

# HIGH-DIMENSIONAL LIMIT THEOREMS FOR SGD: EFFECTIVE DYNAMICS AND CRITICAL SCALING

GÉRARD BEN AROUS, REZA GHEISSARI, AND AUKOSH JAGANNATH

**ABSTRACT.** We study the scaling limits of stochastic gradient descent (SGD) with constant step-size in the high-dimensional regime. We prove limit theorems for the trajectories of summary statistics (i.e., finite-dimensional functions) of SGD as the dimension goes to infinity. Our approach allows one to choose the summary statistics that are tracked, the initialization, and the step-size. It yields both ballistic (ODE) and diffusive (SDE) limits, with the limit depending dramatically on the former choices. We show a critical scaling regime for the step-size, below which the effective ballistic dynamics matches gradient flow for the population loss, but at which, a new correction term appears which changes the phase diagram. About the fixed points of this effective dynamics, the corresponding diffusive limits can be quite complex and even degenerate. We demonstrate our approach on popular examples including estimation for spiked matrix and tensor models and classification via two-layer networks for binary and XOR-type Gaussian mixture models. These examples exhibit surprising phenomena including multimodal timescales to convergence as well as convergence to sub-optimal solutions with probability bounded away from zero from random (e.g., Gaussian) initializations. At the same time, we demonstrate the benefit of overparametrization by showing that the latter probability goes to zero as the second layer width grows.

## Part 1. Introduction and main results

### 1. INTRODUCTION

Stochastic gradient descent (SGD) is the go-to method for large-scale optimization problems in modern data science. It is often used to train complex parametric models on high-dimensional data. Since its introduction in [62], there has been a tremendous amount of work in analyzing its evolution.

In fixed dimensions, the asymptotic theory of SGD, and stochastic approximations more broadly, is by now classical. There have been works on path-wise limit theorems, such as functional central limit theorems and even large deviations principles [62, 49, 46, 39, 26, 13, 25, 11]. At the core of this line of work is the idea that in the limit where the step-size, or learning rate, tends to zero, the trajectory of SGD with a fixed loss function (appropriately rescaled in time) converges to the solution of gradient flow for the population loss with the same initialization. Recently there has been considerable interest in quantifying the rate of this trajectory-wise convergence to higher order, in terms of a diffusion approximation. Namely, there are many works developing asymptotic expansions of the trajectory in the learning rate [47, 41, 43, 2, 44]. Motivated by this, there is a rich line of work bounding the time to equilibrium for the associated diffusion approximation (as well as Langevin–type modifications) under uniform ellipticity assumptions [47, 59, 18, 77]. There is also an interesting line of work obtaining PDE limits in the “shallow network” regime where the dimension of the parameter space diverges but the dimension of the data remains constant: see e.g., [50, 63, 19, 69, 4].

In recent years, there has been considerable interest in understanding the *high-dimensional setting*, where one is constrained in the amount of data or the run-time of the algorithm due to the high-dimensional nature of the data and the complexity of the model being trained. In these regimes, one cannot simply take the learning rate to be arbitrarily small as this would force an unlimited sample

size and run-time. This is a common issue in high-dimensional statistics and the standard analytic approach is to study regimes where the sample size scales with the dimension of the problem [74, 75].

For SGD with constant learning rate, there has been recent progress on quantifying the dimension dependence of the sample complexity for various tasks on general (pseudo or quasi-) convex objectives [14, 15, 68, 53, 33, 24] and special classes of non-convex objectives [31, 71, 6]. There has also been important work on scaling limits as the dimension tends to infinity for the specific problems of linear regression [76, 55], Online PCA [76, 42], and phase retrieval [71] from random starts, and teacher-student networks [64, 65, 32, 73] and two-layer networks for XOR Gaussian mixtures [60] from warm starts. We also note that the study of high-dimensional regimes of gradient descent and Langevin dynamics have a history from the statistical physics perspective, e.g., in [21, 22, 67, 48, 17, 45].

We develop a unified approach to the scaling limits of SGD in high-dimensions with constant learning rate that allows us to understand a broad range of estimation tasks. One of course cannot develop a high-dimensional scaling limit for the full trajectory of SGD as the dimension of the underlying parameter space is growing. On the other hand, in practice, one is rarely interested in the full trajectory; instead one typically tracks the trajectory of various summary statistics of the algorithm's evolution, such as the loss, the amplitude of various weights, or correlations between the classifier and the ground truth (in a supervised setting). We show in Theorem 2.3 that under mild regularity assumptions, the evolution of these summary statistics converges as the dimension grows to the solution of a system of (possibly stochastic) differential equations. These *effective dynamics* depend dramatically on the initializations (warm vs. random or cold), the parameter regions in which one is developing the scaling limit, and the scaling of the step-size with the dimension.

In practice, SGD often exhibits two types of phases in training: *ballistic phases* where the summary statistics macroscopically change in value, and *diffusive phases*, where they fluctuate microscopically. (During training, the evolution can start with either, and can even alternate multiple times between these phases.) Our approach allows us to develop scaling limits for both types of phases.

In ballistic phases, the effective dynamics are given by an ordinary differential equation (ODE) and the finite-dimensional intuition that the summary statistics evolve under the gradient flow for the population loss is correct provided the (constant) learning rate is sufficiently small in the dimension. When the learning rate follows a certain *critical* scaling—matching scalings commonly used in the high-dimensional statistics literature—an additional correction term appears. At this critical scaling, the phase portrait deviates significantly from that of the population gradient flow. Furthermore, in microscopic neighborhoods of the fixed points of this ODE, the effective dynamics become diffusive and are given by SDEs which can exhibit a wide range of (possibly degenerate) behaviors. We note that the appearance of the correction term in the ballistic phase was first observed in the setting of teacher-student networks in [64, 65] and very recently investigated in detail in [73].

As a simple, first example of the departure of the effective dynamics in the critical step-size regime from the classical perspective, we study estimation for spiked matrix and tensor models in Section 3. In these models, the effective dynamics are exactly solvable and when the step-size scales critically with the dimension, in the ballistic phase the dynamics have additional fixed points as compared to the population gradient flow. The stability of these fixed points exhibit sharp transitions at special signal-to-noise ratios. When initialized randomly, the SGD starts in a microscopic neighborhood of an uninformative such fixed point, within which its effective dynamics become diffusive and exhibit a sharp transition between mean-reverting and mean-repellent Ornstein–Uhlenbeck (OU) processes.

To demonstrate our approach on more complex classification tasks typically studied using neural networks, we study a Gaussian mixture model analogue of the classical XOR problem in Section 5. (The XOR problem is arguably the canonical example of a decision boundary requiring at least two-layers to represent [51].) Here we find that the natural summary statistics are 22 dimensional, and their (ballistic) effective dynamics exhibit a rich phenomenology between some 39 connected fixed point regions of varying topological dimension. Surprisingly, we find that if we initialize the weights

of the network randomly (following a Gaussian distribution), then the algorithm will converge to a classifier with macroscopic generalization error with probability  $29/32$  and then follow a degenerate diffusion. On the other hand, we demonstrate the benefit of overparametrization, showing that as the width of the second layer grows, the probability of ballistically converging to a Bayes optimal classifier goes to 1; this is a mathematically rigorous example of the *lottery ticket hypothesis* of [30].

Before delving into the XOR problem, we first analyze the classification of a two component Gaussian mixture model in Section 4. This task is of course best solved using a one-layer network i.e., logistic regression, but with a two-layer network it exhibits some similar phenomenologies to the XOR problem while being more amenable to finer analysis. Here, we again find that if with random initial weights, with probability  $1/2$  the SGD will first converge to a classifier with macroscopic generalization error, and then follow a degenerate diffusion in a microscopic neighborhood of that set of unstable fixed points. We demonstrate this both empirically for positive signal-to-noise ratio and theoretically in the limit where the SNR tends to zero after the dimension tends to infinity.

While the above are a few examples that we are able to solve in detail for both their ballistic and diffusive limits, we expect our main theorem to be applicable and lend new insights into a host of other problems including SGD for finite-rank matrix and tensor PCA, and one and two-layer neural networks applied to mixtures of  $k$ -Gaussians for fixed  $k \geq 2$ . We leave this to future investigation. In this paper, we only consider the simplest variant of SGD, namely online SGD; we leave other variants involving batching and re-use to future works.

## 2. MAIN RESULT

Suppose that we are given a sequence of i.i.d. data  $Y_1, Y_2, \dots$  taking values in  $\mathcal{Y}_n \subseteq \mathbb{R}^{d_n}$  with law  $P_n \in \mathcal{M}_1(\mathbb{R}^{d_n})$ , and a loss function  $L_n : \mathcal{X}_n \times \mathcal{Y}_n \rightarrow \mathbb{R}$ , where here  $\mathcal{X}_n \subseteq \mathbb{R}^{p_n}$  is the parameter space. Consider online stochastic gradient descent with constant learning rate,  $\delta_n$ , which is given by

$$X_\ell = X_{\ell-1} - \delta_n \nabla L_n(X_{\ell-1}, Y_\ell),$$

with possibly random initialization  $X_0 \sim \mu_n \in \mathcal{M}_1(\mathcal{X}_n)$ . Our interest is in understanding this evolution,  $(X_\ell)$ , in the regime where both  $p_n$  and  $d_n \rightarrow \infty$  as  $n \rightarrow \infty$ . To this end, suppose that there is a finite collection of summary statistics of  $(X_\ell)$  whose evolution we are interested in. More precisely, suppose that we are given a sequence of functions  $\mathbf{u}_n \in C^1(\mathbb{R}^{p_n}; \mathbb{R}^k)$  for some fixed  $k$ , where  $\mathbf{u}_n(x) = (u_1^n(x), \dots, u_k^n(x))$ , and our goal is to understand the evolution of  $\mathbf{u}_n(X_\ell)$ .

To develop a scaling limit, we need some regularity assumptions on the relationship between how the step-size scales in relation to the loss, its gradients, and the data distribution. To this end let

$$H(x, Y) = L_n(x, Y) - \Phi(x) \quad \text{where} \quad \Phi(x) = \mathbb{E}[L_n(x, Y)].$$

In the following, we suppress the dependence of  $H$  on  $Y$  and instead view  $H$  as a random function of  $x$ , denoted  $H(x)$ . We let  $V(x) = \mathbb{E}[\nabla H(x) \otimes \nabla H(x)]$  be the covariance matrix for  $\nabla H$  at  $x$ .

We make two assumptions on the triple  $(\mathbf{u}_n, L_n, P_n)$  and the step size  $\delta_n$ . The first is an upper bound on the learning rate in terms of the regularity of the summary statistics. The second is our key assumption and asks that the summary statistic evolutions asymptotically close. These assumptions need not hold uniformly over the entire parameter space  $\mathbb{R}^{p_n}$ , only uniformly over pre-images of compact sets under  $\mathbf{u}_n$ . We start with the regularity assumption, ensuring tightness of trajectories of the summary statistics.

**Definition 2.1.** A triple  $(\mathbf{u}_n, L_n, P_n)$  is  $\delta_n$ -localizable with localizing sequence  $(E_K)_K$  if there is an exhaustion by compacts  $(E_K)_K$  of  $\mathbb{R}^k$ , and constants  $C_K$  (independent of  $n$ ) such that

- (1)  $\max_i \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\nabla^2 u_i^n\|_{\text{op}} \leq C_K \cdot \delta_n^{-1/2}$ , and  $\max_i \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\nabla^3 u_i^n\|_{\text{op}} \leq C_K$ ;
- (2)  $\sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\nabla \Phi\| \leq C_K$ , and  $\sup_{x \in \mathbf{u}_n^{-1}(E_K)} \mathbb{E}[\|\nabla H\|^8] \leq C_K \delta_n^{-4}$ ;

$$(3) \max_i \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \mathbb{E}[\langle \nabla H, \nabla u_i^n \rangle^4] \leq C_K \delta_n^{-2}, \text{ and} \\ \max_i \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \mathbb{E}[\langle \nabla^2 u_i^n, \nabla H \otimes \nabla H - V \rangle^2] = o(\delta_n^{-3}).$$

To help the reader parse this assumption, we provide an in-depth discussion of each of these items, along with examples to have in mind in Remark 1. For now, we make the crucial observation that (1)–(3) are all closed under decreasing the step-size  $\delta_n$  so for any reasonable task and family of summary statistics there will be a scaling of the step-size with  $n$  below which they will satisfy the conditions of Definition 2.1. For concreteness, summary statistics that are good to have in mind are correlations of the parameters with certain ground truth vectors,  $\ell^2$  norms of the parameters, and the population loss itself.

We now turn to our second assumption, that the limiting evolution equations for the family of summary statistics chosen close. Define the following first and second-order differential operators,

$$\mathcal{A}_n = \sum_i \partial_i \Phi \partial_i, \quad \text{and} \quad \mathcal{L}_n = \frac{1}{2} \sum_{i,j} V_{ij} \partial_i \partial_j. \quad (2.1)$$

Alternatively written,  $\mathcal{A}_n = \langle \nabla \Phi, \nabla \rangle$  and  $\mathcal{L}_n = \frac{1}{2} \langle V, \nabla^2 \rangle$ .

**Definition 2.2.** A family of summary statistics  $(\mathbf{u}_n)$  are **asymptotically closable** for learning rate  $\delta_n$  if  $(\mathbf{u}_n, L_n, P_n)$  are  $\delta_n$ -localizable with localizing sequence  $(E_K)_K$ , and furthermore there exist locally Lipschitz functions  $\mathbf{h} : \mathbb{R}^k \rightarrow \mathbb{R}^k$  and  $\Sigma : \mathbb{R}^k \rightarrow \mathbb{R}^{k \times k}$ , such that

$$\sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|(-\mathcal{A}_n + \delta_n \mathcal{L}_n)\mathbf{u}_n(x) - \mathbf{h}(\mathbf{u}_n(x))\| \rightarrow 0, \quad (2.2)$$

$$\sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\delta_n J_n V J_n^T - \Sigma(\mathbf{u}_n(x))\| \rightarrow 0. \quad (2.3)$$

In this case we call  $\mathbf{h}$  the *effective drift*, and  $\Sigma$  the *effective volatility*.

We are now ready to present our main result. For a function  $f$  and measure  $\mu$  we let  $f_*\mu$  denote the push-forward of  $\mu$ .

**Theorem 2.3.** Let  $(X_\ell^{\delta_n})_\ell$  be stochastic gradient descent initialized from  $X_0 \sim \mu_n$  for  $\mu_n \in \mathcal{M}_1(\mathbb{R}^{p_n})$  with learning rate  $\delta_n$  for the loss  $L_n(\cdot, \cdot)$  and data distribution  $P_n$ . For a family of summary statistics  $\mathbf{u}_n = (u_i^n)_{i=1}^k$ , let  $(\mathbf{u}_n(t))_t$  be the linear interpolation of  $(\mathbf{u}_n(X_{\lfloor t\delta_n^{-1} \rfloor}^{\delta_n}))_t$ .

Suppose that  $\mathbf{u}_n$  are asymptotically closable with learning rate  $\delta_n$ , effective drift  $\mathbf{h}$ , and effective volatility  $\Sigma$ , and that the pushforward of the initial data has  $(\mathbf{u}_n)_*\mu_n \rightarrow \nu$  weakly for some  $\nu \in \mathcal{M}_1(\mathbb{R}^k)$ . Then  $(\mathbf{u}_n(t))_t \rightarrow (\mathbf{u}_t)_t$  weakly as  $n \rightarrow \infty$ , where  $\mathbf{u}_t$  solves

$$d\mathbf{u}_t = \mathbf{h}(\mathbf{u}_t)dt + \sqrt{\Sigma(\mathbf{u}_t)}d\mathbf{B}_t. \quad (2.4)$$

initialized from  $\nu$ , where  $\mathbf{B}_t$  is a standard Brownian motion in  $\mathbb{R}^k$ .

