High-dimensional asymptotics of Langevin dynamics in spiked matrix models

Tengyuan Liang
University of Chicago, 5807 South Woodlawn Avenue, 60637 IL, USA

Subhabrata Sen*
Harvard University, 1 Oxford Street, 02138 MA, USA
Corresponding author: subhabratasen@fas.harvard.edu

and

Pragya Sur
Harvard University, 1 Oxford Street, 02138 MA, USA

We study Langevin dynamics for recovering the planted signal in the spiked matrix model. We provide a 'path-wise' characterization of the overlap between the output of the Langevin algorithm and the planted signal. This overlap is characterized in terms of a self-consistent system of integro-differential equations, usually referred to as the Crisanti–Horner–Sommers–Cugliandolo–Kurchan equations in the spin glass literature. As a second contribution, we derive an explicit formula for the limiting overlap in terms of the signal-to-noise ratio and the injected noise in the diffusion. This uncovers a sharp phase transition—in one regime, the limiting overlap is strictly positive, while in the other, the injected noise overcomes the signal, and the limiting overlap is zero.

Keywords: langevin dynamics; high-dimensional asymptotics; spiked matrices; spin glasses.

1. Introduction

Gradient descent-based methods and their noisy counterparts are routinely used in modern Machine Learning. For a host of learning problems, it has now been established that gradient-based methods converge to special estimators with attractive generalization properties [4,20,21,31–34,37,38,44,48,49]. Thus the limiting performance of gradient descent and its variants can often be characterized via careful analyses of these special limiting estimators (c.f. [8,19,28,35,36,43] and the references cited therein). However, an understanding of 'path-wise' properties of these algorithms still lies in its infancy. In this paper, we consider Langevin dynamics as a proxy for stochastic gradient descent, and a simple recovery problem with a non-convex objective function—that of recovering a planted rank 1 matrix under additive Gaussian noise. Admittedly, our setting is simple; however, it is a preliminary step in characterizing the inferential properties of specific noisy gradient based algorithms, a challenging problem in the current literature in general.

Formally, we observe a symmetric matrix $\mathbf{J} \in \mathbb{R}^{N \times N}$, given by

$$\mathbf{J} = VV^\top + \mathbf{Z}, \tag{1.1}$$

where $V \in \mathbb{R}^N$. We assume $\mathbf{Z} = \mathbf{G}^\top \mathbf{D} \mathbf{G}$, where $\mathbf{D}$ is the diagonal and $\mathbf{G}$ is Haar distributed. Denote $\mathbf{D} = \text{Diag}(\sigma^1, \ldots, \sigma^N)$ to be a sequence of deterministic diagonal matrices. As N increases, assume that

$\mu_{D_N} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\sigma^i}$ converges weakly to a probability measure μ_D with compact support. Let d_+, d_- denote the upper and lower edges respectively of μ_D. We assume throughout that $\max_{1 \leq i \leq N} \sigma^i$ and $\min_{1 \leq i \leq N} \sigma^i$ converge to d_+ and d_- respectively. We assume that the entries of $V = (V^i)$ are iid $\mathcal{N}(0, \frac{\lambda}{N})$ for some $\lambda > 0$. We seek to recover the planted truth V, given the observation $\mathbf{J}$. The natural estimator in this setting is derived from PCA—one computes the eigenvector $\hat{v}$ corresponding to the largest eigenvalue of $\mathbf{J}$, and uses $\hat{v}\hat{v}^\top$ as an estimator for the latent subspace $VV^\top$. The performance of this estimator can be precisely characterized using recent advances in the Random Matrix Theory [7,14]. In particular, assume that the empirical spectral measure μ_{D_N} converges weakly to a limiting measure μ_D supported on $[d_-, d_+]$. Define $G_{\mu_D} : \mathbb{R} \backslash [d_-, d_+] \to \mathbb{R}$,

$$G_{\mu_D}(z) = \mathbb{E}_{X \sim \mu_D}\Big[\frac{1}{z - X} \Big]$$

as the Cauchy transform of μ_D. Further, define

$$G_{\mu_D}(d^+) = \lim_{z \downarrow d_+} G_{\mu_D}(z), \quad G_{\mu_D}(d^-) = \lim_{z \uparrow d_-} G_{\mu_D}(z).$$

The BBP phase transition establishes that if $\lambda > \frac{1}{G_{\mu_D}(d^+)}$, there exists an 'outlier' eigenvalue at $G_{\mu_D}^{-1}(1/\lambda)$; otherwise, the largest sample eigenvalue sticks to the spectral edge d_+. In the first case, the eigenvector $\hat{v}$ corresponding to the outlying eigenvalue has a non-trivial overlap with the planted signal V, i.e. with high probability, $|V^\top \hat{v}| > 0$. In the latter case, $V^\top \hat{v} \xrightarrow{P} 0$.

To study the Langevin dynamics in this setting, we introduce the following system of Stochastic Differential Equations:

$$dX_t^i = \sum_{j=1}^{N} \mathbf{J}_{ij} X_t^j dt - f'\left(\frac{1}{N} \sum_{j=1}^{N} (X_t^j)^2 \right) X_t^i dt + \beta^{-1/2} dW_t^i . \tag{1.2}$$

We assume that $f : [0, \infty) \to \mathbb{R}$ satisfies f' to be non-negative and Lipschitz. Eventually, we will apply our results to the special case $f(x) = x^2/2$. Note that this SDE can be looked upon as a penalized version of the PCA problem on taking $\beta \to \infty$. The function f acts as a 'confining' potential, so that we can work without a norm constraint on the diffusion. For any $N \geq 1$ and $T \geq 0$, the SDE (1.2) has a strong solution, which we denote by $X_t \equiv \{X_t^i : 1 \leq i \leq N, t \in [0, T]\}$ (see [9][Lemma 6.7] for a proof).

The strong solution to the Langevin diffusion (1.2) defines a natural collection of estimators indexed by $t \in [0, \infty)$. We track the statistical performance of these estimators via the normalized 'overlap' $R^N(t) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{N} V^i X_t^i$. We note that the entries of V are typically of order $1/\sqrt{N}$; the multiplicative factor $\sqrt{N}$ normalizes these entries so that the resultant is of order 1. Armed with these notations, we can pose a natural question of interest:

How does the overlap $R^N(t)$ evolve over time?

In this paper, we provide sharp asymptotics for the evolution of $R^N(t)$ under a high-dimensional asymptotic limit where we send $N \to \infty$ with $t \geq 0$ *fixed*. In practice, one should interpret these limits as approximate characterizers of the overlap when the dimension N is large, while the diffusion has been run for a 'short time'. In particular, we emphasize that the time scales involved are significantly shorter than those involved for 'mixing' of these diffusion processes. We believe that

these asymptotics are particularly relevant for Statistics and Machine Learning. In particular, once the evolution of $R^N(t)$ is explicitly understood, given a fixed $\delta > 0$, one can characterize the first time T_δ such that $|R^N(T_\delta)| > \delta$. We believe that such guarantees can help practitioners understand the evolution of the *overlap* in this problem. The insights obtained can also be helpful for other problems.

Our first result characterizes the limiting behavior of $R^N(t)$. We will see that the behavior of R^N is intricately tied to that of the 'auto-correlation' function $K^N(t, s) = \frac{1}{N} \sum_{i=1}^N X_t^i X_s^i$. Note that for any $T \geq 0$, R^N and K^N are sequences of (random) continuous functions on $[0, T]$ and $[0, T]^2$, respectively. In subsequent discussions, we equip these metric spaces with the sup-norm topology.

In addition, we need to specify the initial conditions for the Langevin diffusion $\{\mathbf{X}_t : t \geq 0\}$. Our main results will hold under either one of the following two initial conditions:

Initial Conditions:

(i) (I.I.D. initial conditions) Assume that $\{(\sqrt{N}V^i, X_0^i) : 1 \leq i \leq N\}$ are i.i.d., independent of $\mathbf{G}$ and $\{\mathbf{W}_t : t \geq 0\}$. Assume furthermore that $\mathbb{E}[(X_0^i)^2] = 1$, $\mathbb{E}[V^i X_0^i] = \frac{\rho \sqrt{\lambda}}{\sqrt{N}}$ for some $\rho \in [0, 1]$.

(ii) (I.I.D. under rotated basis) Define $\mathbf{Y}_t = \mathbf{GX}_t$, $U = \mathbf{G}V$, $\mathbf{B}_t = \mathbf{GW}_t$. Assume that $\{(Y_0^i, \sqrt{N}U^i) : 1 \leq i \leq N\}$ are i.i.d., independent of $\{\mathbf{B}_t : t \geq 0\}$. Assume, in addition, that $Y_0^i \sim F$ i.i.d., $\mathbb{E}[(Y_0^i)^2] = 1$, $\mathbb{E}[U^i Y_0^i] = \frac{\rho \sqrt{\lambda}}{\sqrt{N}}$ for some $\rho \in [0, 1]$.

REMARK 1. Note that the parameter ρ governs the correlation between the initialization and the spike direction. For concreteness, we assume that $\rho \geq 0$. The results generalize directly to $\rho < 0$ if we replace V by $-V$. In addition, we require the second moments of X_0 and Y_0 to be finite. The precise value 1 is chosen merely for convenience.

REMARK 2. The I.I.D. initial condition is arguably the most natural initialization in Statistics and Machine Learning. The I.I.D. under rotated basis condition, although a bit less natural from this perspective, is simpler to analyze with the same theoretical tools. We think of the second initialization throughout as a warm-up, with the first initialization being of principal interest.

THEOREM 1. Assume that either Initial Condition (i) or Initial Condition (ii) holds. Fix $T \geq 0$. As $N \to \infty$, R^N and K^N converge almost surely to deterministic limits R and K, respectively. Furthermore, these limits are the unique solutions to the following system of integro-differential equations:

$$R(t) = \exp\left\{-\int_0^t f'(K(s))\mathrm{d}s\right\} \cdot \mathop{\mathbb{E}}_{(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{Y}_0) \sim \pi^\infty}\left[\exp(\boldsymbol{\sigma} t)\mathbf{Y}_0 \mathbf{u}\right]$$

$$+ \int_0^t \exp\left\{-\int_s^t f'(K(r))\mathrm{d}r\right\} R(s) \mathop{\mathbb{E}}_{(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{Y}_0) \sim \pi^\infty}\left[\exp(\boldsymbol{\sigma}(t-s))\mathbf{u}^2\right]\mathrm{d}s. \qquad (1.3)$$

$$K(t, s) = \beta^{-1}\int_0^{s \wedge t} \exp\left\{-\int_r^t f'(K(w))\mathrm{d}w - \int_r^s f'(K(w))\mathrm{d}w\right\} \mathop{\mathbb{E}}_{(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{Y}_0) \sim \pi^\infty}\left[\exp(\boldsymbol{\sigma}(t+s-2r))\right]\mathrm{d}r$$

$$+ \exp\left\{-\int_0^t f'(K(r))\mathrm{d}r - \int_0^s f'(K(r))\mathrm{d}r\right\} \cdot \mathop{\mathbb{E}}_{(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{Y}_0) \sim \pi^\infty}\left[\exp(\boldsymbol{\sigma}(t+s))\mathbf{Y}_0^2\right]$$

$$+ \exp\left\{ -\int_0^t f'(K(r))\mathrm{d}r \right\} \cdot \int_0^s \exp\left\{ -\int_r^s f'(K(w))\mathrm{d}w \right\} R(r) \mathop{\mathbb{E}}_{(\boldsymbol{\sigma},\mathbf{u},\mathbf{Y}_0)\sim\pi^\infty} \left[\exp(\boldsymbol{\sigma}\,(t+s-r))\mathbf{Y}_0\mathbf{u} \right] \mathrm{d}r$$

$$+ \exp\left\{ -\int_0^s f'(K(r))\mathrm{d}r \right\} \cdot \int_0^t \exp\left\{ -\int_r^t f'(K(w))\mathrm{d}w \right\} R(r) \mathop{\mathbb{E}}_{(\boldsymbol{\sigma},\mathbf{u},\mathbf{Y}_0)\sim\pi^\infty} \left[\exp(\boldsymbol{\sigma}\,(t+s-r))\mathbf{Y}_0\mathbf{u} \right] \mathrm{d}r$$

$$+ \int_0^t \mathrm{d}r_1 \int_0^s \mathrm{d}r_2 \left\{ \exp\left\{ -\int_{r_1}^t f'(K(w))\mathrm{d}w - \int_{r_2}^s f'(K(w))\mathrm{d}w \right\} R(r_1)R(r_2)\cdot \right.$$

$$\left. \mathop{\mathbb{E}}_{(\boldsymbol{\sigma},\mathbf{u},\mathbf{Y}_0)\sim\pi^\infty} \left[\exp(\boldsymbol{\sigma}\,(t+s-r_1-r_2))\mathbf{u}^2 \right] \right\}. \tag{1.4}$$

Here we use the abbreviated notation $K(t) := K(t,t)$. It remains to specify the distribution π^∞; this limit depends on the initial condition:

(i)   Under I.I.D. initial conditions, $\pi^\infty = \mu_D \otimes \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \begin{pmatrix} \lambda, \rho\sqrt{\lambda} \\ \rho\sqrt{\lambda}, 1 \end{pmatrix}$

(ii)   Under I.I.D. under rotated basis initialization, $\pi^\infty = \mu_D \otimes \mathscr{P}_2$, where $\mathscr{P}_2$ denotes the law of a bivariate random vector (L_1, L_2) such that $L_1 \sim \mathcal{N}(0, \lambda), L_2 \sim F$, with $E[L_1 L_2] = \rho\sqrt{\lambda}, E[L_2^2] = 1$.

At this point, some interpretations might be useful. In Theorem (1), $\boldsymbol{\sigma}$ denotes a random variable with law given by the limiting empirical spectral distribution of $\boldsymbol{D}$; $\mathbf{u}$ (resp. $\mathbf{Y}_0$) denotes a similar limiting random variable corresponding to the empirical distribution that puts mass $1/N$ to the N entries in $\sqrt{N}U^i = \sqrt{N}\boldsymbol{G}V^i$ (resp. $Y_0^i = \boldsymbol{G}X_0^i$), $1 \leq i \leq N$, (recall that these were first defined under Initial Condition (ii)). Since $\boldsymbol{G}$ is Haar distributed, we obtain the aforementioned joint laws of $\boldsymbol{u}, \mathbf{Y}_0$ under the two different initial conditions. Note that the covariance between $\mathbf{u}, \mathbf{Y}_0$ remains the same as that between V^i and X_0^i, and their joint distribution naturally depends on ρ and λ.

