathcal{D}^{(\beta)}$ is self-adjoint for the Hilbert space $L^2(\frac{1}{\mu^{(\beta)}}) := L^2(\mathbb{S}^{d-1}, \frac{1}{\mu^{(\beta)}}\mathbf{dm})$. We shall also write $L^2(1) := L^2(\mathbb{S}^{d-1}, \mathbf{dm})$ for the unweighted $L^2$ space.

The equality (25) can also be written as $\int((\mathcal{D}^{(\beta)})^* f)g\mu^{(\beta)}\mathbf{dm} = \int\mathcal{D}^{(\beta)}(f\mu^{(\beta)})g\mathbf{dm}$, which implies

$$(\mathcal{D}^{(\beta)})^* f = \frac{1}{\mu^{(\beta)}}\mathcal{D}^{(\beta)}(f\mu^{(\beta)}). \quad (26)$$

On the other hand, one can directly check the following facts:

**Lemma E.2.** *Let $f_T(\boldsymbol{w}) = \mu^{(\beta)}(\boldsymbol{w})^{-\frac{1}{2}}\tilde{u}_T(\boldsymbol{w})$. If $u$ is a solution to (22) then $f$ satisfies $\partial_T f = \mathcal{D}^{\beta,\#} f$ where*

$$\mathcal{D}^{\beta,\#} f := \beta\overline{\Delta}f - (\frac{|\overline{\nabla}\mathcal{L}|^2}{4\beta} - \frac{\overline{\Delta}\mathcal{L}}{2})f. \quad (27)$$

We also note

$$\mathcal{D}^{(\beta)}\tilde{u} = \mathcal{D}^{(\beta)}((\mu^{(\beta)})^{\frac{1}{2}}f) = (\mu^{(\beta)})^{\frac{1}{2}}\mathcal{D}^{\beta,\#}f. \quad (28)$$

The equalities (25) and (28) guarantee the following commutative diagram of operators:

$$\begin{array}{ccccc} L^2(\mu^{(\beta)}) & \xrightarrow{f\to(\mu^{(\beta)})^{\frac{1}{2}}f} & L^2(1) & \xrightarrow{f\to(\mu^{(\beta)})^{\frac{1}{2}}f} & L^2(\frac{1}{\mu^{(\beta)}}) \\ \downarrow{\scriptstyle (\mathcal{D}^{(\beta)})^*} & & \downarrow{\scriptstyle \mathcal{D}^{\beta,\#}} & & \downarrow{\scriptstyle \mathcal{D}^{(\beta)}} \\ L^2(\mu^{(\beta)}) & \xrightarrow{f\to(\mu^{(\beta)})^{\frac{1}{2}}f} & L^2(1) & \xrightarrow{f\to(\mu^{(\beta)})^{\frac{1}{2}}f} & L^2(\frac{1}{\mu^{(\beta)}}) \end{array} \tag{29}$$

In this diagram, every horizontal arrow is an isometry and every vertical arrow is a self-adjoint operator. In particular, $f \to f(\mu^{(\beta)})^{\frac{1}{2}}$ is a bijection, from the eigenfunctions of $(\mathcal{D}^{(\beta)})^*$ to those of $\mathcal{D}^{\beta,\#}$ with the same eigenvalues, then again a bijection from the later ones to the eigenfunctions of $\mathcal{D}^{(\beta)}$ with the same eigenvalues. In other words, the spectra of $(\mathcal{D}^{(\beta)})^*$ , $\mathcal{D}^{\beta,\#}$ and $\mathcal{D}^{(\beta)}$, as self-adjoint operators in their corresponding spaces, are the same. By self-adjointness, this spectrum is actually contained in $\mathbb{R}$.

Recall that $U_j$ is the attracting basin containing $\boldsymbol{z}_j$. From now on, we will denote the indicator function of $U_j$ by

$$\chi_j := \chi_{U_j}.$$

The low lying eigenvalues of $-\mathcal{D}^{(\beta)}$ correspond to the local minima. This was first proved in (Simon, 1983). The precise version that we need can be found in (Kolokoltsov, 2000).

**Proposition E.3.** *(Kolokoltsov, 2000, p248) Given $\mathcal{L}$, there exist constants $Q_0, \rho > 0$, determined by $\mathcal{L}$, such that:*

1. *The first $m$ eigenvalues (counted with multiplicity) of $-\mathcal{D}^{(\beta)}$ in $L^2(\frac{1}{\mu^{(\beta)}})$ are $\le O(e^{-\frac{Q_0}{\beta}})$;*
2. *With $\Pi^{(\beta)}$ and $(\Pi^{(\beta)})^\perp = \mathrm{Id} - \Pi^{(\beta)}$ respectively denoting the orthogonal projections in $L^2(\frac{1}{\mu^{(\beta)}})$ to the span of the first $m$ eigenfunctions of $-\mathcal{D}^{(\beta)}$ and to its orthogonal complement, $\dfrac{\|(\Pi^{(\beta)})^\perp(\chi_j\mu^{(\beta)})\|_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_j\mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} = O(e^{-\frac{Q_0}{\beta}})$ holds for $1 \le j \le m$;*
3. *All other eigenvalues of $-\mathcal{D}^{(\beta)}$ are greater than or equal to $\rho$.*

*More precisely, $Q_0$ can be any positive value such that $Q_0 < \min_i \sup_{\boldsymbol{w}\in U_i} \big(\mathcal{L}(\boldsymbol{w}) - \mathcal{L}(\boldsymbol{z}_i)\big)$.*

The original formulation in (Kolokoltsov, 2000) was for the spectral decomposition of $-(\mathcal{D}^{(\beta)})^*$. But in light of the correspondence $\mathcal{D}^{(\beta)}(\mu^{(\beta)}f) = \mu^{(\beta)}(\mathcal{D}^{(\beta)})^*f$ (see §E.1), the translation to the $\mathcal{D}^{(\beta)}$ setting is straightforward.

### E.2. Approximate spectral decomposition

For brevity, let $\hat{\chi}_j^{(\beta)} = \dfrac{\chi_j\mu^{(\beta)}}{\|\chi_j\mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}$ and $\psi_j^{(\beta)} = \Pi^{(\beta)}\hat{\chi}_j^{(\beta)}$. Then $\|\hat{\chi}_j^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} = 1$. By Proposition E.3,

$$\|\psi_j^{(\beta)} - \hat{\chi}_j^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \le O(e^{-\frac{Q_0}{\beta}}), \tag{30}$$

and the matrix

$$K = (K_{ij})_{i,j=1}^m := \big(\langle\psi_i^{(\beta)}, \psi_j^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}\big)_{i,j=1}^m$$

is within distance $O(e^{-\frac{Q_0}{\beta}})$ from Id. In consequence,

$$\|K^{-1} - \mathrm{Id}\| \le O(e^{-\frac{Q_0}{\beta}}). \tag{31}$$

The entries of $K^{-1}$ will be denoted by $K^{-1} = (K^{ij})$.