The proof of Theorem 2.3 is provided in Section 6 and can be seen as a version of the classical martingale problem (see [70]) for high-dimensional stochastic gradient descent. We call the solution to (2.4) the *effective dynamics* of the summary statistics  $\mathbf{u}_n$ . The fact that  $\mathbf{h}, \Sigma$  are locally Lipschitz ensures that this solution is unique.

We end this subsection with discussion of the various scalings appearing in Definition 2.1.

*Remark 1.* The kinds of summary statistics that we most frequently have in mind for application are (1) linear functions of the parameter space  $\mathcal{X}_n$ , for instance the correlation with a unit vector, or some ground truth; (2) radial statistics, like the  $\ell^2$ -norm of the parameters, or some subset of the parameters; and (3) rescaled versions (usually blown up by  $\delta_n^{-1/2}$ ) of these near their fixed points, as described in Section 2.2. Regarding the item (1) in Definition 2.1, for linear functions, it trivially holds; for radial statistics, the Hessian is a block identity matrix, so item (1) holds as

long as  $\delta_n$  is  $O(1)$ ; therefore item (1) is most restrictive for rescalings of non-linear statistics, e.g.,  $u(x) = \delta_n^{-\alpha}(\|x\|^2 - 1)$  where it prevents consideration of this statistic with  $\alpha > 1/2$ .

Turning to item (2) of Definition 2.1, we comment that the regularity assumptions made on  $\Phi, L$  here are less restrictive than uniform Lipschitz assumptions common to the literature. In particular, we do not assume the population loss is Lipschitz everywhere, as we may have that  $\bigcup_K \mathbf{u}_n^{-1}(E_K)$  does not cover  $\mathcal{X}_n$ , nor does it imply uniform smoothness of  $H$  (and in turn  $L$ ) as we may (and will) be taking  $\delta_n \rightarrow 0$  with  $n$ .

Let us lastly motivate the scalings appearing in item (3), which ensure there is some independence between  $H$  and the values of  $\nabla u$  and  $\nabla^2 u$  at  $x$ . As a testbed, suppose that  $\nabla H(x)$  is a random vector with i.i.d. entries all of order 1. If  $u$  is a rescaled linear statistic, e.g.,  $\delta_n^{-1/2}\langle x, e_1 \rangle$  then the first bound of item (3) is saturated, and the second of course is trivial due to the linearity of  $u$ . The second bound is saturated by taking a rescaling of a radial statistic, e.g.,  $\delta_n^{-1/2}\|x\|^2$ , again assuming for maximal simplicity that  $\nabla H$  is an i.i.d. random vector with order one entries. In fact, the second part of item (3) could be dropped at the expense of more complicated diffusion coefficients in limiting SDE's: see Remark 2.

*Remark 2.* While we discussed above the reasons for which the various scalings of Definition 2.1 were selected, it is interesting to ask what changes in Theorem 2.3 should certain of the assumptions of Definition 2.1 be violated. Most of the assumed bounds in the definition of localizability are used to establish tightness and ensure higher order terms in Taylor expansions vanish in the  $n \rightarrow \infty$  limit. In principle the second assumption in item (3) of Definition 2.1 could be dropped; in that case, the same quantity is still ensured to be  $O(\delta^{-3})$  by the other localizability assumptions. Then Theorem 2.3 would still apply, but the limiting diffusion matrix would be the  $n \rightarrow \infty$  limit (assuming it exists) of

$$\begin{aligned} & \delta JVJ^T + \delta^2 \mathbb{E}[\langle \nabla H, J \rangle \otimes \langle \nabla^2 \mathbf{u}, \nabla H \otimes \nabla H - V \rangle] + \delta^2 \mathbb{E}[\langle \nabla^2 \mathbf{u}, \nabla H \otimes \nabla H - V \rangle \otimes \langle \nabla H, J \rangle] \\ & + \delta^3 \mathbb{E}[\langle \nabla^2 \mathbf{u}, \nabla H \otimes \nabla H - V \rangle^{\otimes 2}], \end{aligned}$$

as opposed to simply the limit of  $\delta JVJ^T$ .

Generically, the choice of summary statistics to which to apply Theorem 2.3 depends both on the quantities one is interested in, and the specifics of the task. In our examples, the choices are natural: correlations with ground truth vectors, finite numbers of final layer weights, and  $\ell^2$  norms of the parameters. In less structured settings, the choice of summary statistics may be more open-ended. One could start with a summary statistic of interest, like the projection in a principal subspace of an empirical matrix (a covariance or, as suggested experimentally in e.g., [66, 54], a Hessian), or the population loss itself. Then from that statistic, one would determine the other statistics needed to build an asymptotically closed family per Definition 2.2.

**2.1. Comparison to fixed dimensional perspective: critical v.s. subcritical step-sizes.** Let us compare this with the classical limit theory of SGD in fixed dimension. For the sake of this discussion, suppose that not only does (2.2) hold, but each of the two terms  $\mathcal{A}_n \mathbf{u}$  and  $\delta_n \mathcal{L}_n \mathbf{u}$  (recall (2.1)) individually admit  $n \rightarrow \infty$  limits: namely that there exists  $\mathbf{f}, \mathbf{g} : \mathbb{R}^k \rightarrow \mathbb{R}^k$  such that

$$\sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\mathcal{A}_n \mathbf{u}_n(x) - \mathbf{f}(\mathbf{u}_n(x))\| \rightarrow 0, \quad (2.5)$$

$$\sup_{x \in \mathbf{u}_n^{-1}(E_K)} \|\delta_n \mathcal{L}_n \mathbf{u}_n(x) - \mathbf{g}(\mathbf{u}_n(x))\| \rightarrow 0, \quad (2.6)$$

in which case, evidently (2.2) holds with  $\mathbf{h} = -\mathbf{f} + \mathbf{g}$ . When (2.5) and (2.6) both hold, we call  $\mathbf{f}, \mathbf{g}$  and  $\Sigma$  the **population drift**, the **population corrector**, and the **diffusion matrix** of  $\mathbf{u}$  respectively. From the fixed dimensional perspective, when (2.5) holds, one predicts  $\mathbf{u}$  to solve

$$d\mathbf{u}_t = -\mathbf{f}(\mathbf{u}_t)dt, \quad (2.7)$$

with initial data  $\mathbf{u}_0 \sim \mathbf{u}_* \mu$ . as this is its evolution under gradient descent on the population loss  $\Phi$ . Evidently this perspective only applies in the high-dimensional limit of Theorem 2.3 if both the population corrector  $\mathbf{g}$  and the diffusion matrix  $\Sigma$  are zero. We find that for any triple  $(\mathbf{u}_n, L_n, P_n)$ , there is a scaling of the learning rate  $\delta_n$  with  $n$  below which  $\mathbf{g} = \Sigma = 0$ , and the effective dynamics agree with the population dynamics (2.7) (we call this the **sub-critical** scaling regime, where the classical perspective applies), and a **critical** scaling regime in which  $\mathbf{g}$  and  $\Sigma$  may be non-zero, and the high-dimensionality induces non-trivial corrections to  $\mathbf{f}$ . (In the case of teacher–student networks, the terms  $\mathbf{f}$  and  $\mathbf{g}$  can be compared to the “learning” and “variance” terms in Eq. (14a) of [73].)

To see this, notice that if the triple  $(\mathbf{u}_n, L_n, P_n)$  is  $\delta_n$ -localizable for some  $\delta_n \rightarrow 0$ , then it is also  $\delta'_n$ -localizable for every sequence  $\delta'_n = O(\delta_n)$ . If furthermore (2.3) and (2.5)–(2.6) hold for  $\delta_n$  with some  $\mathbf{f}, \mathbf{g}$  and  $\Sigma$ , then these limits also exists for  $\delta'_n = o(\delta_n)$  with the same  $\mathbf{f}$  but with  $\mathbf{g} = \Sigma = 0$ . As such, there can be exactly one scaling of  $\delta_n$  with  $n$  at which  $\mathbf{g}$  or  $\Sigma$  may be non-zero, and for all smaller scales of  $\delta_n$ , the fixed-dimensional perspective of (2.7) applies.<sup>1</sup>

**2.2. Ballistic vs. diffusive behavior of effective dynamics.** In all of our examples, the diffusion matrix for the effective dynamics of the most natural choice of summary statistics is zero even in the critical scaling regime where  $\mathbf{h} \neq \mathbf{f}$ . We call this the **ballistic limit**. In this case, the effective dynamics of the summary statistics is given by the ODE system

$$d\mathbf{u}_t = \mathbf{h}(\mathbf{u}_t)dt. \quad (2.8)$$

In these settings, the phase portrait of the summary statistics is asymptotically that of this flow.

Note that by construction of the scaling limit, the phase portrait of the ballistic limit only describes the evolution of summary statistics on length-scales that are order 1 and number of iterations that are order  $1/\delta_n$ . If one is then interested in the evolution of  $\mathbf{u}_n$  in microscopic  $o(1)$  neighborhoods of the fixed points of the ballistic effective dynamics of (2.8), Theorem 2.3 also allows one to develop separate **diffusive limits** there.

To study diffusive regimes, one must apply Theorem 2.3 to re-centered and re-scaled summary statistics,  $\tilde{\mathbf{u}}_n(t) = \delta_n^{-\alpha}(\mathbf{u}_n(t) - \mathbf{u}_*)$  where  $\mathbf{u}_*$  is a fixed point of (2.8).<sup>2</sup>

To apply Theorem 2.3,  $\alpha$  must be chosen appropriately so that the triple  $(\tilde{\mathbf{u}}_n(t), L_n, P_n)$  is  $\delta_n$ -localizable and to pick out the next order drifts for  $\tilde{u}$ —the first order term being zero microscopically close to  $\mathbf{u}_*$ —and such that the initial data still converges  $(\tilde{\mathbf{u}}_n)_*\mu_n \rightarrow \tilde{\nu}$ .

This then leads to the **rescaled effective dynamics** of the summary statistics  $\mathbf{u}_n$  near  $\mathbf{u}_*$ :

$$d\tilde{\mathbf{u}}_t = \tilde{\mathbf{h}}(\tilde{\mathbf{u}}_t)dt + \tilde{\Sigma}^{1/2}(\tilde{\mathbf{u}}_t)d\mathbf{B}_t \quad \text{with } \tilde{\mathbf{u}}_0 \sim \tilde{\nu}. \quad (2.9)$$

The rescaled effective dynamics are similar in spirit to diffusion one typically finds for the evolution of SGD near critical points in fixed dimensions. However, we note two important differences as compared to this perspective. Firstly, since this is a high-dimensional limit of general summary statistics, (2.9) applies in a neighborhood of a fixed point of the effective ODE system (2.8), rather than the population dynamics (2.7). Secondly, in many examples (indeed all the ones we study) the SDE’s we get are degenerate, so that uniform ellipticity assumptions typically used to understand hitting and mixing times in these regimes do not apply. The degeneracies can take various forms, with  $\tilde{\Sigma}$  sometimes being rank deficient in the entire  $\tilde{\mathbf{u}}$ -space, and sometimes vanishing completely as  $\tilde{\Sigma}$  approaches certain distinguished points, for instance  $\mathbf{u}_*$ . The implications of such degeneracies can be severe, as degenerate diffusions can be absorbed for arbitrarily long times by their *unstable* fixed points (c.f. the simple case of a 1D geometric Brownian motion).

*Remark 3* (Training at the edge of stability and critical scaling). In [20], it was empirically observed that the best training for neural networks does not occur when step-sizes are small enough for the

---

<sup>1</sup>Note that if  $\delta_n = o(\delta'_n)$ , then limiting  $\mathbf{g}, \Sigma$  may not exist for  $\delta'_n$ , so there is no super-critical regime.

<sup>2</sup>One might also wish to rescale time like  $\delta_n^{-\beta}$ , where  $\beta$  may depend on  $t$ ; we leave this to future work.

classical gradient flow approximation to be valid. Instead, it occurs *at the edge of stability* where the step size is just small enough for the training to remain stable. Here, the loss fluctuates for some time before eventually converging to lower values than it would with smaller step size. This critical step size scaling is defined via the *sharpness*, namely the largest eigenvalue of the training loss Hessian. For a selection of recent theoretical investigations of this phenomenon see, e.g., [1, 23, 5, 78].

While sharpness and edge of stability do not have direct analogues in the context of online SGD, a qualitatively similar phenomenon can be seen by taking the population loss as a summary statistic. The critical scaling of the learning rate with dimension discussed in Sections 2.1–2.2 constrains the step size in terms of the top eigenvalue of the Hessian of the loss. With this scaling, the population loss fluctuates near critical regions of its ballistic flow, allowing it to escape the critical region, whereas with a sub-critical learning rate the population loss stays stuck. We leave more detailed investigation of this connection to edge-of-stability phenomena for SGD to future investigation.

## Part 2. Examples

In the following sections, we demonstrate Theorem 2.3 on a range of popular examples of high-dimensional statistical tasks. We begin first in Section 3 by presenting an application to a widely studied problem of high-dimensional estimation: namely, de-noising a rank one tensor that has been corrupted additively by Gaussian noise. We then turn to classification. Our aim in these examples is to demonstrate the applicability of our result to the analysis of multi-layer neural networks. To this end we analyze the training dynamics of a two-layer neural network for two canonical classification tasks, namely classification of a symmetric, binary gaussian mixture model (Section 4) and classification of a Gaussian analogue of the XOR problem of Minsky–Papert (Section 5).

### 3. MATRIX AND TENSOR PCA

**3.1. Model and background.** Consider the problem of de-noising a rank one tensor that has been corrupted additively by Gaussian noise via SGD. A popular statistical model of this task is the *spiked tensor model* [61]. Suppose that we are given i.i.d. samples of data of the form

$$Y^\ell = \lambda v^{\otimes k} + W^\ell$$

where  $W^\ell$  are i.i.d. copies of a  $k$ -tensor whose entries are i.i.d. standard Gaussians,  $v \in \mathbb{R}^n$  is a unit vector, and  $\lambda = \lambda_n > 0$  is the signal-to-noise ratio. Our goal is to infer  $v$ .

In the case  $k = 2$ , this is a version of the well-known spiked matrix model of PCA [37] for which there is, by now, a substantial literature regarding the statistical thresholds. For a necessarily small selection see, e.g., [7, 56, 52, 58, 27]. For related work on online learning in this context see, e.g., [68]. Of particular interest in this direction is the well-known phase transition at  $\lambda = 1$  for estimation in this problem, which was determined first for Wishart ensembles in [7] and subsequently for this setting in [29, 16, 12]. As we will see in Section 3.3 below, we find a dynamical analogue of this transition at  $\lambda = 1$ .

The case  $k \geq 3$  was introduced by Montanari and Richard [61] as a natural generalization of the spiked matrix models for estimation (and testing) problems where the data has multiple indices or requires higher moments. Here there has been a large literature on the statistical thresholds for estimation and testing, see, e.g., [52, 57, 10, 40, 36]. In this setting, there has also been a tremendous literature on the computational aspects of this problem as it is viewed as a important example of a model with a *statistical-computational gap*, namely, a setting where there is a gap between the regimes of statistical and computational tractability. See, e.g., [61, 35, 34, 40, 38, 8, 6].

We begin with these examples as their effective dynamics are particularly simple to analyze. In particular, they are exactly solvable and only require two summary statistics, a correlation observable and a radial term. Even with this relative simplicity, we encounter a wide range of

ODE and SDE limits. In particular, as mentioned above, we find dynamical phase transitions corresponding to the aforementioned thresholds in these models. For our analysis we will focus exclusively on the most interesting, critical step-size scaling which corresponds to the *proportional asymptotics* regime from the random matrix theory literature.

**3.2. Analysis.** We take as loss the (negative) log-likelihood<sup>3</sup> namely,

$$L(x, Y) = \|Y - x^{\otimes k}\|^2.$$

The pair

$$m = m(x) := \langle x, v \rangle \quad \text{and} \quad r_{\perp}^2 = r_{\perp}^2(x) := \|x - mv\|^2 = \|x\|^2 - m^2$$

are such that  $\Phi(x) = -2\lambda m^k + (r_{\perp}^2 + m^2)^k + c$ , and the law of  $L$  only depends on them.