Theorem (1) holds for any $t \in [0, T]$. This allows one to characterize the overlap at any time t, for sufficiently large N. However, though the theorem provides a precise characterization of $R^N(t)$ under a high-dimensional limit, it is a priori unclear whether this description yields an explicit understanding regarding the behavior of $R^N(t)$. This is primarily due to the mathematically involved nature of the fixed point equations (1.3)–(1.4). To gain a better understanding of the Crisanti–Horner–Sommers–Cugliandolo–Kurchan (CHSCK) equations, our next theorem illustrates that Theorem (1) can yield explicit results on the limiting correlation between the output $\mathbf{X}_t$ and the latent vector $\sqrt{N}V$ under the double limit $t \to \infty$, following $N \to \infty$. To this end, first note that this correlation can be captured through the ratio $R^2(t)/(\lambda K(t,t))$, since

$$\frac{\left(\sum_{i=1}^N \sqrt{N}V^i X_t^i \right)^2}{[\sum_{i=1}^N (X_t^i)^2][N\sum_{i=1}^N (V^i)^2]} = \frac{(R^N(t))^2}{[K^N(t,t)][\sum_{i=1}^N (V^i)^2]}.$$

and $\|V\|_2^2 \to \lambda$ almost surely under our initial conditions.

To precisely characterize the limiting correlation, we specialize to the following setting: consider $f(x) = x^2/2$ and μ_D to be the scaled semi-circle distribution, supported on $[-\sigma_\star, \sigma_\star]$, for some $\sigma_\star > 0$. This corresponds to a setting where the additive noise $\mathbf{Z}$ in (1.1) is a symmetric Gaussian matrix. Note that $f(x) = x^2/2$ satisfies the regularity conditions required for Theorem 1. Formally, the semi-circle law

on $[-\sigma_\star, \sigma_\star]$ has a density

$$\frac{\mathrm{d}\mu_D}{\mathrm{d}x} = \frac{2}{\pi\sigma_\star^2}\sqrt{\sigma_\star^2 - x^2}, \quad -\sigma_\star \leq x \leq \sigma_\star. \tag{1.5}$$

Let $\mathscr{S} \colon \mathbb{R}\backslash[-\sigma_\star, \sigma_\star] \to \mathbb{R}$ denote the Stieljes transform of μ_D, i.e.,

$$\mathscr{S}(z) = \mathop{\mathbb{E}}_{\boldsymbol{\sigma} \sim \mu_D}\left[\frac{1}{z - \boldsymbol{\sigma}}\right] = \frac{2}{z + \sqrt{z^2 - \sigma_\star^2}}. \tag{1.6}$$

THEOREM 2. Assume $\lambda > \sigma_\star/2$, and set $\tilde{\lambda} = \lambda + \frac{\sigma_\star^2}{4\lambda}$. If $\beta < \frac{1}{\sigma_\star^2}$, the equation $z = \beta^{-1}\mathscr{S}(z/2)$ has two real roots. Set s_β to be the largest real root of this equation if $\beta < \frac{1}{\sigma_\star^2}$, otherwise set $s_\beta = 2\sigma_\star$.

  (i) If $2\tilde{\lambda} < s_\beta$ or $\rho = 0$, $\lim_{t\to\infty}\frac{R(t)^2}{\lambda K(t,t)} = 0$.

  (ii) If $2\tilde{\lambda} > s_\beta$ and $\rho > 0$,

$$\lim_{t\to\infty}\frac{R(t)^2}{\lambda K(t,t)} = \left(1 - \frac{\beta^{-1}}{2\lambda(\lambda + \sigma_\star^2/4\lambda)}\right)\left(1 - \frac{\sigma_\star^2}{4\lambda^2}\right). \tag{1.7}$$

Several remarks are in order regarding Theorem 2. Note that it is information theoretically impossible to recover the planted vector V below the BBP threshold (i.e., when $\lambda < \sigma_\star/2$) [7]. Thus we focus on the interesting regime $\lambda > \sigma_\star/2$. In this setting, Theorem 2 can be interpreted as follows: first, if $\rho = 0$ i.e., if the initialization has $o(1)$ correlation, then the limiting correlation stays at zero. We note that our asymptotics is *different* from non-asymptotic results (in N) that allow vanishing (in N) overlap of the initialization with the truth. This is specifically the notion considered in [13], [11]. We defer a detailed comparison of our results with this recent line of work to Section 1.1. Our result establishes that if the diffusive noise is large ($2\tilde{\lambda} < s_\beta$), the initial correlation washes off, and the correlation converges to zero in the limit $t \to \infty$. On the other hand, if the diffusive noise is small ($2\tilde{\lambda} > s_\beta$), we obtain a non-trivial correlation in the limit.

We set $\lambda^\star(\beta)$ to denote the smallest λ (as a function of β) for weak recovery. Direct computation yields

$$\lambda^\star(\beta) = \sqrt{\frac{\beta^{-1}}{2} - \frac{\sigma_\star^2}{4}} \vee \frac{\sigma_\star}{2}.$$

To further interpret this threshold, we consider the two regimes

  (1) Low-temperature regime $\beta^{-1} \in [0, \sigma_\star^2]$. In this case, the transition point is the same as the classical BBP phase transition. However, the limiting correlation is worse than the $\beta = \infty$ case, by a multiplicative factor depending on the injected Langevin noise, shown in Theorem 2.

  (2) High-temperature regime $\beta^{-1} \in (\sigma_\star^2, \infty)$. In this case, the transition point is strictly larger than the classical BBP phase transition.

To develop a physical understanding for this finite temperature transition, note that there are two ways in which the noise can wipe out the signal (which in turn defines the transition $\lambda^\star(\beta)$): (1) the
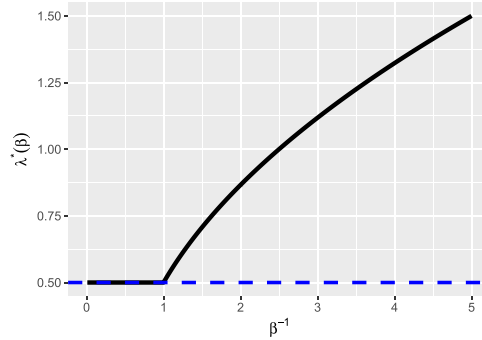
FIG. 1. We plot the minimum signal $\lambda^*(\beta)$ required for signal recovery, as a function of β. We fix $\sigma_\star = 1$ for this plot. For $\beta^{-1} \leq \sigma_\star^2$, the minimum signal required is $\lambda^*(\beta) = \frac{\sigma_\star}{2}$, which matches the BBP threshold (indicated by the dashed line in blue).

For $\beta^{-1} > \sigma_\star^2$, the minimum signal required is $\lambda^*(\beta) = \sqrt{\frac{\beta^{-1}}{2} - \frac{\sigma_\star^2}{4}}$, which is strictly larger than the BBP threshold.

normalized overlap function $R^N(t)$ retains no signal as $\lambda < \frac{\sigma_\star}{2}$, or (2) the norm of the solution/auto-covariance $K^N(t,t)$, which is affected by the Langevin noise level β^{-1}. The injected Langevin noise will not affect the normalized overlap function $R^N(t)$ as $N \to \infty$ for any $\beta^{-1} \neq 0$, yet it will increase the norm of the solution/auto-covariance $K^N(t,t)$.

Consider the case $\lambda > \frac{\sigma_\star}{2}$ and β^{-1} close to 0. In this regime, when β^{-1} is small, and λ is above the BBP phase transition, the growth of the norm $K^N(t)$ will not wipe out the signal, thus explaining why the transition point on λ stays the same as that in BBP; however, the limiting correlation will be smaller than the case of $\beta^{-1} = 0$. Now consider the regime β^{-1} very large: in this case, the growth of the norm $K^N(t,t)$ is asymptotically determined by s_β, which will wipe out the signal component in the overlap $R^N(t)$ even if $\lambda > \frac{\sigma_\star}{2}$, namely $R^2(t)/K(t,t) \to 0$. We illustrate this threshold in Fig. 1.

While our setup is quite simple, we can precisely quantify the tradeoff between the SNR (captured by λ) and the strength of the injected noise (captured by β^{-1}) in the limiting correlation—we consider this to be one of the main contributions of our work. In particular, the limiting correlation increases as a function of λ, as well as β, as one would naturally expect.

1.1   *Related literature*

(i) Dynamical Mean Field Theory and CHSCK equations: Dynamical Mean-Field Theory originated in the theory of spin glasses in the 80s [46,47]. In this approach, dynamics are characterized in terms of the 'correlation' and 'response' functions. In special cases, these functions satisfy a system of integro-differential equations [22,23] (we refer to such systems as CHSCK equations henceforth for simplicity). In general settings, the effective dynamics is described in terms of a non-Markovian stochastic process with long-term memory.

Recently, this framework has been employed in the statistical physics literature to study high-dimensional inference problems such as Gaussian mixture models, max-margin classification and tensor PCA [1,17,39–42,45]. In these papers, versions of CHSCK equations have been proposed, and analyzed numerically to track the performance of specific algorithms. Our work differs from these earlier inquiries in some crucial ways—first, the derivations of the CHSCK equations in these works are non-rigorous, and the subsequent analysis is also numerical. In sharp contrast, our results are fully rigorous, furthermore, we characterize the precise tradeoff

between the SNR and the injected noise in the Langevin algorithm. However, it should be noted that our model is, in some sense, simpler than the other models described above. The main technical difference is that the spiked matrix model does not require a 'response function'—this considerably simplifies the CHSCK system, and the subsequent analysis.

(ii) Prior rigorous results: Dynamical Mean-field Theory for mean-field spin glasses was established on rigorous footing in the works of [9], [5,6], [29] and [30], among others. The CHSCK equations were formally derived in [10]. While some useful information could be extracted from these equations under special settings (e.g. Langevin dynamics for matrix models [9] and at high-temperature [25]), a general analysis of these equations has been quite challenging.

Recently, there has been renewed interest in this area. [27] derived the CHSCK equations for spherical spin glasses starting from disorder dependent initial conditions. [26] examined the universality of these equations to the law of the disorder variables. [24] also introduced alternative techniques for establishing universality of such dynamical algorithms.

We note that in sharp contrast to our setting, the models analyzed in this line of work correspond to 'null' models, i.e. without any planted signal. As a result, our analysis is not directly comparable to the aforementioned papers. Despite this difference, our derivation of the CHSCK equations and its subsequent analysis relies partially on techniques from [9], which studies the Langevin dynamics in matrix models in the absence of a spike. That said, we emphasize that planted models are closer to models typically observed in Statistics and Machine Learning problems. Thus, despite the simplicity of (1.1), Theorem (2) provides useful insights regarding the interplay between the SNR and the noise magnitude in determining the exact value of the limiting correlation. To the best of our knowledge, our work is the first to characterize this tradeoff for a planted model. We hope that this precise analysis would spark further investigations into Langevin-type dynamics for more complex planted models, as seen in contemporary Statistics and Machine Learning problems.

(iii) Recent flow-based analyses: There has been considerable interest in understanding gradient descent algorithms (i.e. $\beta = \infty$ dynamics) in Statistics and Machine Learning. [15] and [16] derive novel systems of integro-differential equations for gradient descent dynamics for certain spiked models. Although the equations are similar in spirit, there are crucial differences between these results, and the ones proved in this paper. First, for the spiked matrix model, the result in [16] applies only to the $\beta = \infty$ case, and recovers the BBP phase transition. In contrast, we discover additional phase transitions depending on the strength of the injected noise. Further, the proof techniques are also completely different—[15] and [16] use recent advances in the random matrix literature on Green functions to derive their results. In a different line of work, [3] and [2] have also used random matrix asymptotics to study the role of 'early stopping' for gradient descent algorithms for linear regression. In this problem, the solution at any fixed time is available in closed form, which considerably simplifies the analysis.

Recently, [18] derived the CHSCK equations for gradient descent (i.e. $\beta = \infty$) for empirical risk minimization. The approaches in these two papers are completely different—[18] first use Approximate Message Passing style algorithms to study first order methods, and subsequently take a continuous time limit. Our approach, on the other hand, is grounded in random matrix theory.

(iv) We note that our results assume a 'warm-start', i.e. an $O(1)$-correlation between the initialization and the planted signal. This situation is different from a random initialization, where the initialization has $O(1/\sqrt{N})$ correlation with the planted signal. The approach pursued in

this paper does not seem suited to analyze the evolution of the correlation in this regime. Some answers are provided by the recent theory of 'bounding flows' [12] and the subsequent applications of this machinery to planted models [13], [11]. The main restriction of this approach is that it yields 'non-sharp' answers, in constrast to the approach outlined in our paper.

Outline: The rest of the paper is structured as follows: we prove Theorem 2 in Section 2, while the proof of the CHSCK equations Theorem 1 is deferred to Section 3.

2. Proof of Theorem 2

Throughout this section, for notational convenience, we use the shorthand $K(t) = K(t, t)$. Let $E(t) := \exp\{\int_0^t f'(K(s))\mathrm{d}s\}$. For our subsequent analysis, it will be convenient to transform the functions $R(t)$, $K(t)$ into a new set of functions $g(t)$, $h(t)$, defined as follows:

$$g(t) := E(t)R(t)$$
$$h(t) := E^2(t)K(t). \tag{2.1}$$

Observe that $(R(t))^2/K(t) = (g(t))^2/h(t)$, and thus it suffices to track $g(t)$, $h(t)$. In turn, we observe that (1.3) and (1.4) imply that $g(t)$, $h(t)$ are uniquely specified as the solutions to the following fixed-point system:

$$g(t) = \mathbb{E}[\mathbf{Y}_0\mathbf{u}] \cdot \mathbb{E}[e^{t\sigma}] + \mathbb{E}[\mathbf{u}^2] \cdot \int_0^t g(s)\mathbb{E}[e^{(t-s)\sigma}]\mathrm{d}s, \tag{2.2}$$

$$h(t) = \beta^{-1} \int_0^t E^2(s)\mathbb{E}[e^{(2t-2s)\sigma}]\mathrm{d}s + \mathbb{E}[\mathbf{Y}_0^2] \cdot \mathbb{E}[e^{2t\sigma}] \tag{2.3}$$

$$+ 2\mathbb{E}[\mathbf{Y}_0\mathbf{u}] \cdot \int_0^t g(s)\mathbb{E}[e^{(2t-s)\sigma}]\mathrm{d}s + \mathbb{E}[\mathbf{u}^2] \cdot \int_0^t\int_0^t g(s_1)g(s_2)\mathbb{E}[e^{(2t-s_1-s_2)\sigma}]\mathrm{d}s_1 \, \mathrm{d}s_2. \tag{2.4}$$

Our first lemma characterizes the behavior of $g(\cdot)$.

LEMMA 3. With σ drawn from the semi-circle distribution with parameter $\sigma_\star$, that is, with density (1.5), the function $g(\cdot)$ defined in (2.1) satisfies the following:

$$g(t) = \sqrt{\lambda}\rho \left\{ \left(1 - \frac{\sigma_\star^2}{4\lambda^2}\right)_+ \exp\left\{(\lambda + \frac{\sigma_\star^2}{4\lambda})t\right\} + \frac{1}{2\pi\lambda} \int_{-\sigma_\star}^{\sigma_\star} \frac{e^{xt}\sqrt{\sigma_\star^2 - x^2}}{(\lambda + \frac{\sigma_\star^2}{4\lambda}) - x} \, \mathrm{d}x \right\}, \tag{2.5}$$

where $x_+ = \max\{x, 0\}$.