Suppose $\tilde{u}_T(\cdot) = \tilde{u}(T,\cdot)$ is the density function of the distribution of trajectories $\widetilde{\boldsymbol{W}}_T$ that start from an intial point $\boldsymbol{w}_0$. We are interested in the spectral decomposition of $\tilde{u}_T$, or more precisely the projection

$$\Pi^{(\beta)}f = \sum_{i=1}^m \Big(\sum_{j=1}^m K^{ij}\langle\tilde{u}_T, \psi_j^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}\Big)\psi_i^{(\beta)}. \tag{32}$$

The strategy is to show that the coefficient $\sum_{j=1}^m K^{ij}\langle\tilde{u}_T, \psi_j^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}$ in the decomposition above has little dependence on the choice of $\boldsymbol{w}_0$ as long as it is supported in $U_{k,Q_1}$.

### E.3. Regularity bound

From now on, suppose $U_k$ is the attracting basin containing $\boldsymbol{w}_0$. By Theorem D.1, after running the SDE (21) for $T_1 = O_{Q_1,\epsilon}(1)$ time starting from a randomly sampled point $\boldsymbol{w}_0$, the resulting point $\widetilde{\boldsymbol{W}}_{T_1}$ is with probabilty $1-\epsilon$ in the "well" $U_{k,Q_1}$ near the local minimum $\boldsymbol{z}_k$ at the bottom of $U_k$.

We will start the SDE again, on a longer time scale, from the restricted probability distribution $\big(\widetilde{\mathcal{F}}_{T_1}^{(\beta)}\delta_{\boldsymbol{w}_0}\big)|_{U_{k,Q_1}}$. In the next stage, gradient descent slows down because $\overline{\nabla}\mathcal{L}$ is small in $U_{k,Q_1}$, and diffusion behavior plays a more important role than in the previous stage.

It is noteworthy that $\big(\widetilde{\mathcal{F}}_{T_1}^{(\beta)}\delta_{\boldsymbol{w}_0}\big)^{U_{k,Q_1}}$ is absolutely continuous because $T_1 > 0$ and the parabolic PDE (22) is non-degenerate.

Instead of bounding $\tilde{u}(T,\cdot)$, we allow the stochastic process (21) to run for a duration of $\frac{R}{\beta}$ before estimating the regularity of the new marginal distribution $u_{\frac{R}{\beta}}\mathbf{dm}$. Let $p : (0,\infty)\times\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}$ be the heat kernel of the operator $\mathcal{D}^{(\beta)}$, i.e. $p(T,\boldsymbol{x},\boldsymbol{y})\mathbf{dm}(\boldsymbol{y}) = \mathrm{d}\big(\widetilde{\mathcal{F}}_T^{(\beta)}\delta_{\boldsymbol{x}}\big)(\boldsymbol{y})$. In other words, for each fixed $\boldsymbol{x}$, $p(T,\boldsymbol{x},\cdot)$ solves (22) and $p(T,\boldsymbol{x},\boldsymbol{y})\mathbf{dm}(\boldsymbol{y})$ converges weakly to $\delta_{\boldsymbol{x}}$ as $T\to 0$.

**Lemma E.4.** *There exist constants $C_1, C_2 > 0$ determined by $\mathcal{L}$, such that for every $R \geq 1$ and sufficiently small $\beta$, $|p(\frac{R}{\beta}, \cdot, \cdot)|$ is uniformly bounded by $O(e^{\frac{C_1 R + C_2}{\beta}})$.*

*Proof.* First remark that, by (29), $e^{-\frac{\mathcal{L}(x)-\mathcal{L}(y)}{2\beta}} p(T, \boldsymbol{x}, \boldsymbol{y}) = (\mu^{(\beta)})^{\frac{1}{2}}(\boldsymbol{x})(\mu^{(\beta)})^{-\frac{1}{2}}(\boldsymbol{y}) p(T, \boldsymbol{x}, \boldsymbol{y})$ is the heat kernel of the operator $\mathcal{D}^{\beta,\#}$.

We then renormalize $\mathcal{D}^{\beta,\#}$ to $\frac{\mathcal{D}^{\beta,\#}}{\beta} = \overline{\Delta} - V^{(\beta)}$ where $V^{(\beta)} := \frac{|\overline{\nabla}\mathcal{L}|^2}{4\beta^2} - \frac{\overline{\Delta}\mathcal{L}}{2\beta}$. The heat kernel of $\frac{\mathcal{D}^{\beta,\#}}{\beta}$ at time $R$ takes the form $e^{-\frac{\mathcal{L}(x)-\mathcal{L}(y)}{2\beta}} p(\frac{R}{\beta}, \boldsymbol{x}, \boldsymbol{y})$.

Note that $V^{(\beta)}$ may take negative values, but $\min V^{(\beta)} \geq -\frac{\|\overline{\Delta}\mathcal{L}\|_{L^\infty}}{2\beta}$. Define the Schrödinger operator

$$\mathcal{S}^{(\beta)} := \overline{\Delta} - (V^{(\beta)} - \min V^{(\beta)}),$$

whose heat kernel at time $R$ is $e^{(\min V^{(\beta)})R} e^{-\frac{\mathcal{L}(x)-\mathcal{L}(y)}{2\beta}} p(\frac{R}{\beta}, \boldsymbol{x}, \boldsymbol{y})$.