In our normalization with  $\lambda > 0$  fixed, the regime  $\delta_n = o(1/n)$  is sub-critical and the regime  $\delta_n = \Theta(1/n)$  is critical.<sup>4</sup> We focus our presentation on the most interesting regime, namely the critical scaling regime of  $\delta_n = c_{\delta}/n$  for some constant  $c_{\delta}$ . Recalling the relation between number of samples and step-size, we see that this regime corresponds to the proportional asymptotics regime most studied in the random matrix theory literature where the above-mentioned transition for the top eigenvalue occurs. Note, however, that the limits in the subcritical regime are in all cases recovered by taking the  $c_{\delta} \downarrow 0$  limits of the ODE's/SDE's of the critical regime.

For notational simplicity, let  $R^2 := m^2 + r_{\perp}^2$ . We consider the pair  $\mathbf{u}_n = (u_1, u_2) = (m, r_{\perp}^2)$ , for which Theorem 2.3 yields the following effective dynamics.

**Proposition 3.1.** *Fix  $k \geq 2$ ,  $\lambda > 0$ ,  $c_{\delta} > 0$  and let  $\delta_n = c_{\delta}/n$ . Then  $\mathbf{u}_n(t)$  converges as  $n \rightarrow \infty$  to the solution of the following ODE initialized from  $\lim_{n \rightarrow \infty} (\mathbf{u}_n)_{*} \mu_n$ :*

$$dm = 2m(\lambda km^{k-2} - kR^{2(k-1)})dt, \quad dr_{\perp}^2 = -4kR^{2(k-1)}(r_{\perp}^2 - c_{\delta})dt. \quad (3.1)$$

We are able to identify and classify the set of fixed points of this effective dynamics. We focus on the critical step-size regime with  $c_{\delta} = 1$  where one sees from (3.1) that  $r_{\perp}^2 \rightarrow 1$ , where the problem in the matrix case is most directly related to an eigenvalue problem (see Section 7 for the generic  $c_{\delta}$  dependencies). Throughout the following, we use the following notion of stability/unstability of a set of fixed points of an ODE.

**Definition 3.2.** We call a set of fixed points  $U$  for an ODE *stable* if for every  $\epsilon > 0$ , for every  $u \in B_{\epsilon}(U)$ , the solution of the ODE with initialization  $u$  converges to some point in  $U$  as  $t \rightarrow \infty$ . Otherwise, we call  $U$  *unstable*.

**Proposition 3.3.** *Eq. (3.1) has isolated fixed points classified as follows. Let  $\lambda_c(k)$  be as in (7.6) and  $m_{\dagger}(k, \lambda) \leq m_{\star}(k, \lambda)$  be as in (7.7) (if  $k = 2$ ,  $\lambda_c = 1$  and  $m_{\dagger} = m_{\star} = \sqrt{\lambda - 1}$ ):*

- (1) *An unstable fixed point at  $(0, 0)$  and a fixed point at  $(0, 1)$ ; if  $k = 2$ ,  $(0, 1)$  is stable if  $\lambda < \lambda_c(2)$  and unstable if  $\lambda > \lambda_c(2)$ ; if  $k > 2$   $(0, 1)$  is always stable.*
- (2) *If  $\lambda > \lambda_c(k)$ : when  $k = 2$ , two stable fixed points at  $(\pm m_{\star}(2), 1)$ . When  $k \geq 3$ , two unstable fixed points at  $(\pm m_{\dagger}(k), 1)$  and two stable fixed points at  $(\pm m_{\star}(k), 1)$ .*

*Remark 4.* The presence of two pairs of fixed points when  $k \geq 3$  with non-zero correlation with  $v$  may seem surprising—indeed it indicates that even some warm starts will fail to attain good correlation with the signal when  $\lambda$  is finite. This is an interesting consequence of the corrector in (3.1) and if one tracks the  $c_{\delta}$  dependence in the above, the fixed point  $m_{\dagger}$  goes to zero as  $c_{\delta} \rightarrow 0$  and this barrier to recovery from warm starts vanishes as one approaches sub-critical step-sizes.

---

<sup>3</sup>Note that one might also add additional penalty terms. The case of a ridge penalty is treated in Section 7.

<sup>4</sup>Note that with different scalings of  $\lambda_n$ , the critical learning rate changes.

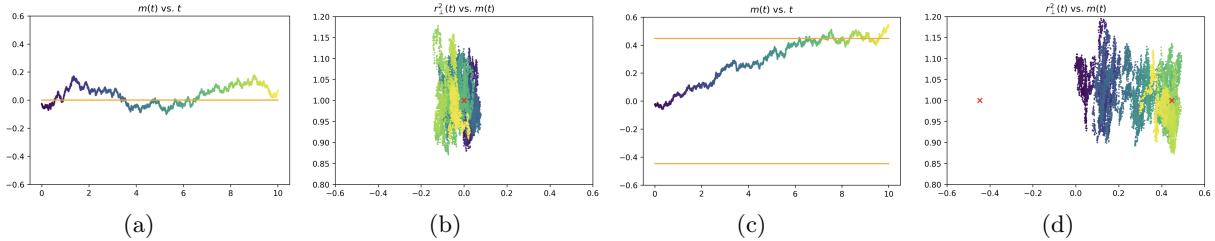


FIGURE 1. Matrix PCA summary statistics in dim.  $n = 1500$  run for  $10n$  steps at  $\lambda = 0.8 < \lambda_c$  in (a)–(b) and  $\lambda = 1.2 > \lambda_c$  in (c)–(d). Here,  $\times$  and  $-$  mark the stable fixed points of the systems. (a) and (c) demonstrate the mean-reverting and mean-repellent OU processes that arise as diffusive limits of the  $m$  variable, and (b) and (d) depict the trajectories in  $(m, r_\perp^2)$  space.

**3.3. A dynamical analogue of the BBP transition.** Let us now consider a rescaling of  $\mathbf{u}_n$  in a microscopic neighborhood of the saddle set  $m = 0$ . This captures the initial phase from a random start: if  $\mu_n \sim \mathcal{N}(0, I_n/n)$ , then  $(\mathbf{u}_n)_*\mu_n \rightarrow \delta_{(0,1)}$  weakly. Now rescale and let  $\tilde{\mathbf{u}}_n = (\tilde{m}, r_\perp^2) = (\sqrt{n}m, r_\perp^2)$ . Evidently,  $\tilde{\nu} = \lim_n (\tilde{\mathbf{u}}_n)_*\mu_n = \mathcal{N}(0, 1) \otimes \delta_1$ .

**Proposition 3.4.** *Fix  $k \geq 2$ ,  $\lambda > 0$  and  $\delta_n = 1/n$ . Then  $\tilde{\mathbf{u}}_n(t)$  converges as  $n \rightarrow \infty$  to the solution of the following SDE initialized from  $\tilde{\nu}$ :*

$$d\tilde{m} = 2\tilde{m}(2\lambda\mathbf{1}_{k=2} - kr_\perp^{2(k-1)})dt + 2(kr_\perp^{2(k-1)})^{1/2}dB_t \quad dr_\perp^2 = -4kr_\perp^{2(k-1)}(r_\perp^2 - 1)dt. \quad (3.2)$$

We see that  $r_\perp^2$  now solves an autonomous ODE which converges exponentially to 1. When  $k = 2$ , as  $t$  tends to  $\infty$ , the equation for  $\tilde{m}$  behaves like

$$d\tilde{m} = 4(\lambda - 1)\tilde{m}dt + 2\sqrt{2}dB_t.$$

This is an OU process which is mean-reverting when  $\lambda < 1$  and mean-repellent when  $\lambda > 1$ . By stitching together the prelimits of these OU processes at a sequence of scales interpolating between that of  $\tilde{\mathbf{u}}_n$  and  $\mathbf{u}_n$ , we expect that one could show that for any  $\lambda > 1$ , SGD reaches the stable fixed points at  $(\pm m_*(2), 1)$  in  $O(n \log n)$  steps (with precise asymptotics, etc.), while when  $\lambda < 1$ , the mean-reverting nature of the OU suggests it needs a much larger number of samples in order to correlate with the vector  $v$ . See Figure 1 for an overview, and Figures 2–3 for more refined numerical verification of this intuition.

**3.4. On the sample complexity of tensor PCA.** When  $k \geq 3$ , SGD is known to require a polynomially diverging sample complexity or  $\lambda$  in order to solve the tensor PCA problem [6]. Accordingly, when  $\lambda$  is kept finite in  $n$ , the expression for  $\tilde{m}$  in (3.2) is *always* a mean-reverting OU-type process. Interestingly, one can also capture the (diverging) signal-to-noise threshold for SGD to recover  $v$  in tensor PCA by our methods. Indeed, for  $k \geq 3$  if one considers  $\lambda_n = \Lambda n^{(k-2)/2}$  (matching the predicted gradient-based algorithm threshold from [8]),  $\tilde{\mathbf{u}}_n$  would instead converge to the solution of

$$d\tilde{m} = 2\tilde{m}(k\Lambda - kr_\perp^{2(k-1)})dt + 2(kr_\perp^{2(k-1)})^{1/2}dB_t \quad dr_\perp^2 = -4kr_\perp^{2(k-1)}(r_\perp^2 - 1)dt, \quad (3.3)$$

which transitions between mean-reverting and mean-repellent at  $\Lambda_c(k) = 1$ , as in  $k = 2$ .

We only considered a few specific choices of summary statistics in the above, and the strength of Theorem 2.3 derives from its general applicability. As demonstrations, let us mention a few other examples that we would expect to be of interest in the study of SGD for matrix and tensor PCA. The first example is a limiting ballistic limit theorem for the evolution of the population loss  $\Phi(x)$ .

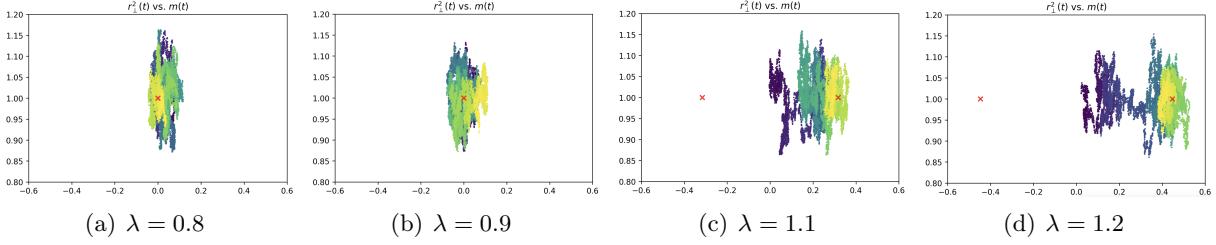


FIGURE 2. Matrix PCA in dimension  $n = 2000$  with various values of  $\lambda$  near the critical  $\lambda = 1$ . Depicted is the evolution of summary statistics  $(m, r_{\perp}^2)$  for  $10n$  steps of SGD initialized randomly.

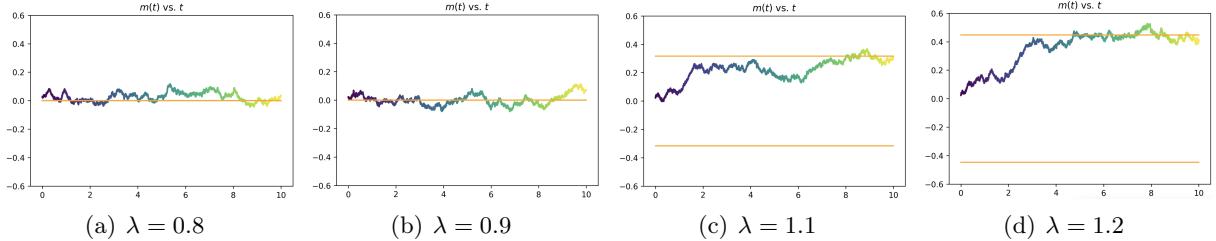


FIGURE 3. Matrix PCA in dimension  $n = 2000$  with various values of  $\lambda$  near the critical  $\lambda = 1$ . Depicted is the evolution of  $m(t)$  for  $10n$  steps of SGD initialized randomly.

The population loss can be taken added to the family of summary statistics in our  $\delta_n$ -localizable triple; in the case of  $k$ -tensor PCA, this yields,

$$d\Phi = \left( -4k^2m^2(\lambda^2m^{2(k-2)} - 2\lambda m^{k-2}R^{2k-2} + R^{4k-4}) - 4k^2R^{4(k-1)}(r_{\perp}^2 - c_{\delta}) \right) dt. \quad (3.4)$$

**3.5. A finer diffusive limit theorem at a random start.** The second example is a diffusive limit theorem near the fixed point  $(m, r_{\perp}^2) = (0, 1)$  (as opposed to (3.3) where we blew up only the  $m$  variable about the saddle set  $m = 0$  and therefore only  $\tilde{m}$  was moving diffusively). In order to do so, we consider the scaling limit of the pair  $(\tilde{m}, \tilde{r}_{\perp}) = (\sqrt{n}m, \sqrt{n}(r_{\perp}^2 - 1))$  and find the following limit:

$$d\tilde{m} = 2k(\lambda \mathbf{1}_{k=2}\tilde{m}^{k-1} - 1)dt + 2\sqrt{k}dB_t^{(1)}, \quad d\tilde{r}_{\perp}^2 = -4k\tilde{r}_{\perp}^2 dt + 2\sqrt{k(k-1)}dB_t^{(2)}. \quad (3.5)$$

Interestingly, with this double rescaling, the  $n \rightarrow \infty$  limit yields a pair of OU processes that are decoupled, namely, each of their drifts are autonomous and their stochastic parts independent. This pair of independent OU processes is depicted in Figures 4–5.

#### 4. TWO-LAYER NETWORKS FOR CLASSIFYING A BINARY GAUSSIAN MIXTURE

**4.1. Model and Background.** As a warm-up to the XOR problem that we will consider in Section 5, we consider the problem of supervised classification of a binary Gaussian mixture model (binary GMM) which is defined as follows. Suppose that we are given i.i.d. samples of the form  $Y = (y, X)$ , where  $y$  is a  $\{0, 1\}$ -valued  $Ber(1/2)$  random variable and, conditionally on  $y$ , we have

$$X \sim \mathcal{N}((2y - 1)\mu, I/\lambda),$$

where  $\mu \in \mathbb{R}^N$  is a fixed unit vector,  $I$  is the identity on  $\mathbb{R}^N$ , and  $\lambda > 0$  is the signal-to-noise ratio. Here,  $y$  is called the class label and  $X$  is called the data. Our goal is to construct an estimator,  $\hat{y} = \hat{y}(X)$ , of the class label,  $y$ , which depends on the data,  $X$ , alone.

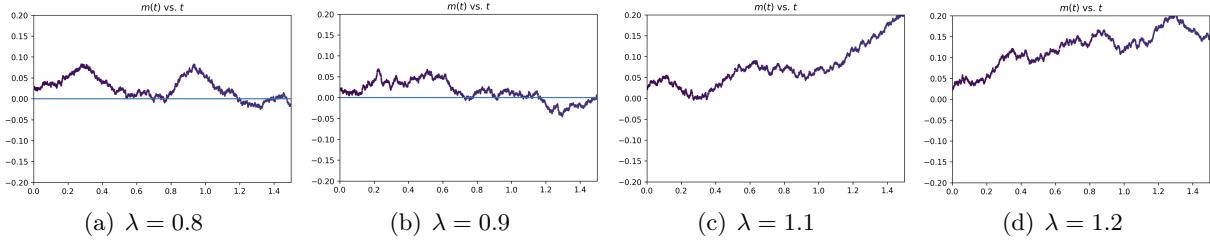


FIGURE 4. Matrix PCA in dimension  $n = 2000$  with various values of  $\lambda$  near the critical  $\lambda = 1$ . Depicted is the evolution of summary statistic  $m(t)$  zoomed in about an  $O(n^{-1/2})$  window of  $m = 0$  for  $1.5 * n$  steps of SGD initialized randomly. In (a)–(b) one sees stable OU processes, and in (c)–(d) one sees unstable OU processes.