Proof. (Proof of Lemma 3) Note that the last term in the RHS of (2.2) contains a convolution, suggesting that the Laplace transform of $g(\cdot)$ should provide useful information. For $\mathrm{Re}(z) > \sigma_\star$, recall that the Laplace transform of g is given by

$$\mathcal{L}_g(z) := \int_0^\infty e^{-zt}g(t)\mathrm{d}t, \tag{2.6}$$

and the Stieltjes transform of μ_D is given by

$$\mathscr{S}(z) = \mathbb{E}\Big[\frac{1}{z - \boldsymbol{\sigma}}\Big]. \tag{2.7}$$

Plugging in the formula for $g(\cdot)$ from (2.2), we obtain that (2.6) can be decomposed into two terms, where the first one equals

$$\mathbb{E}[\mathbf{Y}_0\mathbf{u}] \int_0^\infty e^{-zt}\,\mathbb{E}[e^{t\sigma}]dt = \mathbb{E}[\mathbf{Y}_0\mathbf{u}] \int_0^\infty \mathbb{E}[e^{t(\sigma-z)}]dt = \mathbb{E}[\mathbf{Y}_0\mathbf{u}]\mathscr{S}(z).$$

The last equality above follows by exchanging the order of integration using Fubini's theorem. This is justified since

$$\int_{(0,\infty)\times[-\sigma_\star,\sigma_\star]} |e^{t(\sigma-z)}|d(t,\sigma) \le \int_{(0,\infty)\times[-\sigma_\star,\sigma_\star]} |e^{-t(z_1-\sigma)}|d(t,\sigma),$$

where $z = z_1 + iz_2$. Since $z_1 > \sigma_\star$, the above can be upper bounded by $2\sigma_\star \int_0^\infty e^{-t(z_1-\sigma_\star)}dt < \infty$. By analogous arguments, the term obtained by plugging in the second term from (2.2) into (2.6) is given by $E[\mathbf{u}^2]\mathscr{L}_g(z)\mathscr{S}(z)$. Thus, we have $\mathscr{L}_g(z) = \mathbb{E}[\mathbf{Y}_0\mathbf{u}]\mathscr{S}(z) + E[\mathbf{u}^2]\mathscr{L}_g(z)\mathscr{S}(z)$, which yields

$$\mathscr{L}_g(z) = \frac{\mathbb{E}[\mathbf{Y}_0\mathbf{u}] \cdot \mathscr{S}(z)}{1 - \mathbb{E}[\mathbf{u}^2] \cdot \mathscr{S}(z)}. \tag{2.8}$$

By the Fourier–Mellin formula (inverse Laplace transform), we know that

$$g(t) = \frac{1}{2\pi i} \lim_{T\to+\infty} \int_{\gamma-iT}^{\gamma+iT} e^{zt}\mathscr{L}_g(z)dz. \tag{2.9}$$

When μ_D follows a semi-circle law, using (1.6), we have

$$\mathscr{L}_g(z) = \frac{2\sqrt{\lambda}\rho}{z + \sqrt{z^2 - \sigma_\star^2 - 2\lambda}}. \tag{2.10}$$

Fix $\gamma \in \mathbb{R}$ with $\gamma > \lambda + \frac{\sigma_\star^2}{4\lambda}$, we need to evaluate the Fourier–Mellin integral of the form

$$g_R(t) := \frac{1}{2\pi i} \int_{\gamma-iR}^{\gamma+iR} e^{zt}\mathscr{L}_g(z)\,\mathrm{d}z$$

$$= \frac{1}{2\pi i} \int_{S_1} e^{zt}\mathscr{L}_g(z)\,\mathrm{d}z.$$

In other words, we need to know the integral along the line segment S_1 in Fig. 2. First, we observe that the function

$$f_t(z) := e^{zt}\mathscr{L}_g(z) = e^{zt}\frac{2\sqrt{\lambda}\rho}{z + \sqrt{z^2 - \sigma_\star^2 - 2\lambda}} \tag{2.11}$$
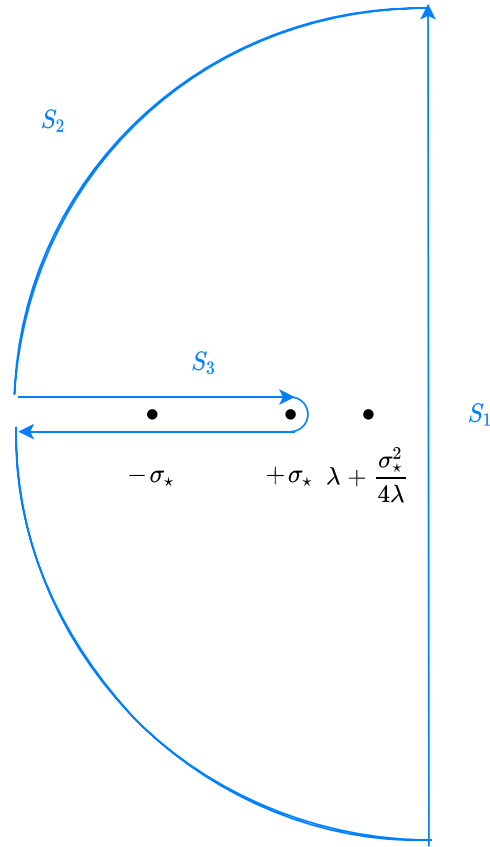
Fig. 2. The contour integral for evaluating the Fourier–Mellin formula.

$$= e^{zt} \frac{2\sqrt{\lambda}\rho}{4\lambda} \frac{\sqrt{z^2 - \sigma_\star^2} - (z - 2\lambda)}{z - (\lambda + \frac{\sigma_\star^2}{4\lambda})} \tag{2.12}$$

has a simple pole at $z = \lambda + \frac{\sigma_\star^2}{4\lambda}$. Therefore, using the Cauchy Residue Theorem, we have

$$\frac{1}{2\pi i} \oint_{S_1 + S_2 + S_3} f_t(z)\, dz = \sqrt{\lambda}\rho \cdot \left(1 - \frac{\sigma_\star^2}{4\lambda^2}\right)_+ e^{(\lambda + \frac{\sigma_\star^2}{4\lambda})t}. \tag{2.13}$$

$\square$

Second, we evaluate the integral over the arc S_2^ϵ. Here the arc is defined as $S_2^\epsilon = S_2 \cap \{z(\theta) : z(\theta) = \gamma - (R\cos\theta + iR\sin\theta), \theta \in [\frac{\pi}{2} + \epsilon, \frac{3\pi}{2} - \epsilon]\}$, where we parametrize the $z(\theta) \in S_2$ by $z(\theta) =$

$\gamma - (R\cos\theta + iR\sin\theta)$ (part of S_2) with $\theta \in [\frac{\pi}{2} + \epsilon, \frac{3\pi}{2} - \epsilon]$. Therefore,

$$\left| \frac{1}{2\pi i} \int_{S_2^\epsilon} f_t(z)\,dz \right| \leq \frac{R}{2} e^{t(\gamma - R\cos(\epsilon))} \sup_{\theta \in [\frac{\pi}{2}+\epsilon, \frac{3\pi}{2}-\epsilon]} \frac{2\sqrt{\lambda}\rho}{|\gamma + Re^{i\theta} + \sqrt{(\gamma + Re^{i\theta})^2 - \sigma_\star^2 - 2\lambda}|} \tag{2.14}$$

and thus for any fixed ϵ,

$$\lim_{R \to \infty} \frac{1}{2\pi i} \int_{S_2^\epsilon} f_t(z)\,dz = 0 . \tag{2.15}$$

For the part of the integral $S_2 \backslash S_2^\epsilon$, one can show that

$$\left| \frac{1}{2\pi i} \int_{S_2 \backslash S_2^\epsilon} f_t(z)\,dz \right| \leq \epsilon \frac{R}{\pi} e^{t\gamma} \sup_{\theta \in [\frac{\pi}{2}, \frac{\pi}{2}+\epsilon]} \frac{2\sqrt{\lambda}\rho}{|\gamma + Re^{i\theta} + \sqrt{(\gamma + Re^{i\theta})^2 - \sigma_\star^2 - 2\lambda}|} \tag{2.16}$$

and thus with some universal constant $C(\lambda, \sigma_\star, \rho, \gamma)$

$$\lim_{R \to \infty} \frac{1}{2\pi i} \int_{S_2^\epsilon} f_t(z)\,dz = C(\lambda, \sigma_\star, \rho, \gamma) \cdot \epsilon . \tag{2.17}$$

Putting these estimates together and sending $\epsilon \to 0$, we have shown

$$\lim_{R \to \infty} \frac{1}{2\pi i} \int_{S_2} f_t(z)\,dz = 0 \tag{2.18}$$

Third, we evaluate the integral over the segment S_3 (corresponding to the branch point of $\sqrt{z^2 - \sigma_\star^2}$), with $z = x + i\epsilon$ where $x \in [-\sigma_\star, \sigma_\star]$. We consider this since the function $z \mapsto \sqrt{z}$ is a well-defined single-valued function only on $z \in \mathbb{C}\backslash(-\infty, 0)$. One can show that

$$\lim_{\epsilon \to 0} \frac{1}{2\pi i} \int_{S_3^\epsilon} f_t(z)\,dz \tag{2.19}$$

$$= \frac{1}{2\pi i} \int_{-\sigma_\star}^{\sigma_\star} e^{xt} \frac{2\sqrt{\lambda}\rho}{(x - 2\lambda) + i\sqrt{\sigma_\star^2 - x^2}}\,dx - \frac{1}{2\pi i} \int_{-\sigma_\star}^{\sigma_\star} e^{xt} \frac{2\sqrt{\lambda}\rho}{(x - 2\lambda) - i\sqrt{\sigma_\star^2 - x^2}}\,dx \tag{2.20}$$

$$= \frac{2\sqrt{\lambda}\rho}{2\pi i} \int_{-\sigma_\star}^{\sigma_\star} e^{xt} \frac{-2i\sqrt{\sigma_\star^2 - x^2}}{-4\lambda x + 4\lambda^2 + \sigma_\star^2}\,dx \tag{2.21}$$

$$= -\sqrt{\lambda}\rho \frac{1}{2\pi\lambda} \int_{-\sigma_\star}^{\sigma_\star} e^{xt} \frac{\sqrt{\sigma_\star^2 - x^2}}{(\lambda + \frac{\sigma_\star^2}{4\lambda}) - x}\,dx . \tag{2.22}$$

Finally, we integrate over the semi-circular arc in S_3. Parametrizing $z = \sigma_\star + \epsilon e^{i\theta}$, we have $\mathrm{d}z = \epsilon i e^{i\theta} \mathrm{d}\theta$. In turn, the integral reduces to

$$\int_{\pi/2}^{-\pi/2} e^{(\sigma_\star + \epsilon e^{i\theta})t} \frac{2\sqrt{\lambda}\rho}{\sigma_\star + \epsilon e^{i\theta} + \sqrt{(\sigma_\star + \epsilon e^{i\theta})^2 - \sigma_\star^2 - 2\lambda}} \epsilon i e^{i\theta} \mathrm{d}\theta.$$

We note that this integral converges to zero as $\epsilon \to 0$. Putting three pieces together, we conclude the proof.

We next turn our attention to h. Recall $E(t) = \exp\{\int_0^t K(s)\mathrm{d}s\}$, and set $F(t) = E^2(t)$. Differentiating, we obtain that $F'(t) = 2K(t)F(t) = 2h(t)$, defined in (2.1). Coupled with (2.4), we have, F satisfies the integro-differential equation,

$$F'(t) = 2\beta^{-1} \int_0^t F(s)\mathbb{E}\Big[\exp(2(t-s)\boldsymbol{\sigma})\Big]\mathrm{d}s + \Phi(t),$$

$$\Phi(t) = 2\mathbb{E}[\boldsymbol{Y_0}^2]\mathbb{E}[\exp(2t\boldsymbol{\sigma})] + 4\mathbb{E}[\boldsymbol{Y_0 u}]\int_0^t g(s)\mathbb{E}[\exp(\boldsymbol{\sigma}(2t-s))]\mathrm{d}s \qquad (2.23)$$

$$+2\mathbb{E}[\boldsymbol{u}^2]\int_0^t \int_0^t g(s_1)g(s_2)\mathbb{E}\Big[\exp\big(\boldsymbol{\sigma}(2t - s_1 - s_2)\big)\Big]\mathrm{d}s_1\mathrm{d}s_2.$$

LEMMA 4. The function h has Laplace transform

$$\mathscr{L}_h(z) = \frac{1}{2}\Big[\frac{z\mathscr{L}_\Phi(z) + \beta^{-1}\mathscr{S}(z/2)}{z - \beta^{-1}\mathscr{S}(z/2)}\Big].$$

Proof of Lemma 4. Taking Laplace transforms in (2.23), we obtain

$$z\mathscr{L}_F(z) - 1 = 2\beta^{-1}\mathscr{L}_F(z)\mathscr{L}_M(z) + \mathscr{L}_\Phi(z),$$

where we set $M(t) = \mathbb{E}[\exp(2t\sigma)]$. Transposing, we obtain

$$\mathscr{L}_F(z) = \frac{1 + \mathscr{L}_\Phi(z)}{z - 2\beta^{-1}\mathscr{L}_M(z)}.$$

Recall that $F'(t) = 2h(t)$, and thus $\mathscr{L}_h(z) = \frac{1}{2}[z\mathscr{L}_F(z) - 1]$. This allows us to obtain the Laplace transform of h. Finally, we note that by direct computation, $\mathscr{L}_M(z) = \frac{1}{2}\mathscr{S}(z/2)$. This completes the calculation. □

We set $\tilde{\lambda} = \lambda + \frac{\sigma_\star^2}{4\lambda}$. Let s_β denote the largest real solution to the equation $z = \beta^{-1}\mathscr{S}(z/2)$.