Because the new potential function $V^{(\beta)} - \min V^{(\beta)}$ is non-negative, by a standard maximal principle argument, the heat kernel of $\mathcal{S}^{(\beta)}$ is bounded by the Gaussian heat kernel, i.e. the one of the Laplacian $\overline{\Delta}$ on $\mathbb{S}^{d-1}$. This in particular shows that

$$\begin{aligned} &e^{(\min V^{(\beta)})R} e^{-\frac{\mathcal{L}(x)-\mathcal{L}(y)}{2\beta}} p(\frac{R}{\beta}, \boldsymbol{x}, \boldsymbol{y}) \\ \leq &O\Big(R^{-\frac{d-1}{2}} e^{-\frac{\mathrm{cdist}(\boldsymbol{x},\boldsymbol{y})^2}{R}}\Big) \end{aligned}$$

for some constant $c > 0$. Thus

$$p(\frac{R}{\beta}, \boldsymbol{x}, \boldsymbol{y}) \leq O\Big(R^{-\frac{d-1}{2}} e^{-(\min V^{(\beta)})R} e^{\frac{\mathcal{L}(x)-\mathcal{L}(y)}{2\beta}}\Big).$$

To establish the lemma, it suffices to take $C_1 = \frac{\|\overline{\Delta}\mathcal{L}\|_{L^\infty}}{2}$ and $C_2 = \max \mathcal{L} - \min \mathcal{L}$. □

**Corollary E.5.** *For $R \geq 1$, small $\beta$ and any initial measure $\nu_0$ on $\mathbb{S}^{d-1}$, $\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta}} \nu_0$ is absolutely continuous. Moreover, $\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta}} \nu_0 = q \mathrm{d}\boldsymbol{m}$ for a function $q$ with $\|q\|_{L^\infty} \leq O(e^{\frac{C_1 R + C_2}{\beta}}) \nu_0(\mathbb{S}^{d-1})$.*

The operator $\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta}}$ was defined in §C.

*Proof.* This follows from Lemma E.4 and the direct decomposition $q(\boldsymbol{y}) = \int p(\frac{R}{\beta}, \boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\nu_0(\boldsymbol{x})$. □

### E.4. Non-escaping from well

We will make a choice of the parameter $Q_1$ in Theorem D.1 that depends only on $\mathcal{L}$. The goal is to assert that the a trajectory of the SDE (21) with starting point $\boldsymbol{w} \in U_{k,Q_1}$ will be trapped in $U_{k,Q}$ for an exponentially long period with high probability. The argument below is essentially due to (Freidlin & Wentzell, 2012, §4.4). However, while their proof works for all $\boldsymbol{w} \in U_{k,Q_0}$, the purpose of reproducing it here is to demonstrate the uniformity of the estimate for all $\boldsymbol{w} \in U_{k,Q_1}$.

For every $\boldsymbol{w} \in U_{k,Q_0}$, its first exit time with respect to $U_{k,Q_0}$ is denoted by

$$\tau^{(\beta)}(\boldsymbol{w}) := \inf\Big\{T > 0 : \widetilde{\boldsymbol{W}}^{(\beta)}_T(\boldsymbol{w}) \in \partial U_{i,Q_0}\Big\}. \quad (33)$$

**Lemma E.6.** *Given $\mathcal{L}$ and $Q_0$, there exists a constant $Q_1$, such that the bound $\mathbb{P}(\tau^{(\beta)}(\boldsymbol{w}) \leq e^{\frac{Q_0}{3\beta}}) \leq O(e^{-\frac{Q_0}{3\beta}})$ holds uniformly for all $\boldsymbol{w} \in U_{k,Q_1}$.*

*Proof.* Together with another constant $r$, also determined by $\mathcal{L}$, we can make $Q_1$ satisfy the following condition:

$$\forall 1 \leq i \leq m, \overline{U_{i,Q_1}} \subset B_r(\boldsymbol{z}_i) \text{ and } \overline{B_{3r}(\boldsymbol{z}_i)} \subset U_{i,Q_0} \quad (34)$$

For every $\boldsymbol{w} \in U_{i,Q_0} \cup \partial U_{i,Q_0}$, and let $\tau_1^{(\beta)}(\boldsymbol{w})$ be the following stopping time:

$$\begin{aligned} &\tau_1^{(\beta)}(\boldsymbol{w}) \\ :=&\inf\Big\{T > 0 : \widetilde{\boldsymbol{W}}^{(\beta)}_T(\boldsymbol{w}) \in \partial B_r(\boldsymbol{z}_i) \cup \partial U_{i,Q_0} \\ &\quad \text{and } \exists T' \in [0, T) \text{ s.t. } \widetilde{\boldsymbol{W}}^{(\beta)}_{T'}(\boldsymbol{w}) \in \partial B_{2r}(\boldsymbol{z}_i)\Big\}. \end{aligned} \quad (35)$$

Fix constants $Q_0', Q_0''$ such that $0 < Q_0'' < Q_0' < Q_0$ . In addition to (34), by (Freidlin & Wentzell, 2012, Ch. 4, (4.6)), one can choose $r$ to be sufficiently small, such that for all sufficiently small $\beta$,

$$\mathbb{P}\Big(\widetilde{\boldsymbol{W}}^{(\beta)}_{\tau_1^{(\beta)}(\boldsymbol{w})}(\boldsymbol{w}) \in \partial U_{i,Q_0}\Big) \leq e^{-\frac{Q_0'}{\beta}} \quad (36)$$

for all $1 \leq i \leq m$ and $\boldsymbol{w} \in \partial B_r(\boldsymbol{z}_i) \cup \partial U_{i,Q_0}$. In addition, define for every $\boldsymbol{w} \in \overline{U_{i,Q_0}}$ recursively the $n$-th stopping times for all $n \geq 1$:

$$\tau_n^{(\beta)}(\boldsymbol{w}) := \tau_{n-1}^{(\beta)}(\boldsymbol{w}) + \tau_1^{(\beta)}(\widetilde{\boldsymbol{W}}^{(\beta)}_{\tau_{n-1}^{(\beta)}(\boldsymbol{w})}). \quad (37)$$

Note that if $\boldsymbol{w} \in B_{2r}(\boldsymbol{z}_i)$, then $\tau^{(\beta)}(\boldsymbol{w}) = \tau^{(\beta)}_{N^{(\beta)}(\boldsymbol{w})}(\boldsymbol{w})$ for a random variable $N = N^{(\beta)}(\boldsymbol{w}) \in \mathbb{N}$.