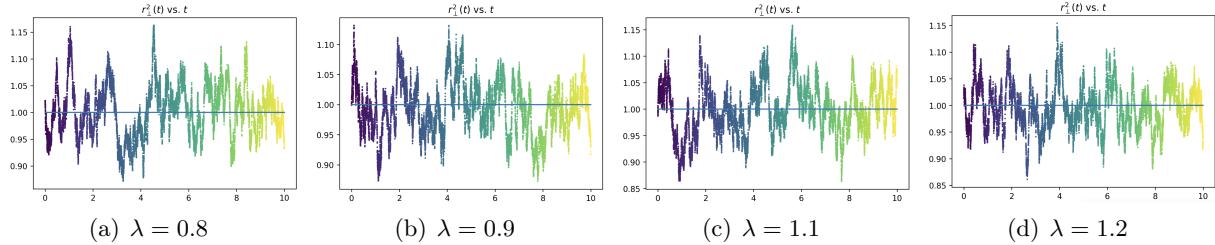


FIGURE 5. Matrix PCA in dimension  $n = 2000$  with various values of  $\lambda$  near the critical  $\lambda = 1$ . Depicted is the evolution of summary statistic  $r_{\perp}^2(t)$  for  $10n$  steps of SGD initialized randomly. This follows a stable OU process independent of  $\lambda$ .

It is classical [3] that the Bayes optimal estimator in this setting is given by  $\hat{y} = \text{sgn}(\mu \cdot x)$ . Furthermore, this estimator can be achieved by (a rounding of) the output of a single layer neural network trained using the binary-cross-entropy loss (4.1). This is also called logistic regression. The single-layer setting can be easily analyzed via our framework. Our focus here, however, is to demonstrate our analysis on multi-layer neural networks.

To that end, we consider now the same setting, except that we will estimate the class labels using a simple two-layer neural network. (Note that the Bayes' optimal estimator is still expressible by this architecture.) At first glance, this may seem an elementary setting with little to say. However as we will see, even in this simple setting surprising behaviour can occur in the high-dimensional setting which runs counter to common intuition. Furthermore, as we will see in Section 5, the phenomena occurring here also appear in richer problems such as the XOR problem.

**4.2. Analysis.** For the sake of concreteness, we consider classification via the following architecture (though our techniques generalize to other settings *mutatis mutandis*): The first layer has weights  $(W_1, W_2) \in \mathbb{R}^N \times \mathbb{R}^N$  and ReLu activation,  $g(x) = x \vee 0$ ; and the second layer has weights  $v_1, v_2 \in \mathbb{R}$  and sigmoid activation,  $\sigma(x) = 1/(1 + e^{-x})$ . The output of the multi-layer network is then  $\sigma(v \cdot g(WX))$ . Our parameter space is then  $\mathcal{X}_n = \mathbb{R}^{2N+2}$  and we therefore take  $n = 2N + 2$  when applying Theorem 2.3.

As we are interested in supervised classification, we take the usual *binary cross-entropy loss* with  $\ell^2$  regularization. In our setting, this reduces to optimizing

$$L((v_i, W_i)_{i \in \{1,2\}}; (y, X)) = -yv \cdot g(WX) + \log(1 + e^{v \cdot g(WX)}) + p(v, W), \quad (4.1)$$

where  $g$  is applied component wise and  $p(v, W) := (\alpha/2)(\|v\|^2 + \|W\|^2)$ .

It can be shown (see Lemma 8.1) that the law of the loss at a given point,  $(v, W) \in \mathcal{X}_n$ , depends only on the 7 summary statistics,

$$\mathbf{u}_n = (v_1, v_2, m_1, m_2, R_{11}^\perp, R_{12}^\perp, R_{22}^\perp), \quad (4.2)$$

where  $m_i = W_i \cdot \mu$  and  $R_{ij}^\perp = W_i^\perp \cdot W_j^\perp$  with  $W_i^\perp = W_i - m_i \mu$  denoting the part of  $W_i$  orthogonal to  $\mu$ . For a point,  $(v, W) \in \mathcal{X}_n$ , let

$$\begin{aligned} \mathbf{A}_i^\mu &= \mathbb{E}[X \cdot \mu \mathbf{1}_{W_i \cdot X \geq 0}(\sigma(v \cdot g(WX)) - y)], & \mathbf{A}_{ij}^\perp &= \mathbb{E}[X \cdot W_j^\perp \mathbf{1}_{W_i \cdot X \geq 0}(\sigma(v \cdot g(WX)) - y)], \\ \mathbf{B}_{ij} &= \mathbb{E}[\mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0}(\sigma(v \cdot g(WX)) - y)^2]. \end{aligned} \quad (4.3)$$

By similar reasoning to Lemma 8.1, it can be seen that these are functions only of  $\mathbf{u}_n$ , and we denote them as such, e.g.,  $\mathbf{A}_i^\mu = \mathbf{A}_i^\mu(\mathbf{u}_n)$ . See Section 8. The critical scaling for  $\delta$  is then of order  $\Theta(1/n)$  and we obtain the following.

**Proposition 4.1.** *Let  $\mathbf{u}_n$  be as in (4.2) and fix any  $\lambda > 0$  and  $\delta_n = c_\delta/N$ . Then  $\mathbf{u}_n(t)$  converges to the solution of the ODE system,  $\dot{\mathbf{u}}_t = -\mathbf{f}(\mathbf{u}_t) + \mathbf{g}(\mathbf{u}_t)$ , initialized from  $\lim_{n \rightarrow \infty} (\mathbf{u}_n)_* \mu_n$ , with:*

$$\begin{aligned} f_{v_i} &= m_i \mathbf{A}_i^\mu(\mathbf{u}) + \mathbf{A}_{ii}^\perp(\mathbf{u}) + \alpha v_i, & f_{m_i} &= v_i \mathbf{A}_i^\mu(\mathbf{u}) + \alpha m_i, \\ f_{R_{ij}^\perp} &= v_i \mathbf{A}_{ij}^\perp(\mathbf{u}) + v_j \mathbf{A}_{ji}^\perp(\mathbf{u}) + 2\alpha R_{ij}^\perp, \end{aligned}$$

and correctors  $g_{v_i} = g_{m_i} = 0$ ,  $g_{R_{ij}^\perp} = c_\delta \frac{v_i v_j}{\lambda} \mathbf{B}_{ij}$  for  $i, j = 1, 2$ .

**4.3. Low variance asymptotics.** Due to the Gaussian integrals defining  $\mathbf{f}, \mathbf{g}$ , it is difficult to analyze the ODE system defined by Proposition 4.1, let alone any rescaled effective dynamics. For ease of analysis, we next send  $\lambda \rightarrow \infty$  corresponding to a small noise regime for the Gaussian mixture. We emphasize that this limit is taken after  $n \rightarrow \infty$  and therefore is still approximately on the critical scale of  $\lambda = \Theta(1)$  at which there is a transition in the existence of any fixed point which is a good classifier. In particular, if  $\lambda = \lambda_n$  is any diverging sequence, then the limiting effective dynamics would exactly match that attained by now sending  $\lambda \rightarrow \infty$ . In Figure 6, we demonstrate numerically that the following predicted fixed points from the  $\lambda \rightarrow \infty$  limit match those arising at finite large  $N$  and  $\lambda > 0$ .<sup>5</sup>

**Proposition 4.2.** *The  $\lambda \rightarrow \infty$  limit of the ODE system of Proposition 4.1 is given by*

$$\dot{m}_i = \begin{cases} \frac{v_i}{2} \sigma(-v \cdot m) - \alpha m_i & m_1 m_2 > 0 \\ \frac{v_i}{2} \sigma(-v_i m_i) - \alpha m_i & \text{else} \end{cases}, \quad \dot{v}_i = \begin{cases} \frac{m_i}{2} \sigma(-v \cdot m) - \alpha v_i & m_1 m_2 > 0 \\ \frac{m_i}{2} \sigma(-v_i m_i) - \alpha v_i & \text{else} \end{cases},$$

and  $\dot{R}_{ij}^\perp = -2\alpha R_{ij}^\perp$ . The fixed points of this system are classified as follows. All fixed points have  $R_{ij}^\perp = 0$  and  $m_i = v_i$  for  $i, j = \{1, 2\}$ . In  $(v_1, v_2)$ , the coordinates are classified by

- (1) A fixed point at  $(v_1, v_2) = (0, 0)$  that is stable if  $\alpha > 1/4$ ;
- (2) If  $\alpha < 1/4$ , two unstable sets of fixed points at the quarter-circles given by  $(v_1, v_2)$  having  $v_1 v_2 > 0$  such that  $v_1^2 + v_2^2 = C_\alpha$  for  $C_\alpha := \log(1 - 2\alpha) - \log(2\alpha)$ .
- (3) If  $\alpha < 1/4$ , two stable fixed points at  $(v_1, v_2)$  equals  $(\sqrt{C_\alpha}, -\sqrt{C_\alpha})$  and  $(-\sqrt{C_\alpha}, \sqrt{C_\alpha})$ .

If  $\mu_n$  is e.g., given by  $(v_1, v_2) \sim \mathcal{N}(0, I_2)$  and  $W_1, W_2 \sim \mathcal{N}(0, I_N/(\lambda N))$  then  $\nu := \lim(\mathbf{u}_n)_* \mu_n$  is  $\mathcal{N}(0, I_2)$  in the  $v_1, v_2$  coordinates, and is in the basin of attraction of the quarter-circles of item (2) with probability  $1/2$  and the basin of attraction of the stable fixed points of (3) with probability  $1/2$ .

---

<sup>5</sup>For large  $\lambda$ , this is indeed a quantitative approximation as  $\mathbf{f}, \mathbf{g}$  exhibit locally Lipschitz dependence on  $\lambda^{-1}$ , so the corresponding dynamics converges as  $\lambda \rightarrow \infty$  by classical well-posedness results (see, e.g., [72])

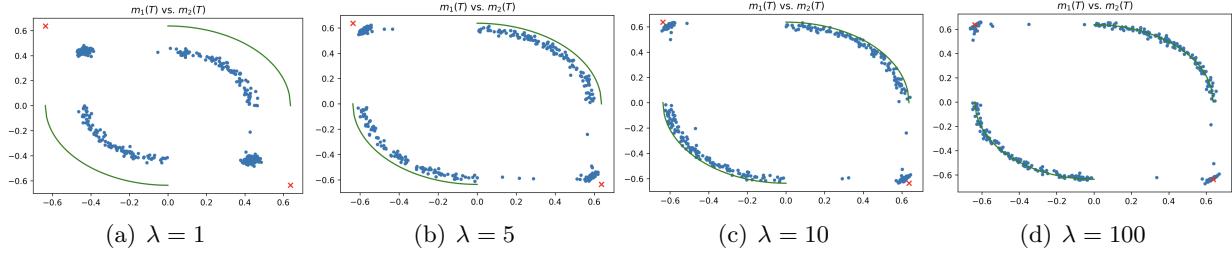


FIGURE 6. GMM in dimension  $N = 500$  with  $\alpha = 0.1$  at various values of  $\lambda$ . Depicted are  $(m_1, m_2)$  values that 500 runs of SGD converge to after  $100N$  steps from a random Gaussian initialization. The — and  $\times$  are the unstable and stable fixed points of the  $\lambda = \infty$  ballistic effective dynamics. The fixed points of the limiting effective dynamics have the same structure at finite  $\lambda$  as  $\lambda = \infty$ , and that as  $\lambda$  gets large quantitatively approach the  $\lambda = \infty$  ones.

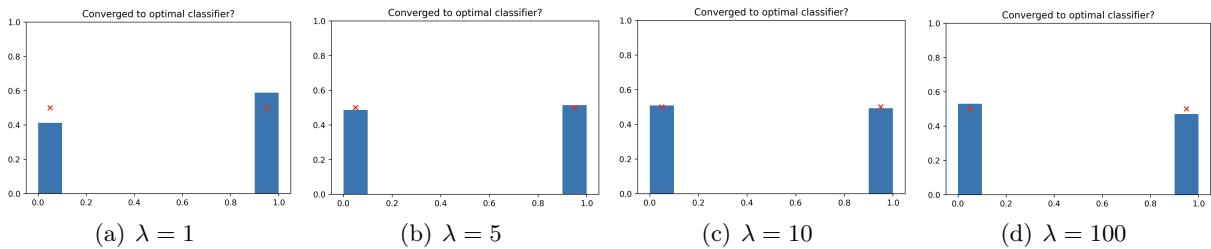


FIGURE 7. GMM in dimension  $N = 500$  with  $\alpha = 0.1$ . Depicted is the fraction of endpoints (SGD after  $100N$  steps from a random Gaussian initialization) with  $m_1 m_2 < 0$ , corresponding to the stable fixed points of the  $\lambda = \infty$  dynamics; it matches the predicted  $\frac{1}{2} - \frac{1}{2}$  fraction.

**4.4. Convergence to spurious solutions.** Let us pause to interpret this result. The stable fixed points when  $\alpha < 1/4$  are the optimal classifiers, whereas the unstable set of fixed points given by item (2) misclassify half of the data. Therefore, the above indicates that when solving the above task with randomly initialized weights, one of the following two scenarios occur, each with probability  $1/2$  (with respect to the initialization): the algorithm will converge to the optimal classifier in linear time or it will appear to have converged to a macroscopically sub-optimal classifier on the same timescale, see Figures 6–7 for numerical verification of this at finite  $N$  and  $\lambda$ .

**4.5. Degeneracy of diffusive limits.** It is then natural to ask about the behaviour of the SGD in the latter regime, after it converges to the sub-optimal classifiers which lie on the aforementioned quarter-circles. Proposition 4.2 rigorously justified the exchange of  $n \rightarrow \infty$  and  $\lambda \rightarrow \infty$  limits in the ballistic phase. In the diffusive phase, one could in principle find the quarter circle of fixed points of the ODE in Proposition 4.1 and consider rescaled observables  $\tilde{v}_i, \tilde{m}_i$  corresponding to blowing up  $v_i, m_i$  in diffusive  $O(n^{-1/2})$  neighborhoods about them to get SDE limits from Theorem 2.3. In order to have explicit formulae, in what follows, we consider the diffusive limits obtained when taking  $\lambda = \infty$ , for which we know the precise locations of these fixed points from Proposition 4.2. This also captures the limit obtained by taking any  $\lambda_n$  diverging faster than  $O(n^{1/2})$ ; the numerics of Figure 8 demonstrate its qualitative consistency with the behavior in microscopic neighborhoods of fixed points at  $\lambda$  finite.

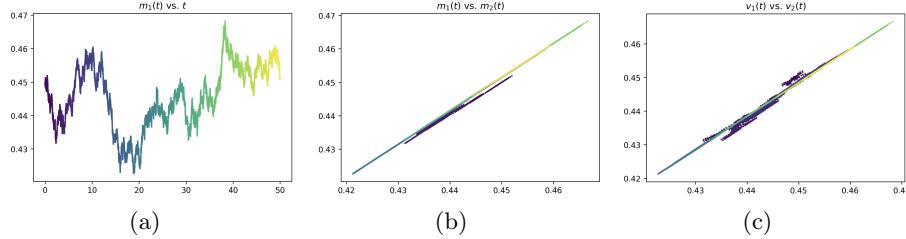


FIGURE 8. Binary GMM in dim.  $N = 250$  with  $\lambda = 100$  and  $\alpha = 0.1$ . Diffusive limits for (a)  $m_1$  individually, and (b)–(c) the pairs  $(m_1, m_2)$  and  $(v_1, v_2)$  where the diffusions can be seen to not be of full rank.