LEMMA 5. If $2\tilde{\lambda} > s_\beta$, we have that

$$\lim_{z \to 0} \left|z\mathscr{L}_\Phi(2\tilde{\lambda} + z) - 2\lambda^2\rho^2\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)_+^2 \mathbb{E}\left[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\right]\right| = 0.$$

Proof. By definition,

$$\mathscr{L}_\Phi(2\tilde{\lambda} + z) = \int_0^\infty \exp(-(2\tilde{\lambda} + z)t)\Phi(t)\mathrm{d}t := T_1 + T_2 + T_3,$$

$$T_1 := 2\mathbb{E}[Y_0{}^2] \int_0^\infty \exp(-(2\tilde{\lambda} + z)t)\mathbb{E}[\exp(2t\sigma)]\mathrm{d}t,$$

$$T_2 := 4\mathbb{E}[Y_0 u] \int_0^\infty \int_0^t \exp(-(2\tilde{\lambda} + z)t)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\mathrm{d}s\mathrm{d}t,$$

$$T_3 := 2\mathbb{E}[u^2] \int_0^\infty \int_0^t \int_0^t \exp(-(2\tilde{\lambda} + z)t)g(s_1)g(s_2)\mathbb{E}\Big[\exp\Big(\sigma(2t - s_1 - s_2)\Big)\Big]\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t.$$

We first claim that there exists a constant $C > 0$ such that

$$\Big|\int_0^\infty \exp(-(2\tilde{\lambda} + z)t)\mathbb{E}[\exp(2t\sigma)]\mathrm{d}t\Big| \leq C,$$

$$\Big|\int_0^\infty \int_0^t \exp(-(2\tilde{\lambda} + z)t)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\mathrm{d}s\mathrm{d}t\Big| \leq C.$$

We establish each bound in turn. Set $z = z_1 + iz_2$ with $z_1, z_2 \in \mathbb{R}$, and note that $z_1, z_2 \to 0$. Choose z_1 small enough so that $2\tilde{\lambda} + z_1 > 4\sigma_\star$, since $2\tilde{\lambda} > 4\sigma_\star$ by assumption. Thus, we have

$$\Big|\int_0^\infty \exp(-(2\tilde{\lambda} + z)t)\mathbb{E}[\exp(2t\sigma)]\mathrm{d}t\Big| \leq \int_0^\infty \exp(-(2\tilde{\lambda} + z_1)t)\mathbb{E}[\exp(2t\sigma)]\mathrm{d}t$$

$$\leq \int_0^\infty \exp\Big(-(2\tilde{\lambda} - 2\sigma_\star + z_1)t\Big)\mathrm{d}t$$

$$\leq \frac{1}{2\tilde{\lambda} - 2\sigma_\star + z_1}.$$

This remains bounded as $z_1 \to 0$. To analyze the second term, first observe from Lemma 3 that

$$\lim_{t\to\infty} \exp(-\tilde{\lambda}t)g(t) = \sqrt{\lambda}\rho\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)_+.$$

Thus, for any $\varepsilon > 0$, there exists t_0 large enough such that for all $t \geq t_0$,

$$(1 - \varepsilon)\sqrt{\lambda}\rho\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)_+ \exp(\tilde{\lambda}t) \leq g(t) \leq (1 + \varepsilon)\sqrt{\lambda}\rho\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)_+ \exp(\tilde{\lambda}t).$$

Then we have, setting $z = z_1 + iz_2$ with z_1 small enough,

$$\left| \int_0^\infty \int_0^t \exp(-(2\tilde{\lambda} + z)t)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\mathrm{d}s\mathrm{d}t \right|$$

$$\leq \int_0^\infty \int_0^t \exp(-(2\tilde{\lambda} + z_1)t)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\mathrm{d}s\mathrm{d}t$$

$$\leq C_1 + \int_{t_0}^\infty \Big[\int_0^{t_0} \exp(-(2\tilde{\lambda} + z_1)t)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\mathrm{d}s +$$

$$\int_{t_0}^t \exp(-(2\tilde{\lambda} + z_1)t)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\mathrm{d}s\Big]\mathrm{d}t.$$

The control of this term is complete once we control the two integrals above. To this end,

$$\int_{t_0}^\infty \int_0^{t_0} \exp\Big(-(2\tilde{\lambda} + z_1)t\Big)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\mathrm{d}s\mathrm{d}t$$

$$\leq \Big(\max_{0 \leq s \leq t_0} g(s) \Big) \int_{t_0}^\infty \exp\Big(-(2\tilde{\lambda} + z_1 - 2\sigma_\star)t\Big)\mathrm{d}t$$

$$\leq \Big(\max_{0 \leq s \leq t_0} g(s) \Big) \frac{\exp(-t_0(2\tilde{\lambda} + z_1 - 2\sigma_\star))}{2\tilde{\lambda} + z_1 - 2\sigma_\star}$$

which remains bounded as $z_1 \to 0$. Finally,

$$\int_{t_0}^\infty \int_{t_0}^t \exp\Big(-(2\tilde{\lambda} + z_1)t)g(s)\mathbb{E}[\exp(\sigma(2t - s))]\Big)\mathrm{d}s\mathrm{d}t$$

$$\leq (1 + \varepsilon)\sqrt{\lambda}\rho\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)_+ \int_{t_0}^\infty \int_{t_0}^t \exp\Big(-(2\tilde{\lambda} + z_1)t\Big) \exp(\tilde{\lambda}s + \sigma_\star(2t - s))\mathrm{d}s\mathrm{d}t$$

$$\leq \frac{(1 + \varepsilon)}{\tilde{\lambda} - \sigma_\star}\sqrt{\lambda}\rho\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)_+ \int_{t_0}^\infty \exp\Big(-(\tilde{\lambda} + z_1 - \sigma_\star)t\Big)\mathrm{d}t$$

which remains bounded as $z_1 \to 0$.

Finally, we turn to the term T_3. Recall that, using (2.5), $g(s) = \sqrt{\lambda}\rho(1 - \frac{\sigma_\star^2}{4\lambda^2})_+ \exp(\tilde{\lambda}s) + g_1(s)$, where

$$g_1(s) = \frac{\sigma_\star^2}{4\lambda}\mathbb{E}_{x \sim \mu_D}\left[\frac{e^{xs}}{\tilde{\lambda} - x} \right].$$

Further recalling that $\mathbb{E}[\boldsymbol{u}^2] = \lambda$, we have

$$T_3 = 2\mathbb{E}[\boldsymbol{u}^2] \int_0^\infty \int_0^t \int_0^t \exp(-(2\tilde{\lambda} + z)t)g(s_1)g(s_2)\mathbb{E}\Big[\exp\Big(\boldsymbol{\sigma}(2t - s_1 - s_2)\Big)\Big]\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t$$

$$= 2\lambda^2\rho^2\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)_+^2 \int_0^\infty \int_0^t \int_0^t \exp(-(2\tilde{\lambda} + z)t + \tilde{\lambda}(s_1 + s_2))\mathbb{E}\Big[\exp\Big(\boldsymbol{\sigma}(2t - s_1 - s_2)\Big)\Big]\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t$$

$$+ \text{Rem} \tag{2.24}$$

where Rem is the remainder term which we will later show to be bounded as $z \to 0$. Computing the first term, we obtain,

$$\int_0^\infty \int_0^t \int_0^t \exp(-(2\tilde{\lambda} + z)t + \tilde{\lambda}(s_1 + s_2))\mathbb{E}\Big[\exp\Big(\boldsymbol{\sigma}(2t - s_1 - s_2)\Big)\Big]\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t$$

$$= \mathbb{E}\Big[\int_0^\infty \int_0^t \int_0^t \exp\Big(-(2\tilde{\lambda} + z - 2\boldsymbol{\sigma})t + (\tilde{\lambda} - \boldsymbol{\sigma})s_1 + (\tilde{\lambda} - \boldsymbol{\sigma})s_2\Big)\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t\Big]$$

$$= \mathbb{E}\Big[\int_0^\infty \exp(-(2\tilde{\lambda} + z - 2\boldsymbol{\sigma})t)\frac{\Big(\exp(t(\tilde{\lambda} - \boldsymbol{\sigma})) - 1\Big)^2}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\mathrm{d}t\Big]$$

$$= \mathbb{E}\Big[\int_0^\infty \exp(-(2\tilde{\lambda} + z - 2\boldsymbol{\sigma})t)\frac{\Big(\exp(2t(\tilde{\lambda} - \boldsymbol{\sigma})) - 2\exp(t(\tilde{\lambda} - \boldsymbol{\sigma})) + 1\Big)}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\mathrm{d}t\Big]$$

$$= \frac{1}{z}\mathbb{E}\Big[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\Big] + \text{Rem}_1. \tag{2.25}$$

Note that Rem_1 involves two terms and remains bounded when $z \to 0$ since

$$\text{Rem}_1 = -2\mathbb{E}\Big[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\int_0^\infty \exp((-\tilde{\lambda} - z + \boldsymbol{\sigma})t)\Big] + \mathbb{E}\Big[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\int_0^\infty \exp((-2\tilde{\lambda} - z + 2\boldsymbol{\sigma})t)\Big]$$

$$= -2\mathbb{E}\Big[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2(\tilde{\lambda} + z - \boldsymbol{\sigma})}\Big] + \mathbb{E}\Big[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2(2\tilde{\lambda} + z - 2\boldsymbol{\sigma})}\Big]$$

Thus, the proof is complete on establishing that Rem remains bounded as $z \to 0$.

We note that Rem may be expressed as the sum of three terms, denoted as $\text{Rem}_{T_j}, j = 1, \cdots, 3$, where

$$\text{Rem}_{T_1} = 2\sqrt{\lambda}\rho(1 - \frac{\sigma_\star^2}{4\lambda^2})_+ \frac{\sigma_\star^2}{4\lambda}\int_0^\infty \int_0^t \int_0^t \exp(-(2\tilde{\lambda} + z)t + \tilde{\lambda}s_1)$$

$$\mathbb{E}_{x \sim \mu_D}\Big[\frac{\exp(xs_2)}{\tilde{\lambda} - x}\Big]\mathbb{E}\Big[\exp\Big(\boldsymbol{\sigma}(2t - s_1 - s_2)\Big)\Big]\mathrm{d}s_1\mathrm{d}s_2\mathrm{d}t,$$

Rem_{T_2} is the same as above except the roles of s_1 and s_2 are reversed, and the third term is defined as follows:

$$\mathrm{Rem}_{T_3} = \frac{2\sigma_\star^4}{16\lambda} \int_0^\infty \int_0^t \int_0^t \exp(-(2\tilde{\lambda} + z)t \mathbb{E}_{x_1 \sim \mu_D} \left[\frac{\exp(x_1 s_1)}{(\tilde{\lambda} - x_1)} \right]$$

$$\mathbb{E}_{x_2 \sim \mu_D} \left[\frac{\exp(x_2 s_2)}{(\tilde{\lambda} - x_2)} \right] \mathbb{E} \left[\exp\left(\boldsymbol{\sigma} (2t - s_1 - s_2) \right) \right] ds_1 ds_2 dt$$

Now the support of x, σ is upper bounded by $\sigma_\star$ and $2t - s_1 - s_2 \geq 0, s_1, s_2 \leq t$ in the range of integration. Thus, we may upper bound Rem_{T_1} by

Rem_{T_1}

$$\leq 2\sqrt{\lambda}\rho(1 - \frac{\sigma_\star^2}{4\lambda^2})_+ \mathbb{E}_{(x,\sigma) \sim \mu_D^{\otimes 2}} \left[\frac{1}{\tilde{\lambda} - x} \int_0^\infty \int_0^t \int_0^t \exp(-(2\tilde{\lambda} + z - 2\sigma_\star)t + (\tilde{\lambda} - \sigma_\star)t) ds_1 ds_2 dt \right]$$

$$= 2\sqrt{\lambda}\rho(1 - \frac{\sigma_\star^2}{4\lambda^2})_+ \mathbb{E}_{x \sim \mu_D} \left[\frac{1}{(\tilde{\lambda} - x)} \int_0^\infty t^2 \exp(-(\tilde{\lambda} + z - \sigma_\star)t) \right]$$

$$= 2\sqrt{\lambda}\rho(1 - \frac{\sigma_\star^2}{4\lambda^2})_+ \frac{\Gamma(3)}{(\tilde{\lambda} + z - \sigma_\star)^3} \mathbb{E}_{x \sim \mu_D} \left[\frac{1}{(\tilde{\lambda} - x)} \right],$$

which remains bounded when $z \to 0$. Similarly, Rem_{T_2} can be bounded. Applying a similar trick, Rem_{T_3} can be bounded by

$$\frac{2\sigma_\star^4}{16\lambda} \mathbb{E}_{(x_1,x_2) \sim \mu_D^{\otimes 2}} \left[\frac{1}{(\tilde{\lambda} - x_1)(\tilde{\lambda} - x_2)} \right] \int_0^\infty t^2 \exp(-(2\tilde{\lambda} + z - 2\sigma_\star)t) dt =$$

$$\frac{2\sigma_\star^4}{16\lambda} \frac{\Gamma(3)}{(2\tilde{\lambda} + z - 2\sigma_\star)^3} \mathbb{E}_{(x_1,x_2) \sim \mu_D^{\otimes 2}} \left[\frac{1}{(\tilde{\lambda} - x_1)(\tilde{\lambda} - x_2)} \right],$$

and this once again remains bounded as $z \to 0$. Putting everything together and from (2.25), we have the desired result. $\qquad \square$

We now turn to the proof of Theorem 2.

Proof of Theorem 2. First, using (1.6), we have

$$\lim_{z \downarrow \sigma_\star} \mathscr{S}(z) = \frac{2}{\sigma_\star}, \quad \lim_{z \uparrow -\sigma_\star} \mathscr{S}(z) = -\frac{2}{\sigma_\star}.$$

If z_* is a real root of $z = \beta^{-1}\mathscr{S}(z/2)$, $y = 2z_*$ is a real root of the fixed point equation $2\beta y = \mathscr{S}(y)$. In turn, such a root exists if and only if $2\beta\sigma_* < 2/\sigma_*$, which immediately gives us the desired conclusion.

Consider first the regime $2\tilde{\lambda} > s_\beta$. Observe that (2.5) implies

$$\lim_{t\to\infty} \exp(-2\tilde{\lambda}t)g^2(t) = \lambda\rho^2\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)^2_+.$$

Further, Lemma 5 implies that as $z \to 0$ along a sector,

$$z\mathscr{L}_\Phi(2\tilde{\lambda} + z) \to 2\lambda^2\rho^2\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)^2_+\mathbb{E}\left[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\right]$$

Thus, Lemma 4 implies

$$\lim_{z\to 0} z\mathscr{L}_h(2\tilde{\lambda} + z) = \frac{2\tilde{\lambda}}{2\tilde{\lambda} - \beta^{-1}\mathscr{S}(\tilde{\lambda})}\mathbb{E}\left[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\right]\lambda^2\rho^2\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)^2_+.$$

Thus using [9,Lemma 7.2],

$$\lim_{t\uparrow\infty} \exp(-2\tilde{\lambda}t)h(t) = \frac{2\tilde{\lambda}}{2\tilde{\lambda} - \beta^{-1}\mathscr{S}(\tilde{\lambda})}\mathbb{E}\left[\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\right]\lambda^2\rho^2\Big(1 - \frac{\sigma_\star^2}{4\lambda^2}\Big)^2_+.$$

Noting that $\rho > 0$, we have

$$\lim_{t\uparrow\infty} \frac{g^2(t)}{h(t)} = \frac{2\tilde{\lambda} - \beta^{-1}\mathscr{S}(\tilde{\lambda})}{2\tilde{\lambda}\lambda}\left[\mathbb{E}\left(\frac{1}{(\tilde{\lambda} - \boldsymbol{\sigma})^2}\right)\right]^{-1}.$$

To simplify this, first note that $\mathscr{S}'(z) = -\mathbb{E}\left(\frac{1}{(z-\boldsymbol{\sigma})^2}\right)$. Calculating the exact value of this derivative, we obtain that

$$-\mathbb{E}\left(\frac{1}{(z - \boldsymbol{\sigma})^2}\right) = \frac{2\sqrt{z^2 - \sigma_\star^2} - z}{\sigma_\star^2\sqrt{z^2 - \sigma_\star^2}}.$$

Equating these for $z = \tilde{\lambda}$ yields

$$\left[\mathbb{E}\left(\frac{1}{(\tilde{\lambda} - \sigma)^2}\right)\right]^{-1} = \lambda^2(1 - \frac{\sigma_\star^2}{4\lambda^2}).$$

Putting things together,

$$\lim_{t\uparrow\infty} \frac{g^2(t)}{\lambda h(t)} = \left(1 - \frac{\beta^{-1}\mathscr{S}(\tilde{\lambda})}{2\tilde{\lambda}}\right)\left(1 - \frac{\sigma_\star^2}{4\lambda^2}\right).$$

Now note that $\mathscr{S}(\tilde{\lambda}) = 1/\lambda$, so the final limit becomes

$$\lim_{t \to \infty} \frac{g^2(t)}{\lambda h(t)} = \left(1 - \frac{\beta^{-1}}{2\lambda(\lambda + \sigma_\star^2/4\lambda)}\right)\left(1 - \frac{\sigma_\star^2}{4\lambda^2}\right).$$

It remains to analyze the sub-critical regime $2\tilde{\lambda} < s_\beta$. Using Lemma 4, we note that $\mathscr{L}_h(\cdot)$ has a simple pole at s_β. Thus, there exists $C_1 > 0$ such that

$$\lim_{z \to 0} z\mathscr{L}_h(s_\beta + z) = C_1 \neq 0. \tag{2.26}$$

In turn, this implies

$$\lim_{t \uparrow \infty} \exp(-s_\beta t)h(t) = C_1.$$

This completes the proof in this sub-case, as $\lim_{t \to \infty} g^2(t)/h(t) = 0$ immediately in this case.