Once $r$ is fixed, because the stochastic process (21) is a perturbation of the geodesic flow when $\beta$ is very small and the separation condition (34) holds, there exists a constant $\theta > 0$ such that $\mathbb{P}(\tau_1^{(\beta)}(\boldsymbol{w}) > \theta) > \frac{1}{2}$ for all sufficiently small $\beta$ and $\boldsymbol{w} \in \partial B_{2r}(\boldsymbol{z}_k)$. By construction of $\tau_1^{(\beta)}$, this inequality is also true for all $\boldsymbol{w} \in \partial B_r(\boldsymbol{z}_k)$

From now on suppose $\boldsymbol{w} \in U_{i,Q_1} \subset B_r(\boldsymbol{z}_i)$, Set a target iteration number at $M^{(\beta)} = \lfloor \frac{4}{\theta} e^{\frac{Q_0''}{\beta}} \rfloor$. Then $\tau_1^{(\beta)}(\boldsymbol{w}) = \tau_0^{(\beta)}(\boldsymbol{w}) + \tau_1^{(\beta)}(\widetilde{\boldsymbol{W}}^{(\beta)}_{\tau_0^{(\beta)}(\boldsymbol{w})}(\boldsymbol{w}))$ where $\tau_0^{(\beta)}(\boldsymbol{w}) := \inf\{T > 0 : \widetilde{\boldsymbol{W}}_T^{(\beta)}(\boldsymbol{w}) \in \partial B_r(\boldsymbol{z}_i)\}$. In particular,

$$N^{(\beta)}(\boldsymbol{w}) = N^{(\beta)}\big(\widetilde{\boldsymbol{W}}^{(\beta)}_{\tau_0^{(\beta)}(\boldsymbol{w})}(\boldsymbol{w})\big). \tag{38}$$

Since $\widetilde{\boldsymbol{W}}^{(\beta)}_{\tau_0^{(\beta)}(\boldsymbol{w})}(\boldsymbol{w}) \in \partial B_r(\boldsymbol{z}_i) \cup \partial U_{i,Q_0}$, by (36),

$$\begin{aligned}\mathbb{P}(N^{(\beta)}(\boldsymbol{w}) <& M^{(\beta)}) \le 1 - \big(1 - e^{-\frac{Q_0'}{\beta}}\big)^{\lfloor \frac{4}{\theta} e^{\frac{Q_0''}{\beta}} \rfloor} \\ =& O(e^{\frac{Q_0''-Q_0'}{\beta}}),\end{aligned} \tag{39}$$

where the implied constant is uniform for $\boldsymbol{w} \in U_{i,Q_1} \subset B_r(\boldsymbol{z}_i)$.

On the other hand, since $\tau_n^{(\beta)}(\boldsymbol{w}) = \tau_0^{(\beta)}(\boldsymbol{w}) + \sum_{l=1}^n \tau_1^{(\beta)}\big(\widetilde{\boldsymbol{W}}^{(\beta)}_{\tau_{l-1}^{(\beta)}(\boldsymbol{w})}(\boldsymbol{w})\big)$, and each term in the summation, given all precedent terms, is greater than $\theta$ with at least $\frac{1}{2}$ probability, we know by large deviation principle that for an absolute constant $c > 0$,

$$\mathbb{P}(\tau^{(\beta)}_{M^{(\beta)}}(\boldsymbol{w}) \le \frac{\theta}{3} M^{(\beta)}) \le e^{-cM^{(\beta)}} \tag{40}$$

uniformly for all $\boldsymbol{w} \in U_{i,Q_1}$. Combining (39) and (40) yields that for all sufficiently small $\beta$ (depending only on $\mathcal{L}$), we have uniformly for all $\boldsymbol{w} \in U_{i,Q_1}$,

$$\begin{aligned}&\mathbb{P}(\tau^{(\beta)}(\boldsymbol{w}) \le e^{\frac{Q_0''}{\beta}}) \\ =&\mathbb{P}(\tau^{(\beta)}_{N^{(\beta)}(\omega)}(\boldsymbol{w}) \le e^{\frac{Q_0''}{\beta}}) \le \mathbb{P}(\tau^{(\beta)}_{N^{(\beta)}(\omega)}(\boldsymbol{w}) \le \frac{\theta}{3} M^{(\beta)}) \\ \le& \mathbb{P}(N^{(\beta)}(\boldsymbol{w}) < M^{(\beta)}) + \mathbb{P}(\tau^{(\beta)}_{M^{(\beta)}}(\boldsymbol{w}) \le \frac{\theta}{3} M^{(\beta)}) \\ \le& O(e^{\frac{Q_0''-Q_0'}{\beta}}) + e^{-c\lfloor \frac{4}{\theta} e^{\frac{Q_0''}{\beta}} \rfloor} \\ \le& O(e^{\frac{Q_0''-Q_0'}{\beta}}).\end{aligned} \tag{41}$$

The proof is completed by taking $Q_0'' = \frac{Q_0}{3}$ and $Q_0' = \frac{2Q_0}{3}$. □

### E.5. Short term convergence towards local equilibria

We are now ready to put the pieces together and understand the evolution during the current stage.

Let $T_1 = T_{\mathrm{des}}$ as in Theorem D.1. Fix an arbitrary parameter $R' > 0$, say $R' = 1$. Let

$$C_2' = C_2 + \frac{1}{2}(\max \mathcal{L} - \min \mathcal{L}) = \frac{3}{2}(\max \mathcal{L} - \min \mathcal{L}),$$

which depends only on $\mathcal{L}$. Consider $R \ge R'$ such that the following assumptions are satisfied:

$$R \ge \frac{\rho + C_1}{\rho} R' + \frac{C_2' + Q_0}{\rho}, \tag{42}$$

$$\frac{R'}{\beta} \le \frac{R}{\beta} \le e^{\frac{Q_0}{3\beta}}. \tag{43}$$

We first write

$$\mathcal{P}_{\overline{\boldsymbol{W}}_0 = \boldsymbol{w}_0}(\overline{\boldsymbol{W}}_{\frac{R}{\beta}}) = \widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta}} \delta_{\boldsymbol{w}_0} = \widetilde{\mathcal{F}}^{(\beta)}_{\frac{R-R'}{\beta}} \widetilde{\mathcal{F}}^{(\beta)}_{\frac{R'}{\beta}} \delta_{\boldsymbol{w}_0}. \tag{44}$$

Then decompose

$$\begin{aligned}\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R'}{\beta}} \delta_{\boldsymbol{w}_0} =& \widetilde{\mathcal{F}}^{(\beta)}_{\frac{R'}{\beta} - T_1}\big((\widetilde{\mathcal{F}}^{(\beta)}_{T_1} \delta_{\boldsymbol{w}_0})|_{U_{k,Q_1}}\big) \\ &+ \widetilde{\mathcal{F}}^{(\beta)}_{\frac{R'}{\beta} - T_1}\big((\widetilde{\mathcal{F}}^{(\beta)}_{T_1} \delta_{\boldsymbol{w}_0})|_{U^c_{k,Q_1}}\big).\end{aligned}$$

By Theorem 4.5, the total mass of $\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta} - T_1}\big((\widetilde{\mathcal{F}}^{(\beta)}_{T_1} \delta_{\boldsymbol{w}_0})|_{U^c_{k,Q_1}}\big)$ is less than $\epsilon$.