**Proposition 4.3.** Let  $\delta_n = 1/N$ ,  $(a_1, a_2) \in \mathbb{R}_+^2$  be such that  $a_1^2 + a_2^2 = C_\alpha$  and let  $\tilde{v}_i = \sqrt{N}(v_i - a_i)$  and  $\tilde{m}_i = \sqrt{N}(m_i - a_i)$ . When  $\lambda = \infty$ , the SDE system obtained by applying Theorem 2.3 to  $\tilde{\mathbf{u}}_n$  is

$$\begin{aligned} d\tilde{v}_i &= \alpha(\tilde{m}_i - \tilde{v}_i) + a_i(\alpha - 2\alpha^2) \sum a_k (\tilde{v}_k + \tilde{m}_k) + \tilde{\Sigma}^{1/2} d\mathbf{B}_t \cdot e_{v_i}, & dR_{ii}^\perp &= -2\alpha R_{ii}^\perp dt, \\ d\tilde{m}_i &= \alpha(\tilde{v}_i - \tilde{m}_i) + a_i(\alpha - 2\alpha^2) \sum a_k (\tilde{v}_k + \tilde{m}_k) + \tilde{\Sigma}^{1/2} d\mathbf{B}_t \cdot e_{m_i}, & dR_{ij}^\perp &= -2\alpha R_{ij}^\perp dt, \end{aligned}$$

where  $\tilde{\Sigma}$  is a matrix whose only non-zero entries are  $\tilde{\Sigma}_{\tilde{v}_i \tilde{v}_j} = \tilde{\Sigma}_{\tilde{m}_i \tilde{m}_j} = \tilde{\Sigma}_{\tilde{v}_i \tilde{m}_j} = \alpha^2 a_i a_j$ .

Notice that this diffusion matrix is rank 1, so this diffusion is non-trivial but degenerate even in the rescaled coordinates  $(\tilde{v}_i, \tilde{m}_i)$ . Moreover, the entries of  $\tilde{\Sigma}$  vanish on the axes  $a_1 = 0$  or  $a_2 = 0$ . In particular, crossing from the unstable quarter ring into the quadrants  $v_1 v_2 < 0$  where the stable fixed points lie is *impossible* in the noiseless setting, and happens on a much larger timescale at finite  $\lambda$ .

## 5. TWO-LAYER NETWORKS FOR THE XOR GAUSSIAN MIXTURE

**5.1. Model and Background.** For our final example, consider the problem of supervised learning for an XOR-type Gaussian mixture model in  $\mathbb{R}^N$ . Suppose that we are given i.i.d. samples of the form  $Y = (y, X)$ , where  $y$  is  $Ber(1/2)$  and  $X$  has the following distribution: if  $y = 1$  then  $X$  is a  $1/2$ - $1/2$  mixture of  $\mathcal{N}(\mu, I/\lambda)$  and  $\mathcal{N}(-\mu, I/\lambda)$  and if  $y = 0$  it is a  $1/2$ - $1/2$  mixture of  $\mathcal{N}(\nu, I/\lambda)$  and  $\mathcal{N}(-\nu, I/\lambda)$ , where  $\lambda > 0$ , and  $\mu, \nu$  are orthogonal unit vectors. Here,  $y$  is the class label and  $X$  is the data.

This data model is a Gaussian mixture model analogue of the (in)famous XOR problem of Minsky–Papert [51]. In particular, it is easy to see that the optimal decision boundary is not expressible by a single-layer neural network as the data is not linearly separable. That said, it is also straightforward to see that this decision boundary is realizable by simple two-layer networks.<sup>6</sup>

We focus on this example as a demonstration of the applicability of our techniques to the analysis of the training dynamics for two-layer neural networks on natural data models. While this model is arguably the simplest model requiring a multi-layer network to solve, it nevertheless exhibits very complex phenomenology. We mention that some of these complexities were also observed in a very similar setup in [60] where ballistic limits from warm starts were derived.

**5.2. Analysis.** Consider the corresponding classification problem using a two-layer neural network, taking as our estimator of the class label  $\hat{y}(X)$  to be the natural rounding of  $\sigma(v \cdot g(WX))$ , where  $\sigma$  and  $g$  are the sigmoid and ReLU as in Section 4. We take  $W$  to be a  $K \times N$  matrix and  $v$  to be a  $K$ -vector.

<sup>6</sup>In the notation of the following subsection, this can be realized by taking  $K = 4$ ,  $W_1 = -W_2 = \mu$ ,  $W_3 = W_4 = \nu$ , and  $v_i = c$  for  $i = 1, \dots, 4$  for some  $c > 0$ .

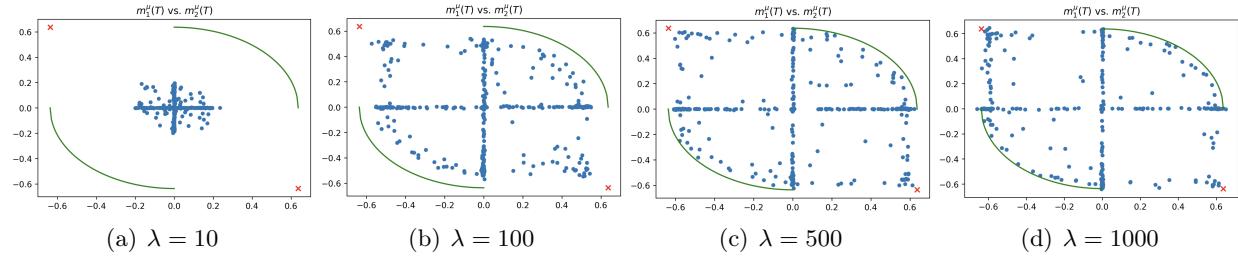


FIGURE 9. XOR in dimension  $N = 500$  with  $\alpha = 0.1$  and  $K = 4$ . Depicted are  $(m_1^\mu, m_2^\mu)$  values that 500 runs of SGD converge to after  $100N$  steps from a random Gaussian initialization. The  $\textcolor{green}{-}$  and  $\textcolor{red}{\times}$  are the unstable and stable fixed points of the  $\lambda = \infty$  ballistic effective dynamics. This demonstrates that the fixed points of the limiting effective dynamics have the same qualitative structure at finite  $\lambda$  as  $\lambda = \infty$ , and approach the  $\lambda = \infty$  ones as  $\lambda$  gets large.

To train the network, we again consider the binary cross-entropy loss with  $\ell_2$ -penalty. This loss is identical to (4.1) *mutatis mutandis*. For the readers convenience, we recall that the loss is of the form

$$L((v_i, W_i)_{i \leq K}; (y, X)) = -yv \cdot g(WX) + \log(1 + e^{v \cdot g(WX)}) + p(v, W),$$

where again  $\sigma, g$  are applied component wise and again  $p(v, W) := (\alpha/2)(\|v\|^2 + \|W\|^2)$ .

In Lemma 9.1 below, we show that the law of the loss at a point  $(v, W)$  depends only on the following  $4K + \binom{K}{2}$  variables: for  $1 \leq i \leq j \leq K$ ,

$$v_i, \quad m_i^\mu = W_i \cdot \mu, \quad m_i^\nu = W_i \cdot \nu, \quad R_{ij}^\perp = W_i^\perp \cdot W_j^\perp \quad (5.1)$$

where  $W_i^\perp = W_i - m_i^\mu \mu - m_i^\nu \nu$  is the part perpendicular to  $\mu, \nu$ . Furthermore, this lemma shows that, if  $\mathbf{u}_n$  given by these variables, then for any fixed  $\lambda > 0$ , the localizability criterion of Definition 2.1 holds as long as  $\delta_n = O(1/n)$ . We can then apply Theorem 2.3 to obtain limits in both the ballistic and diffusive phases. To this end, we need to define the following auxiliary functions analogous to (4.3) above. For a point  $(v, W) \in \mathbb{R}^{K+KN}$ , define the quantity

$$\mathbf{A}_i = \mathbb{E}[X \mathbf{1}_{W_i \cdot X \geq 0}(-y + \sigma(v \cdot g(WX)))],$$

and let

$$\mathbf{A}_i^\mu = \mu \cdot \mathbf{A}_i, \quad \mathbf{A}_i^\nu = \nu \cdot \mathbf{A}_i, \quad \mathbf{A}_{ij}^\perp = W_j^\perp \cdot \mathbf{A}_i.$$

Furthermore, let

$$\mathbf{B}_{ij} = \mathbb{E}[\mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0}(-y + \sigma(v \cdot g(WX)))^2].$$

By similar reasoning, it can be shown that these functions are expressible as functions of  $\mathbf{u}_n$  alone (see Section 9 below). We then find the following effective ballistic dynamics.

**Proposition 5.1.** *Let  $\mathbf{u}_n$  be as in (5.1) and fix any  $\lambda > 0$  and  $\delta_n = c_\delta/N$ . Then  $\mathbf{u}_n(t)$  converges to the solution of the ODE system  $\dot{\mathbf{u}}_t = -\mathbf{f}(\mathbf{u}_t) + \mathbf{g}(\mathbf{u}_t)$ , initialized from  $\lim_n(\mathbf{u}_n)_* \mu_n$  with*

$$\begin{aligned} f_{v_i} &= m_i^\mu \mathbf{A}_i^\mu(\mathbf{u}) + m_i^\nu \mathbf{A}_i^\nu(\mathbf{u}) + \mathbf{A}_{ii}^\perp(\mathbf{u}) + \alpha v_i, & f_{m_i^\mu} &= v_i \mathbf{A}_i^\mu + \alpha m_i^\mu, \\ f_{R_{ij}^\perp} &= v_i \mathbf{A}_{ij}^\perp(\mathbf{u}) + v_j \mathbf{A}_{ji}^\perp(\mathbf{u}) + 2\alpha R_{ij}^\perp, & f_{m_i^\nu} &= v_i \mathbf{A}_i^\nu + \alpha m_i^\nu. \end{aligned}$$

and correctors  $g_{v_i} = g_{m_i^\mu} = g_{m_i^\nu} = 0$ , and  $g_{R_{ij}^\perp} = c_\delta \frac{v_i v_j}{\lambda} \mathbf{B}_{ij}$  for  $1 \leq i \leq j \leq K$ .

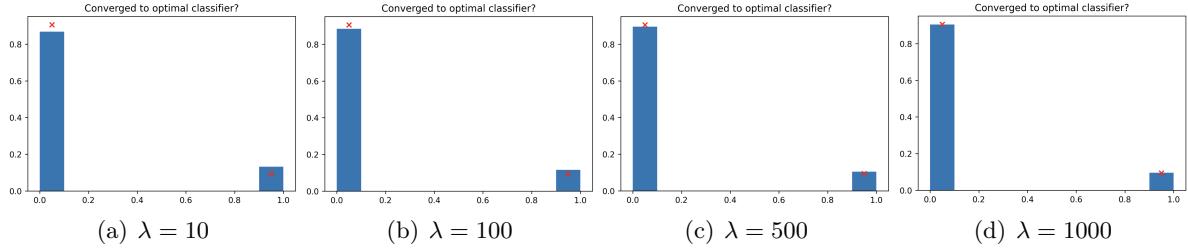


FIGURE 10. XOR in dimension  $N = 500$  with  $\alpha = 0.1$  and  $K = 4$ . The fraction of endpoints (SGD after  $100N$  steps from a random Gaussian initialization) with  $v$  having two positive entries and two negative entries, and with the consequent correct signs on  $m_i^\mu, m_i^\nu$ , corresponding to the stable fixed points of the  $\lambda = \infty$  dynamics; it matches the predicted  $\frac{29}{32}, \frac{3}{32}$  fractions.

**5.3. Low variance asymptotics.** As with the binary GMM, one can develop the large  $\lambda$  limit of these asymptotics after  $n \rightarrow \infty$ . The effective dynamics in this regime are noticeably more tractable. We defer the precise expressions of these dynamics to Proposition 9.1 below. Let us instead classify the corresponding fixed points.

**Proposition 5.2.** *The fixed points of the ODE system of Proposition 9.1 are classified as follows. If  $\alpha > 1/8$ , then the only fixed point is at  $\mathbf{u}_n = \mathbf{0}$ .*

*If  $0 < \alpha < 1/8$ , then let  $(I_0, I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-)$  be any disjoint (possibly empty) subsets whose union is  $\{1, \dots, K\}$ . Corresponding to that tuple  $(I_0, I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-)$ , is a set of fixed points that have  $R_{ij}^\perp = 0$  for all  $i, j$ , and have*

- (1)  $m_i^\mu = m_i^\nu = v_i = 0$  for  $i \in I_0$ ,
- (2)  $m_i^\mu = v_i > 0$  such that  $\sum_{i \in I_\mu^+} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\nu = 0$  for all  $i \in I_\mu^+$ ,
- (3)  $-m_i^\mu = v_i > 0$  such that  $\sum_{i \in I_\mu^-} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\nu = 0$  for all  $i \in I_\mu^-$ ,
- (4)  $m_i^\nu = v_i < 0$  such that  $\sum_{i \in I_\nu^+} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\mu = 0$  for all  $i \in I_\nu^+$ ,
- (5)  $-m_i^\nu = v_i < 0$  such that  $\sum_{i \in I_\nu^-} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\mu = 0$  for all  $i \in I_\nu^-$ .

*In the  $K = 4$  case, these form 39 connected sets of fixed points, and of which  $4! = 24$  are fixed points that are stable, corresponding to the possible permutations in which each of  $I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-$  are singletons.*

Similar to the binary GMM, in Figures 9–10, we demonstrate numerically that the following predicted fixed points from the  $\lambda \rightarrow \infty$  limit match those arising at finite large  $n$  and  $\lambda > 0$ .

In the  $K = 4$  case, we can also exactly calculate the probability that the effective dynamics in the ballistic phase converges to a stable fixed point (as opposed to an unstable one). From a Gaussian initialization  $\mu_n$  where  $v_i \sim \mathcal{N}(0, 1)$  and  $W_i \sim \mathcal{N}(0, I_N/N)$  independently, this converges to  $3/32$ . We refer the reader to Section 9.4 for the proof.

**5.4. Overparametrization in the XOR GMM.** Since the the derivations of the ballistic limiting equations apply for general  $K$ , we can also study the probability of ballistic convergence to a stable vs. unstable fixed point as one varies  $K$ . This addresses the regime of overparametrization for the XOR GMM since  $K = 4$  suffices to express a Bayes-optimal classifier. In this more generic setting, the probability of being in the ballistic domain of attraction of the stable fixed points (corresponding

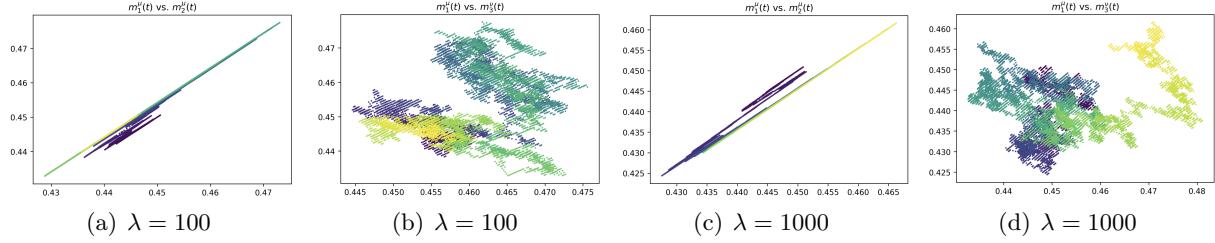


FIGURE 11. XOR GMM in dim.  $N = 250$  and  $\alpha = 0.1$  and  $K = 4$ . (a) and (c) display the degenerate diffusive limits in the regime of Proposition 5.3 in  $(m_1^\mu, m_2^\mu)$  coordinates at  $\lambda = 100$  and  $\lambda = 1000$ . Conversely, (b) and (d) display the diffusive limits in the regime of Proposition 5.3 in  $(m_1^\mu, m_3^\nu)$ , where the limiting diffusions are independent and are of rank 2.

to the Bayes optimal classifiers) is

$$\frac{1}{2^K} \sum_{k=2}^{K-2} \binom{K}{k} (1 - 2^{1-k})(1 - 2^{1+k-K}), \quad (5.2)$$

which goes to 1 exponentially fast as  $K$  grows. This clearly demonstrates the benefits of overparametrization of the landscape in a concrete two-layer network: a random initialization is more likely to contain the “right” initial signature (corresponding to none of  $I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-$  being empty at initialization) in order to be in the basin of a Bayes optimal classifier as the width grows, and as long as the right signature is present in the nodes at initialization, the SGD will ballistically converge to a global minimizer of the population loss. This is a rigorous example of the well-known lottery ticket hypothesis of [30]. Roughly speaking, the lottery ticket hypothesis proposes that the reason for the success of overparametrized networks is that they give more attempts for a sufficiently expressive subnetwork to be initialized well, and succeed at the task on its own.