Finally, we focus on the case $\rho = 0$. Lemma 3 implies that $g = 0$ in this case. On the other hand, in this case, $\Phi(t) = 2\mathbb{E}[Y_0^2]\mathbb{E}[\exp(2t\sigma)]$. Thus, using Lemma 4, we note that the leading asymptotics of $h(\cdot)$ is determined by the pole at s_β. Specifically, $\lim_{t \uparrow \infty} h(t) \neq 0$, which concludes the proof. □

3. Proof of Theorem 1

Setting $U = \mathbf{G}V \in \mathbb{R}^N$, the dynamics under the rotated coordinate system $\mathbf{Y}_t := \mathbf{G}\mathbf{X}_t$ can be expressed as

$$dY_t^i = U^i\langle U, Y_t\rangle dt + \sigma^i Y_t^i dt - f'(\|Y_t\|^2/N)Y_t^i dt + \beta^{-1/2}dB_t^i. \tag{3.1}$$

Setting $u^i = \sqrt{N}U^i$, we note that Eqn. (3.1) reduces to

$$dY_t^i = u^i R^N(t)dt + (\sigma^i - f'(K^N(t)))Y_t^i dt + \beta^{-1/2}dB_t^i. \tag{3.2}$$

Utilizing this expression, one can verify that Y_t^i takes the following integral form:

$$Y_t^i = \exp\left\{\int_0^t [\sigma^i - f'(K^N(s))]ds\right\}\left\{Y_0^i + u^i \int_0^t \exp\left\{-\int_0^s [\sigma^i - f'(K^N(r))]dr\right\}R^N(s)ds\right\} \tag{3.3}$$

$$+ \beta^{-1/2}\int_0^t \exp\left\{\int_s^t [\sigma^i - f'(K^N(r))]dr\right\}dB_s^i. \tag{3.4}$$

We first re-express the solution of our SDE.

LEMMA 6. Define $F_t(K, \lambda) = f'(K(t, t)) - \lambda$. Then we have,

$$Y_t^i = Y_0^i \exp\Big[- \int_0^t F_s(K^N, \sigma^i) \mathrm{d}s \Big] + \int_0^t \exp\Big[- \int_s^t F_{s_1}(K^N, \sigma^i) \mathrm{d}s_1 \Big] u_i R^N(s) \mathrm{d}s + \beta^{-1/2} B_t^i -$$

$$\beta^{-1/2} \int_0^t B_s^i F_s(K^N, \sigma^i) \exp\Big[- \int_s^t F_{s_1}(K^N, \sigma^i) \mathrm{d}s_1 \Big] \mathrm{d}s . \tag{3.5}$$

Proof. The proof follows by an application of the integration by parts formula $\int_0^t f_s \mathrm{d}B_s = f_t B_t - \int_0^t B_s f'_s \mathrm{d}s$ on the solution (3.3). □

DEFINITION 1. We define the empirical measure

$$\nu = \frac{1}{N} \sum_i \delta_{Y_0^i, u^i, \sigma^i, B_\bullet^i}, \tag{3.6}$$

where the fourth marginal is an empirical distribution on the path space $\mathscr{C}[0, T]$.

Consider the following collections of functions, with domain space $\mathbb{R}^3 \times \mathscr{C}[0, T]$ and range space one of $\mathscr{C}[0, T]^j$ for $j = 1, 2, 3$:

$$\mathscr{F}_j \subset \{f : \mathbb{R}^3 \times \mathscr{C}[0, T] \to \mathscr{C}([0, T]^j)\}, \ j = 1, 2, 3, \ \text{with}$$

$$\mathscr{F}_1 = \{f_j, j = 1, \ldots, 5 : f_1(Y_0, u, \sigma, B_\bullet)(w) = u Y_0 \exp(w\sigma), f_2(\cdot)(w) = u^2 \exp(w\sigma), f_3(\cdot)(t) = u B_t,$$

$$f_4(\cdot)(w) = Y_0^2 \exp(w\sigma), f_5(\cdot)(w) = \sigma Y_0 \exp(w\sigma)\},$$

$$\mathscr{F}_2 = \{f_6, \ldots, f_9 : f_6(Y_0, u, \sigma, B_\bullet)(t, w) = u B_t \exp(w\sigma), f_7(\cdot)(t, w) = u B_t \sigma \exp(w\sigma),$$

$$f_8(\cdot)(t, s) = B_t B_s, f_9(\cdot)(s, w) = Y_0 B_s \exp(w\sigma)\},$$

$$\mathscr{F}_3 = \{f_{10}, f_{11}, f_{12} : f_{10}(Y, u, \sigma, B_\bullet)(s_1, s_2, w) = B_{s_1} B_{s_2} \exp(w\sigma),$$

$$f_{11}(\cdot)(s_1, s_2, w) = \sigma B_{s_1} B_{s_2} \exp(w\sigma), f_{12}(\cdot)(s_1, s_2, w) = \sigma^2 B_{s_1} B_{s_2} \exp(w\sigma)\}$$

Finally, define

$$\mathscr{F} = \mathscr{F}_1 \cup \mathscr{F}_2 \cup \mathscr{F}_3. \tag{3.7}$$

All functions in $\mathscr{F}$ have the same domain. To simplify notation, for elements in $\mathscr{F}_1$, we use $f(\cdot)(w)$ to mean $f(Y_0, u, \sigma, B_\bullet)$ evaluated at w for a generic point $(Y_0, u, \sigma, B_\bullet) \in \mathbb{R}^3 \times \mathscr{C}[0, T]$, where $0 \le w \le T$. Similarly for elements in $\mathscr{F}_2$ and $\mathscr{F}_3$.

Define $\mathscr{C}_N = \{\int f d\nu : f \in \mathscr{F}\}$ and note that for $f \in \mathscr{F}_j$, $\int f d\nu \in \mathscr{C}([0,T]^j)$, $j = 1, 2, 3$. We introduce the following convention: for a discrete set $S = \{s_1, \ldots, s_k\}$ and a function $f(\cdot)$, $f(S) = f(s_1, \ldots, s_k)$. The next result establishes that for all $N \geq 1$, (R^N, K^N) depend on ν through the statistics $\mathscr{C}_N$.

LEMMA 7. There exist functions

$$\Phi^{(1)} : \mathscr{C}[0,T] \times \mathscr{C}([0,T]^2) \times \mathscr{C}[0,T]^{|\mathscr{F}_1|} \times \mathscr{C}([0,T]^2)^{|\mathscr{F}_2|} \times \mathscr{C}([0,T]^3)^{|\mathscr{F}_3|} \to \mathscr{C}[0,T],$$

$$\Phi^{(2)} : \mathscr{C}[0,T] \times \mathscr{C}([0,T]^2) \times \mathscr{C}[0,T]^{|\mathscr{F}_1|} \times \mathscr{C}([0,T]^2)^{|\mathscr{F}_2|} \times \mathscr{C}([0,T]^3)^{|\mathscr{F}_3|} \to \mathscr{C}([0,T]^2),$$

such that

$$R^N = \Phi^{(1)}(R^N, K^N, \mathscr{C}_N), \quad K^N = \Phi^{(2)}(R^N, K^N, \mathscr{C}_N). \tag{3.8}$$

Proof. To this end, we combine (3.5) with the definition of R^N to get a fixed point equation

$$\begin{aligned}
R^N(t) =& \frac{1}{N} \sum_{i=1}^N u^i Y_0^i \exp\left[-\int_0^t (f'(K^N(s)) - \sigma^i) ds \right] \\
&+ \frac{1}{N} \sum_{i=1}^N (u^i)^2 \int_0^t \exp\left[-\int_s^t (f'(K^N(s_1)) - \sigma^i) ds_1 \right] R^N(s) ds \\
&+ \frac{\beta^{-1/2}}{N} \sum_{i=1}^N u^i B_t^i - \frac{1}{N} \sum_{i=1}^N u^i \int_0^t B_s^i (f'(K^N(s)) - \sigma^i) \exp\left[-\int_s^t (f'(K^N(s_1)) - \sigma^i) ds_1 \right] ds.
\end{aligned} \tag{3.9}$$

This implicitly specifies the function $\Phi^{(1)}$. The corresponding equation for K^N is more involved. To track this representation systematically, we recall the representation of the solution Y_t^i from (3.5), and denote $Y_t^i := \sum_{j=1}^5 T_j^i(t)$, where the $T_j^i(t)$ represent the respective terms in the RHS of (3.5) (using the definition of F_s, we split the last term in (3.5) into two separate terms for convenience). Now, recall that $K^N(t,s) = \frac{1}{N} \sum_{i=1}^N Y_t^i Y_s^i$, and therefore

$$K^N(t,s) = \sum_{j=1}^5 \frac{1}{N} \sum_{i=1}^N T_j^i(t) T_j^i(s) + \sum_{1 \leq j_1 < j_2 \leq 5} \left[\frac{1}{N} \sum_{i=1}^N T_{j_1}^i(t) T_{j_2}^i(s) + \frac{1}{N} \sum_{i=1}^N T_{j_1}^i(s) T_{j_2}^i(t) \right]. \tag{3.10}$$

We argue that the RHS above is a continuous function of R^N, K^N and ν. To this end, we record these terms explicitly. Define

$$H_\tau^\theta(K) = \exp\left[-\int_\tau^\theta f'(K(\xi)) d\xi \right], \quad DH_\tau^\theta = \frac{dH_\tau^\theta(K)}{d\tau} = f'(K(\tau)) \exp\left[-\int_\tau^\theta f'(K(\xi)) d\xi \right]. \tag{3.11}$$

First, we present the 'diagonal' terms.

$$\frac{1}{N} \sum_{i=1}^{N} T_1^i(t) T_1^i(s) = \frac{1}{N} \sum_{i=1}^{N} (Y_0^i)^2 \exp(\sigma^i(t+s)) H_0^t(K^N) H_0^s(K^N),$$

$$\frac{1}{N} \sum_{i=1}^{N} T_2^i(t) T_2^i(s) = \int_0^t \int_0^s \frac{1}{N} \sum_{i=1}^{N} (u^i)^2 \exp(\sigma^i(t+s-s_1-s_2)) H_{s_1}^t(K^N) H_{s_2}^s(K^N) R^N(s_1) R^N(s_2) \mathrm{d}s_1 \, \mathrm{d}s_2,$$

$$\frac{1}{N} \sum_{i=1}^{N} T_3^i(t) T_3^i(s) = \frac{\beta^{-1}}{N} \sum_{i=1}^{N} B_t^i B_s^i,$$

$$\frac{1}{N} \sum_{i=1}^{N} T_4^i(t) T_4^i(s) = \frac{\beta^{-1}}{N} \sum_{i=1}^{N} \int_0^t \int_0^s B_{s_1}^i B_{s_2}^i DH_{s_1}^t(K^N) DH_{s_2}^t(K^N) \exp(\sigma^i(t+s-s_1-s_2)) \mathrm{d}s_1 \, \mathrm{d}s_2,$$

$$\frac{1}{N} \sum_{i=1}^{N} T_5^i(t) T_5^i(s) = \frac{\beta^{-1}}{N} \sum_{i=1}^{N} (\sigma^i)^2 \int_0^t \int_0^s B_{s_1}^i B_{s_2}^i \exp(\sigma^i(t+s-s_1-s_2)) \mathrm{d}s_1 \, \mathrm{d}s_2.$$

Next, we present the 'off-diagonal' terms:

$$\frac{1}{N} \sum_{i=1}^{N} [T_1^i(t) T_2^i(s) + T_1^i(s) T_2^i(t)] = H_0^t(K^N) \int_0^s \frac{1}{N} \sum_{i=1}^{N} u^i Y_0^i \exp(\sigma^i(t+s-s_1)) H_{s_1}^s(K^N) R^N(s_1) \mathrm{d}s_1$$

$$+ H_0^s(K^N) \int_0^t \frac{1}{N} \sum_{i=1}^{N} u^i Y_0^i \exp(\sigma^i(t+s-s_1)) H_{s_1}^t(K^N) R^N(s_1) \mathrm{d}s_1,$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_1^i(t) T_3^i(s) + T_1^i(s) T_3^i(t)] = H_0^t(K^N) \frac{\beta^{-1/2}}{N} \sum_{i=1}^{N} Y_0^i B_s^i \exp(\sigma^i t) + H_0^s(K^N) \frac{\beta^{-1/2}}{N} \sum_{i=1}^{N} Y_0^i B_t^i \exp(\sigma^i s),$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_1^i(t) T_4^i(s) + T_1^i(s) T_4^i(t)] = -\beta^{-1/2} H_0^t(K^N) \int_0^s \frac{1}{N} \sum_{i=1}^{N} Y_0^i B_{s_1}^i \exp(\sigma^i(t+s-s_1)) DH_{s_1}^s(K^N) \mathrm{d}s_1$$

$$- \beta^{-1/2} H_0^s(K^N) \int_0^t \frac{1}{N} \sum_{i=1}^{N} Y_0^i B_{s_1}^i \exp(\sigma^i(t+s-s_1)) DH_{s_1}^t(K^N) \mathrm{d}s_1$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_1^i(t) T_5^i(s) + T_1^i(s) T_5^i(t)] = \beta^{-1/2} H_0^t(K^N) \int_0^s \frac{1}{N} \sum_{i=1}^{N} \sigma^i Y_0^i \exp(\sigma^i(t+s-s_1)) \mathrm{d}s_1$$

$$+ \beta^{-1/2} H_0^s(K^N) \int_0^t \frac{1}{N} \sum_{i=1}^{N} \sigma^i Y_0^i \exp(\sigma^i(t+s-s_1)) \mathrm{d}s_1,$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_2^i(t) T_3^i(s) + T_2^i(s) T_3^i(t)] = \beta^{-1/2} \int_0^t \frac{1}{N} \sum_{i=1}^{N} u^i B_s^i \exp(\sigma^i(t - s_1)) H_{s_1}^t(K^N) R^N(s_1) \mathrm{d}s_1$$