For simplicity, denote the measure $\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R'}{\beta} - T_1}\big((\widetilde{\mathcal{F}}^{(\beta)}_{T_1} \delta_{\boldsymbol{w}_0})|_{U_{k,Q_1}}\big)$ by $\gamma$. Then its properties can be summarized as follows:

(i) $\mathrm{dist}_{\mathrm{TV}}\big(\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R'}{\beta}} \delta_{\boldsymbol{w}_0}, \gamma\big) \le \epsilon$;

(ii) $\gamma$ is absolutely continuous. Moreover, $\gamma = h(\cdot)\mathrm{d}\mathbf{m}$ with a density function $h$ satisfying $\|h\| \le O(e^{\frac{C_1 R' + C_2}{\beta}})$.

Here the second property follows from Corollary E.5.

In particular, the total mass of $\gamma$ satisfies $1 - \epsilon \le \gamma(\mathbb{S}^{d-1}) \le 1$.

We now run the Fokker-Planck equation (22) starting at initial time $\frac{R'}{\beta}$ and initial data $h(\cdot)$, and denote the solution by $\tilde{h}$; that is, $\tilde{h}_T(\boldsymbol{w}) := \tilde{h}(T, \boldsymbol{w})$ is defined on $T \ge \frac{R'}{\beta}$, $\boldsymbol{w} \in \mathbb{S}^{d-1}$, and solves

$$\begin{cases} \partial_T \tilde{h}_T = & \mathcal{D}^{(\beta)} \tilde{h}_T \\ \tilde{h}_{\frac{R'}{\beta}}(\cdot) = & h(\cdot). \end{cases} \tag{45}$$

In other words, $\tilde{h}_{T - \frac{R'}{\beta}} = e^{(T - \frac{R'}{\beta})\mathcal{D}^{(\beta)}} h$. Note that as the Fokker-Planck equation preserves total mass,

$$\int \tilde{h}_T \mathrm{d}\mathbf{m} = \int h \mathrm{d}\mathbf{m} \in [1 - \epsilon, 1], \forall T \ge \frac{R'}{\beta}. \tag{46}$$

Since $\tilde{h}$ and $\tilde{u}$ both satisfy (22), for $T \ge \frac{R'}{\beta}$, $\tilde{h}_T$ will be viewed as an approximation of $\tilde{u}_T$ in the sense that

$$\begin{aligned}&\mathrm{dist}_{\mathrm{TV}}\big(\widetilde{\mathcal{F}}^{(\beta)}_T \delta_{\boldsymbol{w}_0}, \tilde{h}_T \mathrm{d}\mathbf{m}\big) \\ =& \mathrm{dist}_{\mathrm{TV}}(\tilde{u}_T \mathrm{d}\mathbf{m}, \tilde{h}_T \mathrm{d}\mathbf{m}) \\ =& \mathrm{dist}_{\mathrm{TV}}(\widetilde{\mathcal{F}}^{(\beta)}_{T - \frac{R'}{\beta}}(\tilde{u}_{\frac{R'}{\beta}} \mathrm{d}\mathbf{m}), \widetilde{\mathcal{F}}^{(\beta)}_{T - \frac{R'}{\beta}}(\tilde{h}_{\frac{R'}{\beta}} \mathrm{d}\mathbf{m})) \\ =& \mathrm{dist}_{\mathrm{TV}}(\widetilde{\mathcal{F}}^{(\beta)}_{T - \frac{R'}{\beta}}(\tilde{u}_{\frac{R'}{\beta}} \mathrm{d}\mathbf{m}), \widetilde{\mathcal{F}}^{(\beta)}_{T - \frac{R'}{\beta}} \gamma) \\ \le& \epsilon.\end{aligned} \tag{47}$$

Recall $\psi_j^{(\beta)} = \Pi^{(\beta)}\hat{\chi}_j^{(\beta)}$ and $\hat{\chi}_j^{(\beta)} - \psi_j^{(\beta)} = (\Pi^{(\beta)})^{\perp}\hat{\chi}_j^{(\beta)}$ are respectively the projections of $\hat{\chi}_j^{(\beta)}$ to the span of the first $m$ eigenvalues of $-\mathcal{D}^{(\beta)}$ and its orthogonal complement in $L^2(\frac{1}{\mu^{(\beta)}})$. The conditions (42) and (43) will be assumed on $R$ and $R'$ without further notice.

**Lemma E.7.** *For sufficiently small $\beta$,*

$$\big\|(\Pi^{(\beta)})^{\perp}\tilde{h}_{\frac{R}{\beta}}\big\|_{L^2(\frac{1}{\mu^{(\beta)}})} \le O(e^{-\frac{Q_0}{\beta}}).$$

The projections $\Pi^{(\beta)}$ and $(\Pi^{(\beta)})^{\perp}$ were defined in Proposition E.3.

*Proof.* The projection $(\Pi^{(\beta)})^{\perp}\tilde{h}_{\frac{R}{\beta}}$ can be bounded by

$$\begin{aligned}
&\big\|(\Pi^{(\beta)})^{\perp}\tilde{h}_{\frac{R}{\beta}}\big\|_{L^2(\frac{1}{\mu^{(\beta)}})}\\
=&\big\|(\Pi^{(\beta)})^{\perp}e^{\frac{R-R'}{\beta}\mathcal{D}^{(\beta)}}\tilde{h}_{\frac{R'}{\beta}}\big\|_{L^2(\frac{1}{\mu^{(\beta)}})}\\
=&\big\|e^{\frac{R-R'}{\beta}\mathcal{D}^{(\beta)}}(\Pi^{(\beta)})^{\perp}\tilde{h}_{\frac{R'}{\beta}}\big\|_{L^2(\frac{1}{\mu^{(\beta)}})}\\
\le&e^{-\rho(\frac{R-R'}{\beta})}\big\|\tilde{h}_{\frac{R'}{\beta}}\big\|_{L^2(\frac{1}{\mu^{(\beta)}})}
\end{aligned} \tag{48}$$

because by Proposition E.3.(3), the spectrum of $-\mathcal{D}^{(\beta)}$ on the image of $(\Pi^{(\beta)})^{\perp}$ is in $[\rho,\infty)$.