**5.5. Diffusive limits at unstable fixed points.** As an example of the diffusions that can arise in the rescaled effective dynamics at the unstable fixed points, let us consider the unstable fixed points in which  $v$  has the correct signature (two positive, two negative) but for each of those we are at a corresponding quarter-ring. By way of example, we can set  $K = 4$ , or equivalently focus on a fixed point where all indices beyond the first four have  $v_i = m_i^\mu = m_i^\nu = 0$ . Here, the dynamics effectively becomes a pair of 2 two-layer GMM’s on quarter-rings (as in Section 4), that are anti-correlated. More precisely, let  $(a_{1,\mu}, a_{2,\mu})$  be such that  $a_{1,\mu}^2 + a_{2,\mu}^2 = C_\alpha$  and  $(a_{3,\nu}, a_{4,\nu})$  such that  $a_{3,\nu}^2 + a_{4,\nu}^2 = C_\alpha$ , for  $C_\alpha = -\text{logit}(4\alpha)$ . Take as fixed points about which we expand to be  $v_i = m_i^\mu = a_{i,\mu} > 0$  and  $v_i = m_i^\nu = a_{i,\nu} < 0$  for  $i = 3, 4$ . Namely, we let

$$\tilde{v}_i = \begin{cases} \sqrt{N}(v_i - a_{i,\mu}) & i = 1, 2 \\ \sqrt{N}(v_i - a_{i,\nu}) & i = 3, 4 \end{cases}, \quad \begin{cases} \tilde{m}_i^\mu = \sqrt{N}(m_i^\mu - a_{i,\mu}) & i = 1, 2 \\ \tilde{m}_i^\nu = \sqrt{N}(m_i^\nu - a_{i,\nu}) & i = 3, 4 \end{cases}.$$

(Set  $\tilde{m}_i^\nu = 0$  for  $i = 1, 2$  and  $\tilde{m}_i^\mu = 0$  for  $i = 3, 4$  in  $\tilde{\mathbf{u}}_n$  effectively removing those variables.)

**Proposition 5.3.** Let  $\delta_n = 1/N$  and let  $\tilde{\mathbf{u}}_n = (\tilde{v}_i, \tilde{m}_i^\mu, \tilde{m}_i^\nu, R_{ij}^\perp)$ . When  $\lambda = \infty$ , Theorem 2.3 can be applied and  $\tilde{\mathbf{u}}_n(t)$  converges to the solution of the SDE  $d\tilde{\mathbf{u}}(t) = -\tilde{\mathbf{h}}(\tilde{\mathbf{u}})dt + \sqrt{\Sigma(\tilde{\mathbf{u}})}d\mathbf{B}_t$  where

$$\tilde{h}_{\tilde{v}_i} = \begin{cases} \alpha(\tilde{v}_i - \tilde{m}_i^\mu) - a_{i,\mu}(\alpha - 4\alpha^2) \sum_{k=1,2} a_{k,\mu}(\tilde{v}_k + \tilde{m}_k^\mu) & i = 1, 2 \\ \alpha(\tilde{v}_i - \tilde{m}_i^\nu) - a_{i,\nu}(\alpha - 4\alpha^2) \sum_{k=3,4} a_{k,\nu}(\tilde{v}_k + \tilde{m}_k^\nu) & i = 3, 4 \end{cases}$$

$\tilde{h}_{\tilde{m}_i^\mu}$  (resp.,  $\tilde{h}_{\tilde{m}_i^\nu}$ ) is like  $h_{\tilde{v}_i}$  for  $i = 1, 2$  (resp.,  $i = 3, 4$ ) with  $\tilde{v}_i$  and  $\tilde{m}_i^\mu$  (resp.,  $\tilde{m}_i^\nu$ ) swapped,  $\tilde{h}_{R_{ij}^\perp} = 2\alpha R_{ij}^\perp$ , and  $\tilde{\Sigma}$  is the constant rank-2 matrix whose non-zero entries are

$$\begin{aligned}\tilde{\Sigma}_{\tilde{v}_i \tilde{v}_j} &= \tilde{\Sigma}_{\tilde{m}_i^\mu \tilde{m}_j^\mu} = \tilde{\Sigma}_{\tilde{v}_i \tilde{m}_j^\mu} = 3\alpha^2 a_{i,\mu} a_{j,\mu} \quad \text{if } i, j \in \{1, 2\}, \\ \tilde{\Sigma}_{\tilde{v}_i \tilde{v}_j} &= \tilde{\Sigma}_{\tilde{m}_i^\nu \tilde{m}_j^\nu} = \tilde{\Sigma}_{\tilde{v}_i \tilde{m}_j^\nu} = 3\alpha^2 a_{i,\nu} a_{j,\nu} \quad \text{if } i, j \in \{3, 4\}, \\ \tilde{\Sigma}_{\tilde{v}_i \tilde{v}_j} &= \tilde{\Sigma}_{\tilde{m}_i^\mu \tilde{m}_j^\nu} = \tilde{\Sigma}_{\tilde{m}_i^\nu \tilde{m}_j^\mu} = -\alpha^2 a_{i,\mu} a_{j,\nu} \quad \text{if } i \in \{1, 2\}, j \in \{3, 4\}.\end{aligned}$$

Numerical simulations in Figure 11 confirm these degenerate diffusive limits at finite  $\lambda$ .

### Part 3. Proofs

#### 6. PROOF OF THEOREM 2.3

In this section, we prove our main convergence result, namely Theorem 2.3. The drift terms can be seen from a Taylor expansion out to second order, with the role played by  $\delta_n$ -localizability being to justify neglecting certain negligible second order terms, as well as all higher order terms. The identification of the stochastic term is via the classical martingale problem [70] for summary statistics of stochastic gradient descent in the high-dimensional  $n \rightarrow \infty$  limit.

**Notational remark.** For ease of notation, in the following we say that  $f \lesssim g$  if there is some constant  $C > 0$  such that  $f \leq Cg$  and that  $f \lesssim_a g$  if there is some constant  $C(a) > 0$  depending only on  $a$  such that  $f \leq C(a)g$ . We will often suppress the dependence on  $n$  in subscripts, when it is clear from context.

**Proof of Theorem 2.3.** Our aim is to establish  $\mathbf{u}_n \rightarrow \mathbf{u}$  weakly as random variables on  $C([0, \infty))$  where  $\mathbf{u}$  solves (2.4). It is equivalent to show the same on  $C([0, T])$  equipped with the sup-norm for every  $T > 0$ .

Let  $\tau_K^n$  denote the exit time for the interpolated process  $\mathbf{u}_n(t)$  from  $E_K^n$ . Define its pre-image  $E_{K,n}^* := \mathbf{u}_n^{-1}(E_K^n)$  and let  $L_{K,n}^\infty = L^\infty(E_{K,n}^*)$ . For a function  $f$ , we use the shorthand  $f_\ell$  to denote  $f(X_\ell)$ . By Taylor's theorem, we have that for any  $C^3$  function  $f$  and any  $\ell \leq \tau_K^n/\delta$ ,

$$\begin{aligned}f_\ell &= f(X_{\ell-1} - \delta \nabla \Phi_{\ell-1} - \delta \nabla H_{\ell-1}^\ell) \\ &= f_{\ell-1} + \delta[A_\ell^f - A_{\ell-1}^f] + \delta[M_\ell^f - M_{\ell-1}^f] + O(\delta^3 \|\nabla^3 f\|_{L_{K,n}^\infty} \cdot \|\nabla L\|_{L_{K,n}^\infty}^3),\end{aligned}\tag{6.1}$$

where  $A_\ell^f$  and  $M_\ell^f$  are defined by their increments as follows:

$$\begin{aligned}A_\ell^f - A_{\ell-1}^f &= (-\mathcal{A}_n + \delta \mathcal{L}_n) f_{\ell-1} + \frac{1}{2} \langle \nabla \Phi \otimes \nabla \Phi, \nabla^2 f \rangle_{\ell-1}, \\ M_\ell^f - M_{\ell-1}^f &= -\langle \nabla H^\ell, \nabla f \rangle_{\ell-1} + \delta(\mathcal{E}_\ell^f - \mathcal{E}_{\ell-1}^f), \\ \mathcal{E}_\ell^f - \mathcal{E}_{\ell-1}^f &= \nabla^2 f(\nabla \Phi, \nabla H^\ell)_{\ell-1} + \frac{1}{2} \langle \nabla^2 f, \nabla H^\ell \otimes \nabla H^\ell - V \rangle_{\ell-1},\end{aligned}$$

for  $\mathcal{A}_n = \langle \nabla \Phi, \nabla \rangle$ ,  $\mathcal{L}_n = \frac{1}{2} \sum_{i,j} V_{ij} \partial_i \partial_j$  and  $V = \mathbb{E}[\nabla H \otimes \nabla H]$  as in (2.1). Observe that  $A_\ell^f$  is previsible (with respect to the filtration generated by  $(Y_1, \dots, Y_{\ell-1})$ ), and  $M_\ell^f$  is a martingale. We bound these for  $f = u_j$  among  $\mathbf{u}_n = (u_1, \dots, u_k)$ .

Recalling Definition 2.1, since  $\mathbf{u}_n$  are  $\delta_n$ -localizable, the error term in (6.1) has

$$\delta^3 \sup_{x \in E_{K,n}^*} \mathbb{E}[\|\nabla^3 u_j\| \cdot \|\nabla L\|^3] \lesssim \delta^3 \| \nabla^3 u_j \|_{L_{K,n}^\infty} \left( \|\nabla \Phi\|_{L_{K,n}^\infty}^3 + \sup_{E_{K,n}^*} \mathbb{E}[\|\nabla H\|^3] \right) \lesssim_K \delta^{3/2}.$$

Since  $\delta_n$  goes to 0 as  $n \rightarrow \infty$ , we may thus write  $u_j(X_\ell)$  as

$$u_j(X_\ell) = u_j(0) + \delta \sum_{\ell' \leq \ell} (A_{\ell'}^{u_j} - A_{\ell'-1}^{u_j}) + \delta \sum_{\ell' \leq \ell} (M_{\ell'}^{u_j} - M_{\ell'-1}^{u_j}) + o(1),$$

where the last term is  $o(1)$  in  $L^1$  uniformly for  $\ell \leq \tau_K/\delta$ . Now let us define for  $s \in [0, T]$ ,

$$a'_j(s) = A_{[s/\delta]}^{u_j} - A_{[s/\delta]-1}^{u_j} \quad \text{and} \quad b'_j(s) = M_{[s/\delta]}^{u_j} - M_{[s/\delta]-1}^{u_j}$$

If we let

$$a_j(s) = \int_0^s a'_j(s') ds' = a_j(\delta[s/\delta]) + (s - \delta[s/\delta])(A_{[s/\delta]}^{u_j} - A_{[s/\delta]-1}^{u_j})$$

and similarly  $b_j(s) = \int_0^s b'_j(s') ds'$ , then recalling that  $\mathbf{u}_n(s)$  is the linear interpolation of  $(u_j([s/\delta]))_j$ , we may write

$$\mathbf{u}_n(s) = \mathbf{u}_n(0) + \mathbf{a}_n(s) + \mathbf{b}_n(s) + o(1).$$

where  $\mathbf{a}_n(s) = (a_j(s))_j$  and  $\mathbf{b}_n(s) = (b_j(s))_j$ .

We now prove that the sequence  $(\mathbf{u}_n(s \wedge \tau_K^n))$  is tight in  $C([0, T])$  with limit points which are  $(1/4)$ -Holder for each  $K$ . To this end, let us define

$$\mathbf{v}_n(s) = \mathbf{u}_n(0) + \mathbf{a}_n(s) + \mathbf{b}_n(s).$$

As the  $o(1)$  error above is uniform in  $t$ , we have that

$$\sup_{0 \leq s \leq \tau_K^n} \|\mathbf{u}_n(s) - \mathbf{v}_n(s)\| \rightarrow 0, \quad \text{in } L^1.$$

Thus it suffices to show the claimed tightness and Holder properties of limit points for  $\mathbf{v}_n$  instead of  $\mathbf{u}_n$ . We aim to show that for all  $0 \leq s, t \leq T$ ,

$$\mathbb{E}\|\mathbf{v}_n(s \wedge \tau_K) - \mathbf{v}_n(t \wedge \tau_K)\|^4 \lesssim_{K,T} (t-s)^2, \quad (6.2)$$

from which we will get that the sequence  $\mathbf{v}_n(s \wedge \tau_K)$  is uniformly  $1/4$ -Holder by Kolmogorov's continuity theorem. Evidently, for all  $s, t$  we have that

$$\|\mathbf{v}_n(s) - \mathbf{v}_n(t)\| \leq \|\mathbf{a}_n(s) - \mathbf{a}_n(t)\| + \|\mathbf{b}_n(s) - \mathbf{b}_n(t)\|.$$

We control these terms in turn. We will do this coordinate wise and, for readability, fix some  $j \leq k$  and let  $u = u_j$ ,  $a = a_j$ ,  $b = b_j$  etc.

For the previsible term, we have

$$\begin{aligned} & \mathbb{E}|a(s \wedge \tau_K) - a(t \wedge \tau_K)|^4 \\ & \lesssim \mathbb{E}|\delta \sum_{\ell} ((-\mathcal{A}_n + \delta \mathcal{L}_n)u)_{\ell}|^4 + \mathbb{E}|\delta^2 \sum_{\ell} \langle \nabla \Phi \otimes \nabla \Phi, \nabla^2 u \rangle_{\ell}|^4, \end{aligned} \quad (6.3)$$

where these sums are over steps  $\ell$  ranging from  $[s/\delta] \wedge \tau_K/\delta$  to  $[t/\delta] \wedge \tau_K/\delta$ .

Let  $\mathbf{h} = (h_j)_{j \leq k}$  be as in (2.2). Then the first term in (6.3) satisfies

$$\begin{aligned} & \mathbb{E}|\delta \sum_{\ell} ((-\mathcal{A}_n + \delta \mathcal{L}_n)u)_{\ell}|^4 \lesssim_K \mathbb{E}|\delta \sum_{\ell} h_j(\mathbf{u}_n)_{\ell}|^4 + o((t-s)^4) \\ & \leq (t-s)^4 \left( \|h_j\|_{L^\infty(E_K^n)}^4 + o(1) \right) \lesssim_K (t-s)^4 \end{aligned}$$

by continuity of  $h_j$ . For the second term in (6.3),

$$\mathbb{E}|\delta^2 \sum_{\ell} \langle \nabla \Phi \otimes \nabla \Phi, \nabla^2 u \rangle_{\ell}|^4 \leq \delta^8 \left( |((t-s)/\delta)| \sup_{x \in E_{K,n}^*} \|\nabla \Phi(x)\|^2 \sup_{x \in E_{K,n}^*} \|\nabla^2 u(x)\|_{\text{op}} \right)^4$$

which is  $\lesssim_K \delta^2(t-s)^4$  by items (1)–(2)  $\delta_n$ -localizability. (Applying this bound for  $s=0, t=T$ , the last term in  $a$  is vanishing in the limit for each  $K$  whenever  $\delta_n = o(1)$ .) Combining the above bounds yields

$$\mathbb{E}|a(s \wedge \tau_K) - a(t \wedge \tau_K)|^4 \lesssim_K (t-s)^4.$$

For the martingale term, notice that by Burkholder's inequality,

$$\mathbb{E}|b(s \wedge \tau_K) - b(t \wedge \tau_K)|^4 = \mathbb{E}\left[\left(\delta \sum(M_\ell^u - M_{\ell-1}^u)\right)^4\right] \lesssim \mathbb{E}\left[\left(\delta^2 \sum(M_\ell^u - M_{\ell-1}^u)^2\right)^2\right],$$

where the sum again runs over steps  $\ell$  ranging from  $[s/\delta] \wedge \tau_K$  to  $[t/\delta] \wedge \tau_K$ . Repeatedly using the inequality  $(x+y+z)^2 \lesssim x^2 + y^2 + z^2$ , it suffices to bound the above quantity for each of the three terms defining the martingale difference  $M_\ell^u - M_{\ell-1}^u$  respectively.