$$+ \beta^{-1/2} \int_0^s \frac{1}{N} \sum_{i=1}^{N} u^i B_t^i \exp(\sigma^i(s - s_1)) H_{s_1}^t(K^N) R^N(s_1) \mathrm{d}s_1$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_2^i(t) T_4^i(s) + T_2^i(s) T_4^i(t)] =$$

$$- \beta^{-1/2} \int_0^t \int_0^s \frac{1}{N} \sum_{i=1}^{N} u^i B_{s_2}^i \exp(\sigma^i(t + s - s_1 - s_2)) H_{s_1}^t(K^N) R^N(s_1) D H_{s_2}^s(K^N) \mathrm{d}s_1 \, \mathrm{d}s_2$$

$$- \beta^{-1/2} \int_0^t \int_0^s \frac{1}{N} \sum_{i=1}^{N} u^i B_{s_1}^i \exp(\sigma^i(t + s - s_1 - s_2)) H_{s_2}^s(K^N) R^N(s_2) D H_{s_1}^t(K^N) \mathrm{d}s_1 \, \mathrm{d}s_2 ,$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_2^i(t) T_5^i(s) + T_2^i(s) T_5^i(t)] =$$

$$\beta^{-1/2} \int_0^t \int_0^s \frac{1}{N} \sum_{i=1}^{N} u^i \sigma^i B_{s_2}^i \exp(\sigma^i(t + s - s_1 - s_2)) H_{s_1}^t(K^N) R^N(s_1) \mathrm{d}s_1 \, \mathrm{d}s_2$$

$$+ \beta^{-1/2} \int_0^t \int_0^s \frac{1}{N} \sum_{i=1}^{N} u^i \sigma^i B_{s_1}^i \exp(\sigma^i(t + s - s_1 - s_2)) H_{s_2}^s(K^N) R^N(s_2) \mathrm{d}s_1 \, \mathrm{d}s_2 ,$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_3^i(t) T_4^i(s) + T_3^i(s) T_4^i(t)] = - \frac{\beta^{-1}}{N} \sum_{i=1}^{N} \int_0^s B_t^i B_{s_1}^i \exp(\sigma^i(s - s_1)) D H_{s_1}^s(K^N) \mathrm{d}s_1$$

$$- \frac{\beta^{-1}}{N} \sum_{i=1}^{N} \int_0^t B_s^i B_{s_1}^i \exp(\sigma^i(t - s_1)) D H_{s_1}^t(K^N) \mathrm{d}s_1$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_3^i(t) T_5^i(s) + T_3^i(s) T_5^i(t)] =$$

$$\beta^{-1} \int_0^s \frac{1}{N} \sum_{i=1}^{N} \sigma^i B_t^i B_{s_1}^i \exp(\sigma^i(s - s_1)) \mathrm{d}s_1 + \beta^{-1} \int_0^t \frac{1}{N} \sum_{i=1}^{N} \sigma^i B_s^i B_{s_1}^i \exp(\sigma^i(t - s_1)) \mathrm{d}s_1 ,$$

$$\frac{1}{N} \sum_{i=1}^{N} [T_4^i(t)T_5^i(s) + T_4^i(s)T_5^i(t)]$$

$$= -\frac{\beta^{-1}}{N} \sum_{i=1}^{N} \sigma^i \int_0^t \int_0^s B_{s_1}^i B_{s_2}^i \exp(\sigma^i(t+s-s_1-s_2)) DH_{s_1}^t(K^N) ds_1 \, ds_2$$

$$-\frac{\beta^{-1}}{N} \sum_{i=1}^{N} \sigma^i \int_0^t \int_0^s B_{s_1}^i B_{s_2}^i \exp(\sigma^i(t+s-s_1-s_2)) DH_{s_2}^s(K^N) ds_1 \, ds_2 \, .$$

This specifies the function $\Phi^{(2)}$ implicitly. $\qquad \square$

Given Lemma 7, we next establish that R^N, K^N are, in fact, functions of the low-dimensional statistics $\mathscr{C}_N$.

LEMMA 8. There exist functions

$$\Psi^{(1)} : \mathscr{C}[0,T]^{|\mathscr{F}_1|} \times \mathscr{C}([0,T]^2)^{|\mathscr{F}_2|} \times \mathscr{C}([0,T]^3)^{|\mathscr{F}_3|} \to \mathscr{C}[0,T],$$

$$\Psi^{(2)} : \mathscr{C}[0,T]^{|\mathscr{F}_1|} \times \mathscr{C}([0,T]^2)^{|\mathscr{F}_2|} \times \mathscr{C}([0,T]^3)^{|\mathscr{F}_3|} \to \mathscr{C}([0,T]^2)$$

such that

$$R^N = \Psi^{(1)}(\mathscr{C}_N), \; K^N = \Psi^{(2)}(\mathscr{C}_N).$$

To this end, our main strategy is to apply a Picard iteration scheme on the fixed point equations (3.8). We start with some initial guess R_0^N, K_0^N, and carry out the iterative updates,

$$R_{m+1}^N = \Phi^{(1)}(R_m^N, K_m^N, \mathscr{C}_N), \quad K_{m+1}^N = \Phi^{(2)}(R_m^N, K_m^N, \mathscr{C}_N). \tag{3.12}$$

We will show that this iteration system is contractive, and thus identify R^N, K^N as the unique fixed points of this system. In this endeavor, we will utilize the precise form of the functions $\Phi^{(1)}, \Phi^{(2)}$ as described in the proof of Lemma 7. First, we need some preliminary estimates.

LEMMA 9. Recall that f', defined in (1.2), is non-negative and Lipschitz. Then we have that

(i) For any $m, N \geq 1, 0 \leq H_\tau^\theta(K_m^N) \leq 1$.

(ii) For any $0 \leq t \leq T, N, m \geq 1, \int_0^t |DH_u^t(K_m^N)| du \leq 1$.

(iii) For any $\theta \leq T$,

$$\sup_{\tau \leq \theta} |H_\tau^\theta(K_{m+1}^N) - H_\tau^\theta(K_m^N)| \leq \|f'\|_L \int_0^\theta |K_{m+1}^N(s,s) - K_m^N(s,s)| ds \, . \tag{3.13}$$

(iv)   For any $N, m \geq 1$, $0 \leq \tau \leq \theta \leq T$,

$$|DH_\tau^\theta(K_{m+1}^N) - DH_\tau^\theta(K_m^N)| \leq \|f'\|_L \Big[|K_{m+1}^N(\tau, \tau) - K_m^N(\tau, \tau)| + (DH_\tau^\theta(K_{m+1}^N)$$

$$+ DH_\tau^\theta(K_m^N)) \int_0^\theta |K_{m+1}^N(s, s) - K_m^N(s, s)| ds \Big].$$

(v)   With probability 1, there exist C_0, C_1 (possibly random), depending on T such that

$$\sup_{m \geq 1} \|R_m^N\|_\infty \leq C_0 \exp(C_1 T). \tag{3.14}$$

Proof of Lemma 9. The parts (i)–(iv) are directly adapted from [9], and are just collected here for the convenience of the reader. We prove (v). Note that

$$R_{m+1}^N(t) = H_0^t(K_m^N) \frac{1}{N} \sum_{i=1}^N u^i Y_0^i \exp(\sigma^i t) + \frac{1}{N} \sum_{i=1}^N (u^i)^2 \int_0^t \exp(\sigma^i(t-s)) R_m^N(s) H_s^t(K_m^N) ds$$

$$+ \frac{\beta^{-1/2}}{N} \sum_{i=1}^N u^i B_t^i - \frac{1}{N} \sum_{i=1}^N u^i \int_0^t B_s^i \exp(\sigma^i(t-s)) DH_s^t(K_m^N) ds$$

$$+ \frac{1}{N} \sum_{i=1}^N u^i \sigma^i \int_0^t B_s^i \exp(\sigma^i(t-s)) H_s^t(K_m^N) ds.$$

This implies

$$|R_{m+1}^N(t)| \leq |H_0^t(K_m^N)| \cdot |\frac{1}{N} \sum_{i=1}^N u^i Y_0^i \exp(\sigma^i t)| + \int_0^t |\frac{1}{N} \sum_{i=1}^N (u^i)^2 \exp(\sigma^i(t-s))| \cdot |H_s^t(K_{(m)}^N)| \cdot |R_{(m)}^N(s)| ds$$

$$+ \frac{\beta^{-1/2}}{N} |\sum_{i=1}^N u^i B_t^i| + \int_0^t |\frac{1}{N} \sum_{i=1}^N u^i B_s^i \exp(\sigma^i(t-s))| \cdot |DH_s^t(K_{(m)}^N)| ds$$

$$+ \int_0^t |\frac{1}{N} \sum_{i=1}^N u^i \sigma^i B_s^i \exp(\sigma^i(t-s))| \cdot |H_s^t(K_{(m)}^N)| ds.$$

Thus there exists $C_0, C_1 > 0$ (possibly random), independent of m, such that

$$\sup_{0 \leq t \leq T} |R_{m+1}^N(t)| \leq C_0 + C_1 \int_0^T |R_m^N(s)| ds.$$

Iterating this bound in m, we obtain

$$\sup_{0 \le t \le T} |R_{m+1}^N(t)| \le C_0 \exp(C_1 T).$$

This completes the proof. □

Armed with Lemma 9, we turn to a proof of Lemma 8.

Proof of Lemma 8. Define two sequences of Picard iterations with different initializations, (R_0^N, K_0^N) and $(\tilde{R}_0^N, \tilde{K}_0^N)$:

$$R_{m+1}^N = \Phi^{(1)}(R_m^N, K_m^N, \mathscr{C}_N), \quad K_{m+1}^N = \Phi^{(2)}(R_m^N, K_m^N, \mathscr{C}_N). \tag{3.15}$$

$$\tilde{R}_{m+1}^N = \Phi^{(1)}(\tilde{R}_m^N, \tilde{K}_m^N, \mathscr{C}_N), \quad \tilde{K}_{m+1}^N = \Phi^{(2)}(\tilde{R}_m^N, \tilde{K}_m^N, \mathscr{C}_N). \tag{3.16}$$

Observe that for any $m \ge 0$, (3.9) implies

$$R_{m+1}^N(t) - \tilde{R}_{m+1}^N(t) = \frac{1}{N} \sum_{i=1}^N u^i Y_0^i \exp(\sigma^i t)(H_0^t(K_m^N) - H_0^t(\tilde{K}_m^N))$$

$$+ \int_0^t \frac{1}{N} \sum_{i=1}^N (u^i)^2 \exp(\sigma^i(t-s))(R_m^N(s) H_s^t(K_m^N) - \tilde{R}_m^N(s) H_s^t(\tilde{K}_m^N)) \mathrm{d}s$$

$$- \int_0^t \frac{1}{N} \sum_{i=1}^N u^i B_s^i \exp(\sigma^i(t-s))(DH_s^t(K_m^N) - DH_s^t(\tilde{K}_m^N)) \mathrm{d}s$$

$$+ \int_0^t \frac{1}{N} \sum_{i=1}^N u^i \sigma^i B_s^i \exp(\sigma^i(t-s))(H_s^t(K_m^N) - H_s^t(\tilde{K}_m^N)) \mathrm{d}s.$$

This implies, with Lemma 9, there exists a constant $C > 0$ that potentially depends on T such that, for all $t \le T$:

$$\sup_{0 \le s \le t} |R_{m+1}^N(s) - \tilde{R}_{m+1}^N(s)|$$

$$\le C \Big[\int_0^t |K_m^N(s,s) - \tilde{K}_m^N(s,s)| \mathrm{d}s + \int_0^t |R_m^N(s) H_s^t(K_m^N) - \tilde{R}_m^N(s) H_s^t(\tilde{K}_m^N)| \mathrm{d}s \Big]. \tag{3.17}$$

To control the second term, we observe

$$|R_m^N(s) H_s^t(K_m^N) - \tilde{R}_m^N(s) H_s^t(\tilde{K}_m^N)|$$

$$\le |R_m^N(s)| \cdot |H_s^t(K_m^N) - H_s^t(\tilde{K}_m^N)| + |H_s^t(\tilde{K}_m^N)| \cdot |R_m^N(s) - \tilde{R}_m^N(s)|,$$

and, therefore,

$$\sup_{0 \le s \le t} |R_{m+1}^N(s) - \tilde{R}_{m+1}^N(s)| \le C' \Big[\int_0^t |K_m^N(s,s) - \tilde{K}_m^N(s,s)| \mathrm{d}s + \int_0^t |R_m^N(s) - \tilde{R}_m^N(s)| \mathrm{d}s \Big]. \tag{3.18}$$

Similarly, analyzing the iterative update equation for $K_m^N, \tilde{K}_m^N$,

$$\sup_{0 \le s,s' \le t} |K_{m+1}^N(s,s') - \tilde{K}_{m+1}^N(s,s')| \le C'' \Big[\int_0^t |K_m^N(s,s) - \tilde{K}_m^N(s,s)| \mathrm{d}s + \int_0^t |R_m^N(s) - \tilde{R}_m^N(s)| \mathrm{d}s \Big]. \tag{3.19}$$

Iterating these bounds, we have shown for any $t \le T$,

$$\sup_{0 \le s \le t} |R_{m+1}^N(s) - \tilde{R}_{m+1}^N(s)| + \sup_{0 \le s,s' \le t} |K_{m+1}^N(s,s') - \tilde{K}_{m+1}^N(s,s')| \tag{3.20}$$

$$\le \frac{t^m}{m!} \cdot (C' + C'')^m \Big[\sup_{0 \le s \le T} |R_0^N(s) - \tilde{R}_0^N(s)| + \sup_{0 \le s,s' \le T} |K_0^N(s,s') - \tilde{K}_0^N(s,s')| \Big]. \tag{3.21}$$

This implies the following bound:

$$\max\{\|R_{m+1}^N - \tilde{R}_{m+1}^N\|_\infty, \|K_{m+1}^N - \tilde{K}_{m+1}^N\|_\infty\} \tag{3.22}$$

$$\le \frac{T^m}{m!} \cdot (C' + C'')^m \Big[\sup_{0 \le s \le T} |R_0^N(s) - \tilde{R}_0^N(s)| + \sup_{0 \le s,s' \le T} |K_0^N(s,s') - \tilde{K}_0^N(s,s')| \Big], \tag{3.23}$$

For $m \in \mathbb{N}$ large enough, $\frac{T^m}{m!} \cdot (C' + C'')^m < 1$, the above shows that the iterated operator Φ^m with $\Phi : (R,K) \mapsto (\Phi^{(1)}(R,K,\mathscr{C}_N), \Phi^{(2)}(R,K,\mathscr{C}_N))$ is a contraction under the $\|\cdot\|_\infty$ metric. Thus, by the Banach fixed-point theorem, we know Φ^m has a unique fixed point denoted as (R,K). Note that

$$\Phi^m(\Phi(R,K)) = \Phi(\Phi^m(R,K)) = \Phi(R,K) \tag{3.24}$$

and thus $\Phi(R,K)$ is a fixed point of Φ^m as well, and by uniqueness, we must have $\Phi(R,K) = (R,K)$. This shows the existence of a fixed point for Φ. Using the fact that any fixed point of Φ must be a fixed point of Φ^m, we conclude Φ has the unique fixed point (R,K) as Φ^m. Since (R^N, K^N) satisfies (3.8), this implies $(R,K) = (R^N, K^N)$. Thus, the Picard iteration scheme outlined above uniquely specifies (R^N, K^N) as a function of $\mathscr{C}_N$. □

Finally, we will establish that $\Psi^{(1)}, \Psi^{(2)}$ derived in Lemma 8 are continuous functions. To this end, recall that

$$\Psi^{(1)} : \mathscr{C}[0,T]^{|\mathscr{F}_1|} \times \mathscr{C}([0,T]^2)^{|\mathscr{F}_2|} \times \mathscr{C}([0,T]^3)^{|\mathscr{F}_3|} \to \mathscr{C}[0,T],$$

$$\Psi^{(2)} : \mathscr{C}[0,T]^{|\mathscr{F}_1|} \times \mathscr{C}([0,T]^2)^{|\mathscr{F}_2|} \times \mathscr{C}([0,T]^3)^{|\mathscr{F}_3|} \to \mathscr{C}([0,T]^2)$$

We equip $\mathscr{C}([0,T]^j)$, $j = 1, 2, 3$, with the sup-norm topology. Further, we equip $\mathscr{C}[0,T]^{|\mathscr{F}_1|} \times \mathscr{C}([0,T]^2)^{|\mathscr{F}_2|} \times \mathscr{C}([0,T]^3)^{|\mathscr{F}_3|}$ with the product topology. With this notion of convergence, we can establish the following continuity properties of $\Psi^{(1)}$ and $\Psi^{(2)}$.