By the bound on $\tilde{h}_{\frac{R'}{\beta}} = h$, we have:

$$\begin{aligned}
&\|\tilde{h}_{\frac{R'}{\beta}}\|_{L^2(\frac{1}{\mu^{(\beta)}})}\\
\le&O(e^{\frac{C_1R'+C_2}{\beta}})\|1\|_{L^2(\frac{1}{\mu^{(\beta)}})}\\
=&O(e^{\frac{C_1R'+C_2}{\beta}})\Big(\int\frac{1}{\mu^{(\beta)}}\mathbf{dm}\Big)^{-\frac{1}{2}}\\
\le&O(e^{\frac{C_1R'+C_2}{\beta}})(\min\mu^{(\beta)})^{-\frac{1}{2}}\\
\le&O(e^{\frac{C_1R'+C_2}{\beta}})(e^{\frac{\min\mathcal{L}-\max\mathcal{L}}{\beta}})^{-\frac{1}{2}}\\
=&O(e^{\frac{C_1R'+C_2'}{\beta}}).
\end{aligned} \tag{49}$$

Combining (48) with (49), we obtain

$$\begin{aligned}
&\big\|(\Pi^{(\beta)})^{\perp}\tilde{h}_{\frac{R}{\beta}}\big\|_{L^2(\frac{1}{\mu^{(\beta)}})}\\
\le&O(e^{-\rho(\frac{R-R'}{\beta})}e^{\frac{C_1R'+C_2'}{\beta}})\\
\le&O(e^{-\frac{\rho R-(\rho+C_1)R'-C_2'}{\beta}})\\
\le&O(e^{-\frac{Q_0}{\beta}})
\end{aligned}$$

because of (42). □

**Notation E.8.** In addition to the usual big $O(\cdot)$ notation, we will use $\widetilde{O}(\epsilon)$ to represent an error taking value in $[-\epsilon,\epsilon]$.

**Lemma E.9.** *For each $j$,*

$$\int_{U_{j,Q_0}}\tilde{h}_{\frac{R}{\beta}}\mathbf{dm} = \begin{cases} 1+\widetilde{O}(\epsilon)+O(e^{-\frac{Q_0}{3\beta}}) & j=k;\\ O(e^{-\frac{Q_0}{3\beta}}) & j\neq k.\end{cases}$$

*Proof.* By the fact that $\tilde{h}_{\frac{R}{\beta}}\mathbf{dm} = \widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta}-T_1}\big((\widetilde{\mathcal{F}}^{(\beta)}_{T_1}\delta_{\boldsymbol{w}_0})|_{U^c_{k,Q_1}}\big)$, condition (43), and Lemma E.6, $\int_{U_{j,Q_0}}\tilde{h}_{\frac{R}{\beta}}\mathbf{dm} = \delta_{kj}\int_{\mathbb{S}^{d-1}}\tilde{h}_{\frac{R}{\beta}}\mathbf{dm} + O(e^{-\frac{Q_0}{3\beta}})$. We then conclude with (46). □

**Corollary E.10.** *For sufficiently small $\beta$,*

$$\begin{aligned}
&\langle\tilde{h}_{\frac{R}{\beta}},\psi_j^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}\\
=&\frac{\delta_{kj}+\widetilde{O}(\epsilon)+O(e^{-\frac{Q_0}{3\beta}})}{\|\chi_k\mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}+O(e^{-\frac{2Q_0}{\beta}}).
\end{aligned}$$

*Proof.* It follows from (30) and the lemma above that

$$\begin{aligned}
&\langle\tilde{h}_{\frac{R}{\beta}},\psi_j^{(\beta)}-\hat{\chi}_j^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}\\
=&\langle(\Pi^{(\beta)})^{\perp}\tilde{h}_{\frac{R}{\beta}},\psi_j^{(\beta)}-\hat{\chi}_j^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}\\
\le&O(e^{-\frac{Q_0}{\beta}})\cdot O(e^{-\frac{Q_0}{\beta}}) = O(e^{-\frac{2Q_0}{\beta}}).
\end{aligned} \tag{50}$$

Together with Lemma E.9, this implies

$$\begin{aligned}
&\langle\tilde{h}_{\frac{R}{\beta}},\hat{\chi}_j^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}\\
=&\frac{\langle\tilde{h}_{\frac{R}{\beta}},\chi_j\mu^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_j\mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}\\
=&\frac{\int_{U_{j,Q_0}}\tilde{h}_{\frac{R}{\beta}}\mu^{(\beta)}\cdot\frac{1}{\mu^{(\beta)}}\mathbf{dm}}{\|\chi_j\mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}\\
=&\frac{\int_{U_{j,Q_0}}\tilde{h}_{\frac{R}{\beta}}\mathbf{dm}}{\|\chi_j\mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}\\
=&\frac{\delta_{kj}+\widetilde{O}(\epsilon)+O(e^{-\frac{Q_0}{3\beta}})}{\|\chi_k\mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}.
\end{aligned} \tag{51}$$

The corollary is proved by adding (50) and (51). □

### E.6. Proof of Theorem 4.6

To prove Theorem 4.6, it suffices to prove Theorem E.1.

*Proof of Theorem E.1.* The constant $R_{\mathrm{dif}}$ will be the right hand in (42) and let $r_{\mathrm{dif}} = \frac{Q_0}{3}$. For $T\in[\frac{R_{\mathrm{dif}}}{\beta},e^{\frac{r_{\mathrm{dif}}}{\beta}}]$, write $T=\frac{R}{\beta}$. Then $R$ satisfies both (42) and (43).

*Part (i).* Let $U_k$ be the unique basin containing $\boldsymbol{w}_0$. First assume that $A \subseteq U_k$. In this case, using (30)

$$\begin{aligned}
&\langle \chi_A \mu^{(\beta)}, \psi_j^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\langle \chi_A \mu^{(\beta)}, \hat{\chi}_j^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
&+ \langle \chi_A \mu^{(\beta)}, \psi_j^{(\beta)} - \hat{\chi}_j^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\frac{\langle \chi_A \mu^{(\beta)}, \chi_j \mu^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_j \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} \\
&+ \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \|\psi_j^{(\beta)} - \hat{\chi}_j^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\frac{\delta_{kj} \|\chi_A \mu^{(\beta)}\|^2_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{Q_0}{\beta}}) \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\left( \frac{\delta_{kj} \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{Q_0}{\beta}}) \right) \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}
\end{aligned} \tag{52}$$