For the first term in that martingale difference, observe that

$$\begin{aligned} \mathbb{E}\left[\left(\delta^2 \sum_\ell \langle \nabla H^\ell, \nabla u \rangle_{\ell-1}^2\right)^2\right] &= \delta^4 \sum_{\ell, \ell'} \mathbb{E}\left[\langle \nabla H^\ell, \nabla u \rangle_{\ell-1}^2 \langle \nabla H^{\ell'}, \nabla u \rangle_{\ell'-1}^2\right] \\ &\leq \left(\delta \sum_\ell \left(\delta^2 \mathbb{E}\langle \nabla H^\ell, \nabla u \rangle_{\ell-1}^4\right)^{1/2}\right)^2 \lesssim_K (t-s)^2, \end{aligned} \quad (6.4)$$

where in the second line we used Cauchy-Schwarz and in the last we used item (3) of  $\delta_n$ -localizability.

For the second term in the martingale difference,

$$\begin{aligned} \mathbb{E}\left[\left(\delta^4 \sum_\ell (\nabla^2 u(\nabla \Phi, \nabla H^\ell)_{\ell-1})^2\right)^2\right] &\leq \delta^6(t-s)^2 \sup_{x \in E_{K,n}^*} \|\nabla^2 u(x)\|_{\text{op}}^4 \cdot \|\nabla \Phi(x)\|^4 \cdot \mathbb{E}\|\nabla H(x)\|^4 \lesssim_K \delta^2(t-s)^2, \end{aligned} \quad (6.5)$$

by items (1)–(2) of  $\delta_n$ -localizability. Finally, by the same reasoning, for the third term,

$$\begin{aligned} \mathbb{E}\left[\left(\delta^4 \sum_\ell \langle \nabla^2 u, \nabla H^\ell \otimes \nabla H^\ell - V \rangle_{\ell-1}^2\right)^2\right] &\lesssim \delta^6(t-s)^2 \sup_{x \in E_{K,n}^*} \|\nabla^2 u(x)\|_{\text{op}}^4 \cdot \mathbb{E}\|\nabla H(x)\|^8 \lesssim_K (t-s)^2. \end{aligned} \quad (6.6)$$

All of the above terms are  $O((t-s)^2)$  since  $0 \leq s, t \leq T$ . Thus we have the claimed (6.2), and by Kolmogorov's continuity theorem,  $(\mathbf{v}_n(s \wedge \tau_K))_s$ , are uniformly  $1/4$ -Holder and thus the sequence is tight with  $1/4$ -Holder limit points. Notice furthermore that if we look at  $(\mathbf{v}_n(t \wedge \tau_K) - \mathbf{a}_n(t \wedge \tau_K))_t$ , this sequence is also tight and the limits points are continuous martingales. Let us examine their limiting quadratic variations.

Let  $\mathbf{v}_n^K(t) = \mathbf{v}_n(t \wedge \tau_K)$  and define  $\mathbf{a}_n^K(t)$  and  $\mathbf{b}_n^K(t)$  analogously. Furthermore, let  $\mathbf{v}^K(t)$ ,  $\mathbf{a}^K(t)$  and  $\mathbf{b}^K(t)$  be their respective limits which we have shown to exist and be  $1/4$ -Holder.

We will compute the limiting quadratic variation for  $\mathbf{b}^K(t)$ . For ease of notation, let  $\Delta M_\ell^{u_i} = M_\ell^{u_i} - M_{\ell-1}^{u_i}$  and  $\Delta \mathcal{E}_\ell^{u_i} = \mathcal{E}_\ell^{u_i} - \mathcal{E}_{\ell-1}^{u_i}$ . Notice first that for  $1 \leq i, j \leq k$ ,

$$b_{n,i}^K(t)b_{n,j}^K(t) - \int_0^t \delta \mathbb{E}[\Delta M_{[s/\delta] \wedge \tau_K}^{u_i} \Delta M_{[s/\delta] \wedge \tau_K}^{u_j}] ds,$$

is a martingale. We therefore need to consider the limit as  $n \rightarrow \infty$  of the integral above. Write

$$\begin{aligned} \mathbb{E}[\Delta M_\ell^{u_i} \Delta M_\ell^{u_j}] &= \langle \nabla u_i, V \nabla u_j \rangle + \delta \mathbb{E}[\langle \nabla H^\ell, \nabla u_i \rangle_{\ell-1} \Delta \mathcal{E}_\ell^{u_j}] + \delta \mathbb{E}[\Delta \mathcal{E}_\ell^{u_i} \langle \nabla H^\ell, \nabla u_j \rangle_{\ell-1}] \\ &\quad + \delta^2 \mathbb{E}[\Delta \mathcal{E}_\ell^{u_i} \Delta \mathcal{E}_\ell^{u_j}]. \end{aligned} \quad (6.7)$$

Consider the integrals of  $\delta$  times each of these four terms separately. For the first term,

$$\sup_{t \leq T} \left| \int_0^t \delta \langle \nabla u_i, V \nabla u_j \rangle_{[s/\delta] \wedge \tau_K} - \Sigma_{ij}(\mathbf{v}_n^K(s)) ds \right| \leq T \sup_{x \in E_{K,n}^*} |\delta \langle \nabla u_i, V \nabla u_j \rangle(x) - \Sigma_{ij}(\mathbf{u}_n(x))|,$$

goes to zero as  $n \rightarrow \infty$  by the assumption in (2.3).

We now reason that the integrals of  $\delta$  times the other three terms in (6.7) all go to zero as  $n \rightarrow \infty$ . The second and third are identical: by Cauchy–Schwarz,

$$\sup_{x \in E_{K,n}^*} |\delta^2 \mathbb{E}[\langle \nabla H, \nabla u_i \rangle \Delta \mathcal{E}_\ell^{u_j}]| \leq \delta^2 \mathbb{E}[\langle \nabla H, \nabla u_i \rangle^2]^{1/2} \mathbb{E}[(\Delta \mathcal{E}_\ell^{u_i})^2]^{1/2}.$$

The first expectation contributes  $\delta^{-1/2}$  by the first part of item (3) of localizability. Also,

$$\mathbb{E}[(\Delta \mathcal{E}_\ell^{u_i})^2]^{1/2} \lesssim \mathbb{E}[\langle \nabla^2 u_i, \nabla \Phi \otimes \nabla H \rangle^2]^{1/2} + \mathbb{E}[\langle \nabla^2 u_i, \nabla H \otimes \nabla H - V \rangle^2]^{1/2}. \quad (6.8)$$

The first of these terms is at most  $\delta^{-1}$  as argued in (6.5). The second is  $o(\delta^{-3/2})$  by the second part of item (3) in the definition of localizability. As such, we are able to conclude that

$$\sup_{t \leq T} \left| \int_0^t \delta^2 \mathbb{E}[\langle \nabla H, \nabla u_i \rangle_{[s/\delta] \wedge \tau_K} \Delta \mathcal{E}_{[s/\delta] \wedge \tau_K}^{u_j}] ds \right|,$$

goes to zero as  $n \rightarrow \infty$ .

The integral of  $\delta$  times the fourth term in (6.7) is handled similarly using Cauchy–Schwarz and the bound of  $o(\delta^{-3/2})$  on (6.8).

Altogether, we end up with

$$\lim_{n \rightarrow \infty} \sup_{i,j \leq k} \sup_{t \leq T} \left| \int_0^t \delta \mathbb{E}[\Delta M_{[s/\delta] \wedge \tau_K}^{u_i} \Delta M_{[s/\delta] \wedge \tau_K}^{u_j}] ds - \int_0^t \Sigma_{ij}(\mathbf{v}_n^K(s)) ds \right| = 0.$$

Thus, if we consider the continuous martingales given by  $\mathbf{b}^K(t)$ , its angle bracket is, by definition, given by

$$\langle \mathbf{b}^K \rangle_t = \int_0^t \Sigma(\mathbf{v}^K(s)) ds.$$

By Ito's formula for continuous martingales (see, e.g., [28, Theorem 5.2.9]), we have that  $f(\mathbf{v}_t) - \int_0^t \mathsf{L}f(\mathbf{v}_s) ds$  is a martingale for all  $f \in C_0^\infty(\mathbb{R}^k)$ , where

$$\mathsf{L} = \frac{1}{2} \sum_{ij=1}^k \Sigma_{ij} \partial_i \partial_j - \sum_{i=1}^k h_i \partial_i.$$

Since, by assumption,  $\mathbf{h}, \sqrt{\Sigma}$  are locally Lipschitz—and thus Lipschitz on  $E_K$ —this property uniquely characterizes the solutions to (2.4) (see, e.g., [70, Theorem 6.3.4]). Thus  $\mathbf{v}_K$  converges to the solution of (2.4) stopped at  $\tau_K$ . By a standard localization argument [70, Lemmas 11.1.11–12], every limit point  $\mathbf{v}(t)$  of  $\mathbf{v}_n(t)$  solves the SDE (2.4) (using here that  $E_K$  is an exhaustion by compact sets of  $\mathbb{R}^k$ ).  $\square$

## 7. PROOFS FOR MATRIX AND TENSOR PCA

In this section, we prove the results of Section 3. We will state them in the more general setting where we add a ridge penalty to the loss, so that for  $\alpha \geq 0$  fixed, the loss is given by

$$L(x, Y) = -2(\langle W, x^{\otimes k} \rangle + \lambda \langle x, v \rangle^k) + \|x\|^{2k} + \frac{\alpha}{2} \|x\|^2 + c(Y), \quad (7.1)$$

where  $c(Y)$  only depends on  $Y$ . Note that  $H(x) = -2\langle W, x^{\otimes k} \rangle$ .

Our first aim is to establish Proposition 3.1, showing that the summary statistics  $\mathbf{u}_n = (m, r_\perp^2)$  satisfy the conditions of Theorem 2.3 with the desired  $\mathbf{f}, \mathbf{g}$  and  $\Sigma$ . We begin by checking localizability

for  $\mathbf{u}_n$ . In what follows, for ease of notation we will denote  $r^2 = r_\perp^2$  and  $R^2 = m^2 + r^2$ . In these coordinates,

$$\Phi(x) = -2\lambda m^k + (r^2 + m^2)^k + \frac{\alpha}{2}(r^2 + m^2) + c' \quad (7.2)$$

**Lemma 7.1.** *The distribution of  $L(x, Y)$  depends only on  $\mathbf{u}_n = (m, r^2)$ . Furthermore, if  $\lambda$  is fixed and  $\delta_n = O(1/n)$ , then  $\mathbf{u}_n$  is  $\delta_n$ -localizable for  $E_K$  being the centered balls of radius  $K$  in  $\mathbb{R}^2$ .*

*Proof.* We check the items in Definition 2.1 one by one, beginning with item (1). Express the derivatives for  $\mathbf{u}_n$  as

$$\nabla m = v, \quad \nabla r^2 = 2(x - mv). \quad (7.3)$$

Notice that  $\nabla^2 m = 0$ , while  $\nabla^2 r^2 = 2(I - vv^T)$ , whose operator norm is simply 2, and  $\nabla^\ell u_i = 0$  for all  $\ell \geq 3$ .

For item (2), differentiating (7.2),  $\nabla \Phi = \partial_1 \phi \nabla m + \partial_2 \phi \nabla r^2$ , where

$$\partial_1 \phi = -2\lambda km^{k-1} + (2kR^{2k-2} + \alpha)m \quad \partial_2 \phi = kR^{2k-2} + \frac{\alpha}{2}.$$

Notice that  $\langle \nabla m, \nabla m \rangle = 1$ ,  $\langle \nabla m, \nabla r^2 \rangle = 0$ , and  $\langle \nabla r^2, \nabla r^2 \rangle = 4r^2$ . Consider

$$\|\nabla \Phi\| \leq |\partial_1 \phi| \|\nabla m\| + |\partial_2 \phi| \|\nabla r^2\|;$$

the bounding quantity is evidently a continuous function of  $m, r^2$  and therefore as long as  $x$  is such that  $(m, r^2) \in E_K$ , it is bounded by some  $C(K)$ . Next, if we consider

$$\mathbb{E}[\|\nabla H\|^8] \leq C_k \mathbb{E}[\|W(x, \dots, x, \cdot)\|^8] \leq \mathbb{E}\|W\|_{\text{op}}^8 \cdot R^{8k} \leq C(k, K)n^4$$

where the bound on the operator norm of an i.i.d. Gaussian  $k$ -tensor can be found, e.g., in [9, Lemma 4.7]. Moving on to item (3), by the same reasoning, for every  $w$ ,

$$\mathbb{E}[\langle \nabla H, w \rangle^4] \leq 16k \mathbb{E}[|W(w, x, \dots, x)|^4] \leq C(k, K)n^2 \|w\|.$$

If  $w = \nabla m = v$  then  $\|w\| = 1$  and if  $w = \nabla r^2 = 2(x - mv)$  then  $\|w\| \leq C(K)$ , so in both cases this is at most  $C(k, K)n^2$ . Finally,  $\nabla^2 u$  is only non-zero if  $u = r$  in which case it is  $I - vv^T$ . Then,

$$\mathbb{E}[\langle \nabla^2 r, \nabla H \otimes \nabla H - V \rangle^2] \leq 2\mathbb{E}[\|\nabla H\|^4] \leq C(k, K)n^2$$

by the second item in the definition of localizability, and evidently the right-hand side is  $O(\delta^{-2})$  if  $\delta_n = O(1/n)$ .  $\square$

**Proof of Proposition 3.1.** Having checked localizability for  $\mathbf{u}_n$ , we apply Theorem 2.3. To compute  $\mathbf{f}$ , by the above,

$$f_m = -2\lambda km^{k-1} + (2kR^{2k-2} + \alpha)m, \quad f_{r^2} = 2r^2(2kR^{2k-2} + \alpha).$$

We next turn to calculating the corrector. For this, we first calculate the matrix  $V = \mathbb{E}[\nabla H \otimes \nabla H]$ . Recalling that  $H = -2\langle W, x^{\otimes k} \rangle$  where  $W$  is an i.i.d. Gaussian  $k$ -tensor, we have that

$$V_{ij} = \mathbb{E}[\partial_i H \partial_j H] = 4k(k-1)x_i x_j R^{2k-4} + 4kR^{2k-2} \mathbf{1}\{i=j\}. \quad (7.4)$$

In particular, for  $\delta = c_\delta/n$ , we have  $\delta \mathcal{L}^\delta m = 0$  and

$$\delta \mathcal{L}^\delta r^2 = \frac{4c_\delta}{n} k \left( (n-1)R^{2k-2} + (k-1)r^2 R^{2k-4} \right)$$

from which we obtain in the limit that  $n \rightarrow \infty$  that  $g_m = 0$  and  $g_{r^2} = 4c_\delta k R^{2k-2}$ .