LEMMA 10. The maps $\mathscr{C} \mapsto \Psi^{(1)}(\mathscr{C})$ and $\mathscr{C} \mapsto \Psi^{(2)}(\mathscr{C})$ are continuous.

Proof. We measure the discrepancy between $\mathscr{C}$ and $\tilde{\mathscr{C}}$ using the uniform topology on this space, and denote $\mathrm{d}(\mathscr{C}, \tilde{\mathscr{C}}) = \|\mathscr{C} - \tilde{\mathscr{C}}\|_\infty$. Define $R, K, \tilde{R}, \tilde{K}$ via the fixed point equations:

$$R = \Phi^{(1)}(R, K, \mathscr{C}), \quad K = \Phi^{(2)}(R, K, \mathscr{C}),$$
$$\tilde{R} = \Phi^{(1)}(\tilde{R}, \tilde{K}, \tilde{\mathscr{C}}), \quad \tilde{K} = \Phi^{(2)}(\tilde{R}, \tilde{K}, \tilde{\mathscr{C}}).$$

Observe that

$$\|R - \tilde{R}\|_\infty = \|\Phi^{(1)}(R, K, \mathscr{C}) - \Phi^{(1)}(\tilde{R}, \tilde{K}, \tilde{\mathscr{C}})\|_\infty$$
$$\leq \|\Phi^{(1)}(R, K, \mathscr{C}) - \Phi^{(1)}(\tilde{R}, \tilde{K}, \mathscr{C})\|_\infty + \|\Phi^{(1)}(\tilde{R}, \tilde{K}, \mathscr{C}) - \Phi^{(1)}(\tilde{R}, \tilde{K}, \tilde{\mathscr{C}})\|_\infty.$$

Controlling the second term, we note from (3.9) that

$$\|\Phi^{(1)}(\tilde{R}, \tilde{K}, \mathscr{C}) - \Phi^{(1)}(\tilde{R}, \tilde{K}, \tilde{\mathscr{C}})\|_\infty \leq \|\mathscr{C} - \tilde{\mathscr{C}}\|_\infty.$$

On the other hand, using the same argument as in the proof of Lemma 8, we have that

$$\|\Phi^{(1)}(R, K, \mathscr{C}) - \Phi^{(1)}(\tilde{R}, \tilde{K}, \mathscr{C})\|_\infty \leq C\left[\int_0^T |K(s,s) - \tilde{K}(s,s)|\mathrm{d}s + \int_0^T |R(s) - \tilde{R}(s)|\mathrm{d}s\right].$$

A similar argument for $\Phi^{(2)}$ implies that for any $t \in [0, T]$,

$$\sup_{0 \leq s \leq t} |R(s) - \tilde{R}(s)| \leq \|\mathscr{C} - \tilde{\mathscr{C}}\|_\infty + C\left[\int_0^t |K(s,s) - \tilde{K}(s,s)|\mathrm{d}s + \int_0^t |R(s) - \tilde{R}(s)|\mathrm{d}s\right], \quad (3.25)$$

$$\sup_{0 \leq s,s' \leq t} |K(s,s') - \tilde{K}(s,s')| \leq \|\mathscr{C} - \tilde{\mathscr{C}}\|_\infty + C\left[\int_0^t |K(s,s) - \tilde{K}(s,s)|\mathrm{d}s + \int_0^t |R(s) - \tilde{R}(s)|\mathrm{d}s\right]. \quad (3.26)$$

By Gronwall's lemma, we have

$$\|R - \tilde{R}\|_\infty + \|K - \tilde{K}\|_\infty \leq 2\|\mathscr{C} - \tilde{\mathscr{C}}\|_\infty \exp(C'T). \quad (3.27)$$

This completes the proof. □

Our next lemma establishes that under the two initial conditions introduced in Section 1, the low-dimensional statistics $\mathscr{C}_N$ converge almost surely to deterministic limits. We defer the proof of this lemma to Appendix A.

LEMMA 11. Under the i.i.d. and i.i.d. under rotated basis initial conditions, each element in $\mathscr{C}_N$ converges to deterministic limits almost surely.

Finally, we complete the proof of Theorem 1, assuming Lemmas 8, 10 and 11.

Proof. (Proof of Theorem 1) Fix $T > 0$. Lemmas 8, 10 and 11 together imply that R^N, K^N converge to deterministic functions R and K, respectively. It remains to characterize these limits. Note that from (3.3),

$$Y_t^i = \exp\left\{\int_0^t \left[\sigma^i - f'(K^N(s))\right]\mathrm{d}s\right\} \left\{Y_0^i + u^i \int_0^t \exp\left\{-\int_0^s \left[\sigma^i - f'(K^N(r))\right]\mathrm{d}r\right\} R^N(s)\mathrm{d}s\right\} \quad (3.28)$$

$$+ \beta^{-1/2} \int_0^t \exp\left\{\int_s^t \left[\sigma^i - f'(K^N(r))\right]\mathrm{d}r\right\} \mathrm{d}B_s^i. \quad (3.29)$$

We recall that $R^N(t) = \frac{1}{N}\sum_{i=1}^N u^i Y_t^i$ and $K^N(t,s) = \frac{1}{N}\sum_{i=1}^N Y_t^i Y_s^i$, and that $R^N \to R$, $K^N \to K$ uniformly. Thus calculating R^N, K^N and setting $N \to \infty$, we obtain the desired fixed point equations in the limit. Note that the final limit operation requires that $\nu = \frac{1}{N}\sum_{i=1}^N \delta_{Y_0^i, u^i, \sigma^i, B_\bullet^i}$ converges to the claimed limits under the i.i.d. and rotated i.i.d. initial conditions. This completes the proof. □

Data Availability Statement

No new data were generated or analysed in support of this review.

REFERENCES

1. AGORITSAS, E., BIROLI, G., URBANI, P. & ZAMPONI, F. (2018) Out-of-equilibrium dynamical mean-field equations for the perceptron model. *J. Phys. A: Math. Theor.*, **51**, 085002.
2. ALI, A., DOBRIBAN, E. & TIBSHIRANI, R. (2020) The implicit regularization of stochastic gradient flow for least squares. In: *International Conference on Machine Learning*, PMLR, pp. 233–244.
3. ALNUR, ALI, J., KOLTER, Z. & TIBSHIRANI, R. J. (2019) A continuous-time view of early stopping for least squares regression. In: *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1370–1378.
4. ARORA, S., COHEN, N., WEI, H. & LUO, Y. (2019) Implicit regularization in deep matrix factorization. *Adv. Neural Inf. Process. Syst.*, **32**.
5. BEN AROUS, G. & GUIONNET, A. (1997) Symmetric Langevin spin glass dynamics. *Ann. Probab.*, **25**, 1367–1422.
6. AROUS, G. B. & GUIONNET, A. (1998). Langevin dynamics for Sherrington-Kirkpatrick spin glasses. In: ANTON BOVIER & PIERRE PICCO (eds) *Mathematical Aspects of Spin Glasses and Neural Networks*. Boston, MA: Springer, pp. 323–353.
7. BAIK, J., AROUS, G. B. & PÉCHÉ, S. (2005) Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.*, **33**, 1643–1697.

8. BELKIN, M., RAKHLIN, A. & TSYBAKOV, A. B. (2019). Does data interpolation contradict statistical optimality? In: *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1611–1619. https://proceedings.mlr.press/v89/belkin19a.html.

9. AROUS, G. B., DEMBO, A. & GUIONNET, A. (2001) Aging of spherical spin glasses. *Probab. Theory Related Fields*, **120**, 1–67.

10. AROUS, G. B., DEMBO, A. & GUIONNET, A. (2006) Cugliandolo-Kurchan equations for dynamics of spin-glasses. *Probab. Theory Related Fields*, **136**, 619–660.

11. AROUS, G. B., GHEISSARI, R. & JAGANNATH, A. (2020) Algorithmic thresholds for tensor PCA. *Ann. Probab.*, **48**, 2052–2087.

12. AROUS, G. B., GHEISSARI, R. & JAGANNATH, A. (2020) Bounding flows for spherical spin glass dynamics. *Comm. Math. Phys.*, **373**, 1011–1048.

13. AROUS, G. B., GHEISSARI, R. & JAGANNATH, A. (2021) Online stochastic gradient descent on non-convex losses from high-dimensional inference. *J. Mach. Learn. Res.*, **22**, 1–51.

14. BENAYCH-GEORGES, F. & NADAKUDITI, R. R. (2011) The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.*, **227**, 494–521.

15. BODIN, A. & MACRIS, N. (2021) Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. *Adv. Neural Inf. Process. Syst.*, **34**, 21605–21617.

16. BODIN, A. & MACRIS, N. (2021) Rank-one matrix estimation: analytic time evolution of gradient descent dynamics. In: *Conference on Learning Theory*, PMLR, pp. 635–678. https://proceedings.mlr.press/v134/.

17. CAMMAROTA, C., BIROLI, G., SARAO, S., ZDEBOROVA, L. & KRZAKALA, F. (2020) Who is afraid of big bad minima? analysis of gradient-flow in a spiked matrix-tensor model. In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, vol. **33**, Vancouver: Curran Associates, Inc.

18. CELENTANO, M., CHENG, C. & MONTANARI, A. (2021) The high-dimensional asymptotics of first order methods with random data. arXiv preprint arXiv:2112.07572.

19. CHATTERJI, N. S. & LONG, P. M. (2021) Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *J. Mach. Learn. Res.*, **22**, 1–30.

20. CHATTERJI, N. S., LONG, P. M. & BARTLETT, P. L. (2021) When does gradient descent with logistic loss find interpolating two-layer networks? *J. Mach. Learn. Res.*, **22**, 1–48.

21. CHIZAT, L. & BACH, F. (2020) Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In: *Conference on Learning Theory*, PMLR, pp. 1305–1338.

22. CRISANTI, A. & SOMMERS, H.-J. (1992) The sphericalp-spin interaction spin glass model: the statics. *Z. Phys. B Condens. Matter*, **87**, 341–354.

23. CUGLIANDOLO, L. F. & KURCHAN, J. (1993) Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Phys. Rev. Lett.*, **71**, 173.

24. DEMBO, A. & GHEISSARI, R. (2021) Diffusions interacting through a random matrix: universality via stochastic Taylor expansion. *Probab. Theory Related Fields*, **180**, 1057–1097.

25. DEMBO, A., GUIONNET, A. & MAZZA, C. (2007) Limiting dynamics for spherical models of spin glasses at high temperature. *J. Statist. Phys.*, **126**, 781–815.

26. DEMBO, A., LUBETZKY, E. & ZEITOUNI, O. (2021) Universality for Langevin-like spin glass dynamics. *Ann. Appl. Probab.*, **31**, 2864–2880.

27. DEMBO, A. & SUBAG, E. (2020) Dynamics for spherical spin glasses: disorder dependent initial conditions. *J. Statist. Phys.*, **181**, 465–514.

28. DENG, Z., KAMMOUN, A. & THRAMPOULIDIS, C. (2022) A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, **11**, 435–495.

29. GRUNWALD, M. (1996) Sanov results for Glauber spin-glass dynamics. *Probab. Theory Related Fields*, **106**, 187–232.

30. GRUNWALD, M. (1998). Sherrington-Kirkpatrick spin-glass dynamics. In: BOVIER ANTON & PICCO PIERRE (eds) *Mathematical Aspects of Spin Glasses and Neural Networks*. Boston, MA: Springer, pp. 355–382.

31. GUNASEKAR, S., LEE, J., SOUDRY, D. & SREBRO, N. (2018) Characterizing implicit bias in terms of optimization geometry. In: *International Conference on Machine Learning*, PMLR, pp. 1832–1841.

32. GUNASEKAR, S., LEE, J. D., SOUDRY, D. & SREBRO, N. (2018) Implicit bias of gradient descent on linear convolutional networks. *Adv. Neural Inf. Process. Syst.*, **31**.

33. JI, Z. & TELGARSKY, M. (2019) The implicit bias of gradient descent on nonseparable data. In: *Conference on Learning Theory*, PMLR, pp. 1772–1798.

34. JIN, H. & MONTÚFAR, G. (2023) Implicit bias of gradient descent for mean squared error regression with wide neural networks. *Journal of Machine Learning Research*, **24**, 1–97.

35. LIANG, T., RAKHLIN, A. & ZHAI, X. (2020) On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. *Conference on Learning Theory*, PMLR, pp. 2683–2711.

36. LIANG, T. & SUR, P. (2022) A precise high-dimensional asymptotic theory for boosting and minimum-ℓ1-norm interpolated classifiers. *Ann. Statist.*, **50**, 1669–1695.

37. MA, C., WANG, K., CHI, Y. & CHEN, Y. (2018) Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In: *International Conference on Machine Learning*, PMLR, pp. 3345–3354. https://proceedings.mlr.press/.

38. MA, S., BASSILY, R. & BELKIN, M. (2018) The power of interpolation: understanding the effectiveness of SGD in modern over-parametrized learning. In: *International Conference on Machine Learning*, PMLR, pp. 3325–3334. https://proceedings.mlr.press/.

39. MANNELLI, S. S., BIROLI, G., CAMMAROTA, C., KRZAKALA, F., URBANI, P. & ZDEBOROVÁ, L. (2020) Marvels and pitfalls of the Langevin algorithm in noisy high-dimensional inference. *Phys. Rev. X*, **10**, 011057.