We can now, from (31), (32), (52) and Corollary E.10, deduce

$$\begin{aligned}
&\langle \Pi^{(\beta)} \tilde{h}_{\frac{R}{\beta}}, \chi_A \mu^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\sum_{i=1}^m \sum_{j=1}^m K^{ij} \langle \tilde{h}_{\frac{R}{\beta}}, \psi_i^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
&\cdot \langle \chi_A \mu^{(\beta)}, \psi_j^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\sum_{i=1}^m \sum_{j=1}^m \left( \delta_{ij} + O(e^{-\frac{Q_0}{\beta}}) \right) \\
&\cdot \left( \frac{\delta_{ki} + O(\epsilon + e^{-\frac{Q_0}{3\beta}})}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{2Q_0}{\beta}}) \right) \\
&\cdot \left( \frac{\delta_{kj} \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{Q_0}{\beta}}) \right) \\
&\cdot \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\sum_{j=1}^m \left( \frac{\delta_{kj} + \widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}})}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{2Q_0}{\beta}}) \right) \\
&\cdot \left( \frac{\delta_{kj} \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{Q_0}{\beta}}) \right) \\
&\cdot \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}
\end{aligned} \tag{53}$$

By separating the $j = k$ and $j \neq k$ terms, this becomes

$$\begin{aligned}
&\langle \Pi^{(\beta)} \tilde{h}_{\frac{R}{\beta}}, \chi_A \mu^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\left( \left( \frac{1 + \widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}})}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{2Q_0}{\beta}}) \right) \right. \\
&\cdot \left( \frac{\|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} + O(e^{-\frac{Q_0}{\beta}}) \right) \\
&\left. + \frac{O(e^{-\frac{Q_0}{\beta}})}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} \right) \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\left( \frac{\|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|^2_{L^2(\frac{1}{\mu^{(\beta)}})}} + \frac{\widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}})}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} \right) \\
&\cdot \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\frac{\|\chi_A \mu^{(\beta)}\|^2_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|^2_{L^2(\frac{1}{\mu^{(\beta)}})}} \\
&+ \frac{\left( \widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}}) \right) \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}}{\|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})}} \\
=&\frac{\int_A \mu^{(\beta)} \mathbf{dm}}{\int_{U_k} \mu^{(\beta)} \mathbf{dm}} + \widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}}).
\end{aligned} \tag{54}$$

Here we used the fact that $\|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \leq \|\chi_k \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} = \left( \int_{U_k} \mu^{(\beta)} \mathbf{dm} \right)^{\frac{1}{2}} \leq 1$.

On the other hand, by Lemma E.7,

$$\begin{aligned}
&\langle (\Pi^{(\beta)})^{\perp} \tilde{h}_{\frac{R}{\beta}}, \chi_A \mu^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
\leq& O(e^{-\frac{Q_0}{\beta}}) \|\chi_A \mu^{(\beta)}\|_{L^2(\frac{1}{\mu^{(\beta)}})} \leq O(e^{-\frac{Q_0}{\beta}})
\end{aligned} \tag{55}$$

Adding the last two inequalities gives, for all subset $A \subseteq U_k$,

$$\begin{aligned}
&\langle \tilde{h}_{\frac{R}{\beta}}, \chi_A \mu^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} \\
=&\frac{\int_A \mu^{(\beta)} \mathbf{dm}}{\int_{U_k} \mu^{(\beta)} \mathbf{dm}} + \widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}})
\end{aligned} \tag{56}$$

Assume now $A \subseteq U_k^c$. In this case, by Lemma E.9,

$$\langle \tilde{h}_{\frac{R}{\beta}}, \chi_A \mu^{(\beta)} \rangle_{L^2(\frac{1}{\mu^{(\beta)}})} = \int_A \tilde{h}_{\frac{R}{\beta}} \mathbf{dm} = O(e^{-\frac{Q_0}{3\beta}}) \tag{57}$$

Finally, for a general subset $A \subseteq \mathbb{S}^{d-1}$, we decompose $A$ into $A \cap U_k$ and $A \backslash U_k$ and apply (56) and (57), and conclude

that

$$\begin{aligned}&\int_A \tilde{h}_{\frac{R}{\beta}}\mathbf{dm}\\ =&\langle \tilde{h}_{\frac{R}{\beta}}, \chi_A \mu^{(\beta)}\rangle_{L^2(\frac{1}{\mu^{(\beta)}})}\\ =&\frac{\int_{A\cap U_k}\mu^{(\beta)}\mathbf{dm}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}} + \widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}})\\ =&\frac{\int_A \mu^{(\beta)}|_{U_k}\mathbf{dm}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}} + \widetilde{O}(\epsilon) + O(e^{-\frac{Q_0}{3\beta}})\end{aligned} \tag{58}$$

By (47), the relation (58) is equivalent to

$$\begin{aligned}&\mathrm{dist}_{\mathrm{TV}}\Big(\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta}}\delta_{\boldsymbol{w}_0}, \frac{\mu^{(\beta)}|_{U_k}}{\int_{U_k}\mu^{(\beta)}\mathbf{dm}}\mathbf{dm}\Big)\\ \le&\epsilon + O(e^{-\frac{Q_0}{3\beta}}) \le 2\epsilon.\end{aligned} \tag{59}$$

for sufficiently small $\beta$. Let $T = \frac{R}{\beta}$. After renaming $2\epsilon$ as $\epsilon$ and rewriting $\widetilde{\mathcal{F}}^{(\beta)}_{\frac{R}{\beta}}\delta_{\boldsymbol{w}_0} = \mathcal{P}_{\widetilde{\boldsymbol{W}}^{(\beta)}_0=\boldsymbol{w}_0}(\widetilde{\boldsymbol{W}}^{(\beta)}_T)$, in view of (43), we obtain Part (i) of Theorem E.1.