40. MANNELLI, S. S., KRZAKALA, F., URBANI, P. & ZDEBOROVA, L. (2019) Passed & spurious: descent algorithms and local minima in spiked matrix-tensor models. In: *International Conference on Machine Learning*. PMLR, pp. 4333–4342.

41. MANNELLI, S. S. & ZDEBOROVÁ, L. (2020) Thresholds of descending algorithms in inference problems. *J. Stat. Mech.: Theory Exp.*, **2020**, 034004.

42. MIGNACCO, F., KRZAKALA, F., URBANI, P. & ZDEBOROVÁ, L. (2020) Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Adv. Neural Inf. Process. Syst.*, **33**, 9540–9550.

43. MONTANARI, A., RUAN, F., SOHN, Y. & YAN, J. (2019) The generalization error of max-margin linear classifiers: high-dimensional asymptotics in the overparametrized regime. arXiv preprint arXiv:1911.01544.

44. NACSON, M. S., LEE, J., GUNASEKAR, S., SAVARESE, P. H. P., SREBRO, N. & SOUDRY, D. (2019) Convergence of gradient descent on separable data. In: *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 3420–3428.

45. MANNELLI, S. S. & URBANI, P. (2021) Analytical study of momentum-based acceleration methods in paradigmatic high-dimensional non-convex problems. *Adv. Neural Inf. Process. Syst.*, **34**, 187–199.

46. SOMPOLINSKY, H. & ZIPPELIUS, A. (1981) Dynamic theory of the spin-glass phase. *Phys. Rev. Lett.*, **47**, 359.

47. SOMPOLINSKY, H. & ZIPPELIUS, A. (1982) Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Phys. Rev. B*, **25**, 6860.

48. SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S. & SREBRO, N. (2018) The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, **19**, 2822–2878.

49. ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. & VINYALS, O. (2021) Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, **64**, 107–115.

A. Proof of Lemma 11

We prove Lemma 11 in this section. Note that for $f \in \mathcal{F}_j, j = 1, 2, 3, \int f \mathrm{d}\nu \in \mathcal{C}([0, T]^j)$. Thus, in the lemma above, we mean specifically that $\int f \mathrm{d}\nu$ converges almost surely as a $\mathcal{C}([0, T]^j)$-valued random variable. Throughout, we equip $\mathcal{C}([0, T]^j)$ with a sup-norm.

Proof. Before embarking on a formal proof, we summarize our general proof strategy. The proof follows in two stages—first, we establish that for any fixed $x \in [0, T]^j, \int f \mathrm{d}\nu (x)$ converges almost surely. We next prove an additional Holder continuity property of $\int f \mathrm{d}\nu$, uniformly in n, which allows us to bootstrap the above pointwise a.s. convergence to a functional a.s. convergence statement.

We first consider the case of the initial condition (ii). Here, almost sure convergence of $\int f \mathrm{d}\nu\,(x)$ for any fixed $x \in [0, T]^j$ follows immediately from the Strong Law of Large Numbers. We next establish that for sufficiently large n, the following holds for each element of $\mathcal{C}_N$: there exists $C, \alpha > 0$, independent of n, such that almost surely, for any $x, y \in [0, 1]^j$,

$$\left| \int f \mathrm{d}\nu\,(x) - \int f \mathrm{d}\nu\,(y) \right| \leq C \|x - y\|^{\alpha}. \tag{A1}$$

Before we delve into the proof, we argue that (A1) suffices to establish convergence of $\int f \mathrm{d}\nu$ to $\mathbb{E} \int f \mathrm{d}\nu$ in sup-norm.

To see this, first observe that for every $f \in \mathcal{F}_j, j = 1, 2, 3$, $\mathbb{E}[\int f \mathrm{d}\nu] \in \mathcal{C}([0, T]^j)$, by the explicit form of these functions. Thus, $\mathbb{E}[\int f \mathrm{d}\nu](\cdot)$ is uniformly continuous on $[0, T]^j$. This implies that if we fix $\epsilon > 0$, there exists $\delta_f > 0$ such that whenever $|t_1 - t_2| \leq \delta_f$, we have that

$$\left| \mathbb{E}\left[\int f \mathrm{d}\nu \right](t_1) - \mathbb{E}\left[\int f \mathrm{d}\nu \right](t_2) \right| \leq \epsilon.$$

Define $\delta = \min\{\delta_f : f \in \mathcal{F}\}$ and fix any $0 < \delta' < \delta$. Let $\{x_1, \cdots, x_\ell\}$ denote a δ'-net of $[0, 1]^j$. For any $y \in [0, T]^j$, if we denote x_t to be the nearest point in the net, we know that for sufficiently large n,

$$\left| \left[\int f \mathrm{d}\nu \right](y) - \mathbb{E}\left[\int f \mathrm{d}\nu \right](y) \right| \leq \left| \left[\int f \mathrm{d}\nu \right](y) - \left[\int f \mathrm{d}\nu \right](x_t) \right| \tag{A2}$$

$$+ \left| \left[\int f \mathrm{d}\nu \right](x_t) - \mathbb{E}\left[\int f \mathrm{d}\nu \right](x_t) \right| + \left| \mathbb{E}\left[\int f \mathrm{d}\nu \right](x_t) - \mathbb{E}\left[\int f \mathrm{d}\nu \right](y) \right|$$

$$\leq C(\delta')^{\alpha} + 2\epsilon.$$

Since the RHS does not depend on y, and $\epsilon, \delta' > 0$ can be arbitrary as long as $\delta' < \delta$, we have the required sup-norm convergence of $\int f \mathrm{d}\nu$ as a function on $\mathcal{C}([0, T]^j)$. We shall next establish Holder continuity of $\int f \mathrm{d}\nu$ (uniformly in n), that is, property (A1) for each $f \in \mathcal{F}$.

To begin, let us consider f_1. Observe that

$$\left| \frac{1}{N} \sum_{i=1}^{N} u^i Y_0^i \exp(w_1 \sigma^i) - \frac{1}{N} \sum_{i=1}^{N} u^i Y_0^i \exp(w_2 \sigma^i) \right| \leq \exp(\|\sigma\|_\infty T) \left(\frac{1}{N} \sum_{i=1}^{N} |u^i||Y_0^i| \right) |w_1 - w_2| \quad \text{(A3)}$$

$$\leq C|w_1 - w_2|,$$

where the last inequality is true a.s. with C independent of N only for sufficiently large n on using the SLLN. Thus, f_1 is Lipschitz almost surely for sufficiently large n. For the above upper bound, recall that we always have $\|\sigma_i\|_\infty \leq 2\max\{|d_+|, |d_-|\}$ for n sufficiently large. Similar arguments work for f_2, f_4, f_5, so we skip presenting those details here.

We next turn to f_3. Recall that

$$\frac{1}{N} \sum_{i=1}^{N} u^i B_t^i = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} U^i B_t^i = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} V^i W_t^i,$$

where the last equality follows from the orthogonality of the matrix $\mathbf{G}$. Recall that $V^i \sim \mathcal{N}(0, \frac{\lambda}{N})$. This implies

$$\mathbb{P}\left[\sup_{0 \le t \le T} \Big| \sum_{i=1}^N V^i W_t^i \Big| > \sqrt{N}\varepsilon \right] \le \mathbb{P}\left[\sup_{0 \le t \le T} \Big| \sum_{i=1}^N V^i W_t^i \Big| > \sqrt{N}\varepsilon, \sum_{i=1}^N (V^i)^2 \le C' \right] + \mathbb{P}\left[\sum_{i=1}^N (V^i)^2 \ge C' \right]$$
$$\le 10 \exp(-CN) \tag{A4}$$

using standard deviation bounds for the supremum of a Brownian motion and a Chernoff bound on $\sum_{i=1}^N (V^i)^2$. This completes the proof for f_3.

Next, we move to functions in $\mathcal{F}_2$. First, note that for f_8, we have that $\int f_8 d\nu \to \mathbb{E} \int f_8 d\nu$ on $\mathcal{C}([0,T]^2)$, by the Uniform Law of Large Numbers and properties of Brownian Motion. Next, we present the proof for f_6—the proofs for f_7 and f_9 are similar. To show Holder continuity of f_6, set $\mathbf{D}_w = \mathrm{diag}(\exp(w\sigma^1)\cdots, \exp(w\sigma^N))$. Then we have

$$\frac{1}{N} \sum_{i=1}^N u^i B_t^i \exp(w\sigma^i) = \frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \mathbf{D}_w \mathbf{G} \mathbf{W}_t.$$

Thus, for any $(t_1, w_1), (t_2, w_2) \in [0, T]^2$,

$$\left| \frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \mathbf{D}_{w_1} \mathbf{G} \mathbf{W}_{t_1} - \frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \mathbf{D}_{w_2} \mathbf{G} \mathbf{W}_{t_2} \right|$$
$$\le \frac{1}{\sqrt{N}} \left| V^\top \mathbf{G}^\top \mathbf{D}_{w_1} \mathbf{G}(\mathbf{W}_{t_1} - \mathbf{W}_{t_2}) \right| + \frac{1}{\sqrt{N}} |V^\top \mathbf{G}^\top (\mathbf{D}_{w_1} - \mathbf{D}_{w_2}) \mathbf{G} \mathbf{W}_{t_2}|$$
$$\le \|\mathbf{G}V\|_2 \|\mathbf{D}_{w_1}\|_2 \frac{1}{\sqrt{N}} \|\mathbf{W}_{t_1} - \mathbf{W}_{t_2}\|_2 + \|\mathbf{G}V\|_2 \|\mathbf{D}_{w_1} - \mathbf{D}_{w_2}\|_2 \frac{1}{\sqrt{N}} \|\mathbf{W}_{t_2}\|_2.$$

There exists $C > 0$ such that almost surely, $\|\mathbf{G}V\|_2 = \|V\|_2 < C$, $\sup_{w \in [0,T]} \|\mathbf{D}_{w_1}\|_2 \le \exp(\|\sigma\|_\infty T)$. Further, $\|\mathbf{D}_{w_1} - \mathbf{D}_{w_2}\|_2 \le C|w_1 - w_2|$, and

$$\sup_{t_1, t_2 \in [0,T]} \left| \frac{1}{N} \sum_{i=1}^N (W_{t_1}^i - W_{t_2}^i)^2 - |t_1 - t_2| \right| \overset{as}{\to} 0.$$

Thus, there exists constants $C, C' > 0$ such that almost surely,

$$\left| \frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \mathbf{D}_{w_1} \mathbf{G} \mathbf{W}_{t_1} - \frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \mathbf{D}_{w_2} \mathbf{G} \mathbf{W}_{t_2} \right|$$
$$\le C(\sqrt{|t_1 - t_2|} + |w_1 - w_2|) \le C'(\sqrt{|t_1 - t_2|} + \sqrt{|w_1 - w_2|}),$$

where the last inequality follows from the fact that $w_1, w_2 \in [0, T]$, so that we always have $\sqrt{|w_1 - w_2|} \le \sqrt{2T}$. The other functions in $\mathcal{F}_2$ may be controlled using analogous arguments.

Finally, we move onto the functions in $\mathcal{F}_3$. We sketch the proof for f_{12}; the same proof works for f_{10}, f_{11}. Set $\tilde{\boldsymbol{D}}_w = \mathrm{diag}((\sigma^i)^2 \exp(w\sigma_i))$. Thus, we have, for $(w, s_1, s_2), (w', s_1', s_2') \in [0, T]^3$,

$$
\left| \frac{1}{N} \sum_{i=1}^N (\sigma^i)^2 \exp(w\sigma^i) B_{s_1}^i B_{s_2}^i - \frac{1}{N} \sum_{i=1}^N (\sigma^i)^2 \exp(w'\sigma^i) B_{s_1'}^i B_{s_2'}^i \right|
$$

$$
= \frac{1}{N} \left| \mathbf{B}_{s_1}^\top \tilde{\boldsymbol{D}}_w \mathbf{B}_{s_2} - \mathbf{B}_{s_1'}^\top \tilde{\boldsymbol{D}}_{w'} \mathbf{B}_{s_2'} \right| \leq C(\sqrt{|s_1 - s_1'|} + |w - w'| + \sqrt{|s_2 - s_2'|}) \tag{A5}
$$

$$
\leq C'(\sqrt{|s_1 - s_1'|} + \sqrt{|w - w'|} + \sqrt{|s_2 - s_2'|})
$$

almost surely for some universal constants $C, C' > 0$. This completes the proof for initial condition (ii).

Next, we turn to the initial condition (i). The main difference now lies in the fact that the pointwise almost sure convergence no longer follows directly from SLLN for all choices of $\mathcal{F}$. First, we observe that the difference in initial conditions only affects f_1, f_4, f_5 and f_9—thus, we can restrict to these specific functions. We will use the same strategy to establish functional almost sure convergence, starting from the pointwise a.s. convergence, thus we omit those details. We present the proof for f_1—the proofs for f_4, f_5, f_9 are similar. Fix $w \in [0, T]$. We have

$$
\frac{1}{N} \sum_{i=1}^N u^i Y_0^i \exp(w\sigma^i) = \frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \boldsymbol{D}_w \mathbf{G} \mathbf{X}_0, \tag{A6}
$$

where we use $\boldsymbol{D}_w = \mathrm{diag}(\exp(w\sigma^i))$. We observe that

$$
\mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N u^i Y_0^i \exp(w\sigma^i) \right] = \mathbb{E}\left[\mathbb{E}[\frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \boldsymbol{D}_w \mathbf{G} \mathbf{X}_0 | \mathbf{G}, \mathbf{X}_0] \right] = 0.
$$

since entries of V are i.i.d. with mean 0. Now, there exists $M_0 > 0$ such that $\mathcal{E}_n = \{ \sum_{i=1}^N (X_0^i)^2 < N \cdot M_0 \}$ occurs eventually almost surely. Given $\mathbf{G}$ and $\mathbf{X}_0$, $V^\top \mathbf{G}^\top \boldsymbol{D}_w \mathbf{G} \mathbf{X}_0 \sim \mathcal{N}(0, \lambda(\mathbf{X}_0^\top \mathbf{G}^\top \boldsymbol{D}_w \mathbf{G} \mathbf{X}_0)/N)$. Moreover,

$$
\frac{\mathbf{X}_0^\top \mathbf{G}^\top \boldsymbol{D}_w \mathbf{G} \mathbf{X}_0}{N} \leq \|\boldsymbol{D}_w\|_2 \frac{\sum_{i=1}^N (X_0^i)^2}{N}.
$$

Thus, on the event $\mathcal{E}_n$, for any $\varepsilon > 0$,

$$
\mathbb{P}\left[\left| \frac{1}{\sqrt{N}} V^\top \mathbf{G}^\top \boldsymbol{D}_w \mathbf{G} \mathbf{X}_0 \right| > \varepsilon | \mathbf{G}, \mathbf{X}_0 \right] \leq 2 \exp\left(-\frac{N\varepsilon^2}{2M'} \right)
$$

for some universal constant $M' > 0$. The proof is complete on using the Borel Cantelli lemma. $\qquad\square$