*Part (ii).* The second part of the theorem follows directly from Part (i) by disintegrating $\nu_0$ as $\sum_{j=1}^m \nu_0|_{U_j\cap\Lambda_\epsilon} + \nu_0|_{\Lambda_\epsilon^c}$. □

Corollary 4.7 follows from Theorem 4.6 and the following lemma:

**Lemma E.11.** *In the setting of Theorem 4.6, given $\epsilon > 0$, for sufficiently small $\beta$,*

*(i) For all $i$ with $\mathcal{L}(\boldsymbol{z}_i) > \min\mathcal{L}$,*

$$\mathrm{dist}_{\mathrm{TV}}\Big(\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}, \mu^{(\beta)}\mathbf{dm}\Big) \ge 1-\epsilon;$$

*(ii) For all $i$ with $\mathcal{L}(\boldsymbol{z}_i) = \min\mathcal{L}$,*

$$\begin{aligned}&\mathrm{dist}_{\mathrm{TV}}\Big(\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}, \mu^{(\beta)}\mathbf{dm}\Big)\\ \ge&1 - \frac{(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_i))^{-\frac{1}{2}}}{\sum_{k:\mathcal{L}(\boldsymbol{z}_k)=\min\mathcal{L}}(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_k))^{-\frac{1}{2}}} - \epsilon.\end{aligned}$$

*Proof.* By (Hwang, 1980, Thm. 2.1), as $\beta\to 0$, $\mu^{(\beta)}$ converges to a probability measure $\mu^{(0)}$ supported on $\{\boldsymbol{z}_i : \mathcal{L}(\boldsymbol{z}_i) = \min\mathcal{L}\}$ and $\mu^{(0)}(\{\boldsymbol{z}_i\}) = \frac{(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_i))^{-\frac{1}{2}}}{\sum_{k:\mathcal{L}(\boldsymbol{z}_k)=\min\mathcal{L}}(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_k))^{-\frac{1}{2}}}$ if $\mathcal{L}(\boldsymbol{z}_i) = \min\mathcal{L}$. Therefore:

For all $i$ with $\mathcal{L}(\boldsymbol{z}_i) > \min\mathcal{L}$, $\mu^{(0)}$ is supported outside $U_i$, thus

$$\mathrm{dist}_{\mathrm{TV}}\Big(\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}, \mu^{(\beta)}\mathbf{dm}\Big) \ge 1;$$

For all $i$ with $\mathcal{L}(\boldsymbol{z}_i) = \min\mathcal{L}$,

$$\begin{aligned}&\mathrm{dist}_{\mathrm{TV}}\Big(\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}, \mu^{(0)}\mathbf{dm}\Big)\\ \ge&1 - \frac{(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_i))^{-\frac{1}{2}}}{\sum_{k:\mathcal{L}(\boldsymbol{z}_k)=\min\mathcal{L}}(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_k))^{-\frac{1}{2}}}.\end{aligned}$$

Both inequalities above are obtained by comparing the measures of $U_i$.

It now suffices to take sufficiently small $\beta$ in the limit. □

*Proof of Corollary 4.7.* Suppose $m \ge 2$, choose $\lambda_e < \lambda_{\mathrm{dif}}$ sufficiently small such that $\beta$ is small enough for Lemma E.11. Let $\Omega_{\mathrm{dif}} = \Lambda_\epsilon \cap \bigcup_{i:\mathcal{L}(\boldsymbol{z}_i)>\min\mathcal{L}} U_i$ if there is only one $\boldsymbol{z}_k$ with $\mathcal{L}(\boldsymbol{z}_k) = \min\mathcal{L}$; and $\Omega_{\mathrm{dif}} = \Lambda_\epsilon \cap \bigcup_i U_i$ otherwise. Note that $\mathbf{m}(\bigcup_{i:\mathcal{L}(\boldsymbol{z}_i)>\min\mathcal{L}} U_i) > 0$ in the first case and $\bigcup_i U_i = 1$ in the second case. So in the first case $\mathbf{m}(\Omega_{\mathrm{dif}}) > \kappa_1 - \epsilon$ for some constant $\kappa_1 > 0$, and in the second case $\mathbf{m}(\Omega_{\mathrm{dif}}) > 1-\epsilon$.

Suppse $\boldsymbol{w}\in\Omega_{\mathrm{dif}}$ and $\boldsymbol{w}\in U_i$. Notice that in the second case, $1 - \frac{(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_i))^{-\frac{1}{2}}}{\sum_{k:\mathcal{L}(\boldsymbol{z}_k)=\min\mathcal{L}}(\det\overline{\nabla}^2\mathcal{L}(\boldsymbol{z}_k))^{-\frac{1}{2}}} > 0$ by the non-degeneracy Assumption 4.2. By Lemma E.11, in both cases,

$$\mathrm{dist}_{\mathrm{TV}}\Big(\frac{\mu^{(\beta)}|_{U_i}}{\int_{U_i}\mu^{(\beta)}\mathbf{dm}}, \mu^{(\beta)}\mathbf{dm}\Big) > \kappa_2 - \epsilon$$

for some constant $\kappa_2 > 0$. By Theorem 4.6,

$$\mathrm{dist}_{\mathrm{TV}}\Big(\mathcal{P}_{\overline{\boldsymbol{W}}_0=\boldsymbol{w}_0}(\overline{\boldsymbol{W}}_t), \mu^{(\beta)}\mathbf{dm}\Big) > \kappa_2 - 2\epsilon.$$

The corollary follows by fixing a sufficiently small $\epsilon < \frac{1}{4}\min(\kappa_1,\kappa_2)$ and choose $\lambda_e$ accordingly. □

## F. Proof of the tunneling stage

Using again the time change $T = \zeta t$, Theorem 4.8 would follow from:

**Theorem F.1.** *Under the genericity Assumption 4.2, There exists a constant $Q_{\mathrm{tun}}$ such that for all $T \ge 0$*

$$\begin{aligned}&\mathrm{dist}_{\mathrm{TV}}\Big(\mathcal{P}_{\widetilde{\boldsymbol{W}}^{(\beta)}_0=\boldsymbol{w}_0}(\widetilde{\boldsymbol{W}}^{(\beta)}_T), \mu^{(\beta)}\mathbf{dm}\Big)\\ \le&O(e^{-(e^{-\frac{Q_{\mathrm{tun}}}{\beta}})T})\end{aligned}$$

*holds uniformly for all initial position $\boldsymbol{w}_0\in\mathbb{S}^{d-1}$ and sufficiently small $\beta$.*

*Proof of Theorem F.1.* As noted in the discussion following Theorem 4.8, this part (or the $\mathbb{R}^d$ version of it) has essentially been proved in (Shi et al., 2020). Their proof is for fast growing functions $\mathcal{L}$ on $\mathbb{R}^d$ and is based on the fact that

there exists $Q$ such that for the second eigenvalue $\lambda_2^{(\beta)}$ of $-\mathcal{D}^{(\beta)}$,

$$\lambda_2^{(\beta)} \gtrsim e^{-\frac{Q}{\beta}}. \tag{60}$$

(Recall that the first eigenvalue is 0 as $\mathcal{D}^{(\beta)}\mu^{(\beta)} = 0$.) In our setting of the compact manifold $\mathbb{S}^{d-1}$, (60) was proved in (Michel, 2019, Thm 2.8), and the same argument as in (Shi et al., 2020) applies. □