Together, these yield the ODE system of (3.1),

$$\dot{u}_1 = 2u_1(\lambda k u_1^{k-2} - kR^{2k-2} - \alpha), \quad \dot{u}_2 = -(4u_2 - 4c_\delta)kR^{2k-2} - 2\alpha u_2.$$

which in the  $\alpha = 0$  case matches Proposition 3.1. Finally, to see that  $\Sigma = 0$ , consider

$$JVJ^T = \begin{pmatrix} 4k(k-1)m^2R^{2k-4} + 4kR^{2k-2} & 4k(k-1)m(R^2-m)R^{2k-4} \\ 4k(k-1)m(R^2-m)R^{2k-4} & 4k(k-1)(R^2-m)^2R^{2k-4} \end{pmatrix}, \quad (7.5)$$

which when multiplied by  $\delta = O(1/n)$  evidently vanishes.  $\square$

**7.1. The fixed points of Proposition 3.1.** We now turn to analyzing the ODE of Proposition 3.1.

**Proof of Proposition 3.3.** At the fixed points of the ODE in Proposition 3.1,

$$\lambda k u_1^{k-1} = (kR^{2k-2} + \alpha)u_1, \quad \text{and} \quad 2c_\delta k R^{2k-2} = (2kR^{2k-2} + \alpha)u_2.$$

If  $u_1 = 0$ , then  $R^2 = u_2$  and there are two possible fixed points: either  $u_2 = 0$  or  $u_2$  solves

$$ku_2^{k-2}(2c_\delta - 2u_2) = \alpha.$$

Notice that if  $k = 2$ , this has a nontrivial solution of the form  $c_\delta - \frac{\alpha}{2} = u_2$ , provided  $\alpha < \alpha_c(2) := 2c_\delta$ , and if  $k > 2$ , this has a nontrivial solution provided  $\alpha \leq \max_{x \geq 0} kx^{k-2}(2c_\delta - 2x)$  at  $c_\delta(k-2)x^{k-3} - (k-1)x^{k-2} = 0$  i.e.,  $\frac{c_\delta(k-2)}{k-1} = x$ . This gives

$$\alpha < \alpha_c(k) := 2c_\delta^{k-1}k(k-1)^{-(k-1)}(k-2)^{k-2}.$$

Evidently when we take  $\alpha = 0$ , then its non-trivial solution is at  $u_2 = 1$  for all  $k \geq 2$ .

Alternatively, if  $u_1 \neq 0$  at a fixed point, then we can simplify further and get

$$\lambda u_1^{k-2} = R^{2k-2} + \alpha/k, \quad \text{and} \quad kR^{2k-2} = (kR^{2k-2} + \alpha)u_2,$$

so that at the fixed point,

$$u_1^{k-2} = \frac{kR^{2k-2} + \alpha}{\lambda k}, \quad \text{and} \quad u_2 = \frac{2c_\delta k R^{2k-2}}{2kR^{2k-2} + \alpha}.$$

For simplicity of calculations, set  $\alpha = 0$  as is the case in Proposition 3.1. Then, we simply get  $u_2 = c_\delta$ . In the case of  $k = 2$ , we also find that there is a solution if and only if  $\lambda > c_\delta$ , in which case  $R^2 = \lambda$ , from which together with  $R^2 = u_1^2 + u_2$ , we also get  $u_1 = \pm\sqrt{\lambda - c_\delta}$ .

In the general case of  $k > 2$ , we find that  $R^2 = c_\delta + \lambda^{-\frac{2}{k-2}}R^{\frac{4(k-1)}{k-2}}$ . This has real solutions (all of which have  $R \geq u_2 = c_\delta$  as required) whenever  $\lambda > \lambda_c(k)$  defined as

$$\lambda_c(k) := \left(\frac{c_\delta}{k}\right)^{k/2} \left(\frac{(2k-2)^{k-1}}{(k-2)^{(k-2)/2}}\right). \quad (7.6)$$

(Interpreting  $0^0 = 1$ , this returns  $\lambda_c(2) = c_\delta$ .) With this  $\lambda$ , whenever  $\lambda > \lambda_c(k)$ , the equation for  $R^2$  has exactly two real solutions, both of which are at least  $c_\delta$  which we can denote by

$$\begin{aligned} \rho_\dagger(k, \lambda) &:= \inf\{\rho \geq 1 : \lambda^{-\frac{2}{k-2}}\rho^{\frac{2(k-1)}{k-2}} - \rho + c_\delta = 0\}, \\ \rho_\star(k, \lambda) &:= \sup\{\rho \geq 1 : \lambda^{-\frac{2}{k-2}}\rho^{\frac{2(k-1)}{k-2}} - \rho + c_\delta = 0\}. \end{aligned}$$

When  $\lambda > \lambda_c(k)$ ,  $\rho_\dagger < \rho_\star$  and when  $\lambda = \lambda_c(k)$ , the two are equal. Given this, we can then solve for  $\tilde{u}_1$  at the corresponding fixed point, and find that they occur at

$$m_\dagger(k, \lambda) = \sqrt{\rho_\dagger - c_\delta}, \quad \text{and} \quad m_\star(k, \lambda) = \sqrt{\rho_\star - c_\delta}, \quad (7.7)$$

as claimed.  $\square$

**7.2. Effective dynamics for the population loss.** In practice, one is interested in tracking the loss, or ideally, the generalization error. In this subsection, we add the generalization error  $\Phi$  to our set of summary statistics and obtain limiting equations for its evolution from (3.4).

Recalling (7.2), the fact that  $\Phi$  is a localizable summary statistic follows from the facts that  $\|\nabla m\|, \|\nabla r^2\| \leq C(K)$ , and the fact that  $\Phi$  is a smooth  $n$ -independent function of  $m, r^2$ .

For simplicity of calculations let us stick to  $\alpha = 0$ .

$$\begin{aligned} f_\Phi &= \langle \nabla \Phi, \nabla \Phi \rangle = 4\lambda^2 k^2 m^{2(k-1)} - 8\lambda k^2 m^k R^{2k-2} + 4k^2 R^{4k-4} m^2 + 4k^2 r^2 R^{4k-4} \\ &= 4k^2 m^2 (\lambda^2 m^{2(k-2)} - 2\lambda m^{k-2} R^{2k-2} + R^{4k-4}) + 4k^2 r^2 R^{4k-4}. \end{aligned}$$

Next, consider the corrector for  $\Phi$ . For this, notice that

$$\begin{aligned} \frac{1}{2} \nabla^2 \Phi &= -\lambda k(k-1) m^{k-2} \nabla m^{\otimes 2} + k R^{2k-2} \nabla m^{\otimes 2} + k(k-1) R^{2(k-2)} (2m \nabla m + \nabla r^2) \otimes \nabla m \\ &\quad + k(k-1) R^{2(k-2)} (2m \nabla m \otimes \nabla r^2 + \nabla r^2 \otimes \nabla r^2) + \frac{1}{2} \partial_2 \phi \nabla^2 r^2. \end{aligned}$$

Recalling  $V$  from (7.4), and taking  $\delta = c_\delta/n$ , all the terms in  $\sum_{ij} V_{ij} \partial_i \partial_j \Phi$  vanish in the limit except the contribution from the  $\nabla^2 r^2$ , which yields  $g_\Phi = \lim_{n \rightarrow \infty} \delta \mathcal{L}^\delta \Phi = 4c_\delta k^2 R^{4(k-1)}$ . Finally, we wish to compute the volatility for the stochastic part of the evolution of  $\Phi$ . For this, consider  $\nabla \Phi V \nabla \Phi^T$  and notice that all the entries of that matrix are continuous functions of  $\mathbf{u}_n$  and thus go to zero when multiplied by  $\delta = O(1/n)$ .

**7.3. Diffusive limits at the equator.** In this subsection, we develop the stochastic limit theorems for the rescaled observables about the axis  $m = 0$ . Here we take as variables  $(\tilde{u}_1, \tilde{u}_2) = (\sqrt{nm}, r^2)$ . For simplicity of presentation, we take  $\alpha = 0$  and  $c_\delta = 1$ .

**Proof of Proposition 3.2.** We begin by checking localizability. The change from the original variables is in the  $J$  matrix, in which now  $\nabla \tilde{u}_1 = \sqrt{n} \nabla m = \sqrt{n} v$ . This does not affect items (1)–(2) of localizability; for item (3), notice that

$$\mathbb{E}[\langle \nabla H, \nabla m \rangle^4] = n^2 \mathbb{E}[\langle \nabla H, v \rangle^4] \leq n^2 \mathbb{E}[W_{1,\dots,1}^4] \leq Cn^2.$$

The second part of item (3) is unchanged since  $\nabla^2 \tilde{u}_1 = 0$ .

Computing the drifts,

$$\begin{aligned} \langle \nabla \Phi, \nabla \tilde{u}_1 \rangle &= -2k\lambda\sqrt{nm}^{k-1} + 2k\sqrt{n}R^{2k-2}m = -2k\lambda n^{-\frac{k-2}{2}} \tilde{u}_1^{k-1} + 2k(r^2 + (\tilde{u}_1^2/n))^{k-1} \tilde{u}_1, \\ \langle \nabla \Phi, \nabla r^2 \rangle &= 4kr^2 R^{2k-2} = 4kr^2(r^2 + (\tilde{u}_1^2/n))^{k-1}. \end{aligned}$$

Taking limits as  $n \rightarrow \infty$ , as long as  $\lambda$  is fixed in  $n$ , we see that  $\mathbf{f}$  is given by

$$f_{\tilde{u}_1} = \begin{cases} -2k\lambda \tilde{u}_1^{k-1} + 2k \tilde{u}_2^{k-1} \tilde{u}_1 & k = 2 \\ 2k \tilde{u}_2^{k-1} \tilde{u}_1 & k \geq 3 \end{cases}, \quad \text{and} \quad f_{\tilde{u}_2} = 4k \tilde{u}_2^k.$$

We turn to obtaining the correctors in these rescaled coordinates. Evidently  $\delta \mathcal{L} \tilde{u}_1 = 0$  still by linearity of  $\tilde{u}_1$ . Following the calculation for the corrector, it is now given by  $g_{\tilde{u}_2} = 4k \tilde{u}_2^{k-1}$ .

Next we consider the volatility of the stochastic process one gets in the limit. Recalling  $JVJ^T$  from (7.5), and noticing that the rescaling  $J \rightarrow \tilde{J}$  multiplies its  $(1, 1)$ -entry by  $n$  and its off-diagonal entries by  $\sqrt{n}$ , we find that in the new coordinates,

$$\tilde{J}V\tilde{J}^T = \begin{pmatrix} 4k(k-1)\tilde{u}_1^2 R^{2k-4} + 4knR^{2k-2} & 4k(k-1)\tilde{u}_1(R^2 - m)R^{2k-4} \\ 4k(k-1)\tilde{u}_1(R^2 - m)R^{2k-4} & 4k(k-1)(R^2 - m)^2 R^{2k-4} \end{pmatrix} \quad (7.8)$$

Multiplying by  $\delta = 1/n$  and taking the limit as  $n \rightarrow \infty$ , the only entry of this matrix that survives is from  $\Sigma_{11}$  where we get  $\Sigma_{11} = 4k \tilde{u}_2^{k-1}$  as claimed.  $\square$

Regarding the discussion in the  $k \geq 3$  case of (3.3), when  $\lambda_n = \Lambda n^{(k-2)/2}$ , observe that the first term in  $\langle \Phi, \nabla \tilde{u}_1 \rangle$  above would not vanish and would instead converge to  $-4k\Lambda \tilde{u}_1^{k-1}$ .

**7.4. Diffusive limit for the radius.** We now show how to rescale the radial term  $r^2$  to obtain a diffusive limit for  $r^2$  about  $r^2 = 1$ . (For readability, we take the case  $c_\delta = 1$  though an analogous result works for general  $c_\delta$ .) To this end, consider  $\tilde{\mathbf{u}}_n = (\tilde{u}_1, \tilde{u}_2) = (\sqrt{n}m, \sqrt{n}(r^2 - 1))$ . Now  $J$  is in terms of  $\nabla \tilde{u}_1 = \sqrt{n}\nabla m$  and  $\nabla \tilde{u}_2 = \sqrt{n}\nabla u_2$ . Let us verify localizability for  $\tilde{\mathbf{u}}_n$ ; the only changes as compared to the previous subsection are those entailing  $\tilde{u}_2$ .

For item (1),  $\|\nabla^2 \tilde{u}_2\|_{\text{op}} = O(\sqrt{n})$  and  $\nabla^3 \tilde{u}_2 = 0$ . For the first part of item (3),

$$\mathbb{E}[\langle \nabla H, \nabla \tilde{u}_2 \rangle^4] = n^2 \mathbb{E}[\langle \nabla H, 2(x - mv) \rangle^4] \lesssim n^2(R^{4k} + m^4)\mathbb{E}[W_{1,\dots,1}^4] \lesssim_K n^2,$$

where we used in the first inequality that the law of  $H$  is rotation invariant and  $H$  is a  $k$ -homogenous function. For the second part of item (3),

$$\mathbb{E}[\langle \nabla^2 \tilde{u}_2, \nabla H \otimes \nabla H - V \rangle^2] \leq n \text{Var}(\|\nabla H\|^2).$$

We now express

$$\text{Var}(\|\nabla H\|^2) = \sum_i \text{Var}((\partial_i H)^2) + \sum_{i \neq j} \text{Cov}((\partial_i H)^2, (\partial_j H)^2).$$

The  $\partial_i H$  are Gaussian with mean zero, and by (7.4), variance  $C'_k R^{2(k-2)} x_i^2 + C_k R^{2(k-1)}$  and covariance  $C_k x_i x_j R^{2(k-2)}$ . Recall the following fact about Gaussians: if  $X, Y$  are Gaussians with variances  $\sigma^2$  and covariance  $t$ , then  $\text{Cov}(X^2, Y^2) \leq Ct^2\sigma^4$  for some universal constant  $C$ . Also,  $\text{Var}(X^2) \leq C\sigma^4$ . Applying this to  $\partial_i H$ , we get

$$\text{Var}(\|\nabla H\|^2) \lesssim_K n + \sum_{i,j} x_i^2 x_j^2 \lesssim_K n.$$

Combined with the above, this gives a bound of  $n^2 = O(\delta^{-2})$  on the second part of item (3).

We now calculate the resulting drifts. For  $\mathbf{f}$ , write

$$\begin{aligned} \mathcal{A}_n u_1 &= -2k\lambda \mathbf{1}_{k=2} \tilde{u}_1^{k-1} + 2kr^{2(k-1)} \tilde{u}_1 = -2k\lambda \mathbf{1}_{k=2} \tilde{u}_1^{k-1} + 2k(1 + n^{-1/2} \tilde{u}_2)^{k-1} \\ \mathcal{A}_n \tilde{u}_2 &= 4kn^{1/2} r^2 (r^2 + (\tilde{u}_1^2/n))^{k-1} = 4kn^{1/2} (1 + n^{-1/2} \tilde{u}_2) (1 + n^{-1/2} \tilde{u}_2 + n^{-1} \tilde{u}_1^2)^{k-1} \\ &= 4kn^{1/2} + 4k^2 \tilde{u}_2 + o(1) \end{aligned}$$

We next calculate the prelimits of the corrector. Evidently  $\delta \mathcal{L} \tilde{u}_1 = 0$  still by linearity of  $\tilde{u}_1$  and

$$\delta \mathcal{L} \tilde{u}_2 = \sqrt{n} \delta \mathcal{L}^\delta r^2 = \frac{4}{\sqrt{n}} k \left( (n-1)R^{2k-2} + (k-1)(1 + n^{-1/2} \tilde{u}_2) R^{2k-4} \right)$$

Combining terms and sending  $n \rightarrow \infty$ , we obtain

$$f_{\tilde{u}_1} - g_{\tilde{u}_1} = -2k\lambda \mathbf{1}_{k=2} \tilde{u}_1^{k-1} + 2k, \quad \text{and} \quad f_{\tilde{u}_2} - g_{\tilde{u}_2} = 4k \tilde{u}_2.$$

It remains to compute the volatility of the stochastic process one gets in the limit. Recalling  $JVJ^T$  from (7.5) and noticing that the rescaling  $J$  to  $\tilde{J}$  has now multiplied all four of its entries by  $n$ , we find that in the new coordinates,

$$\tilde{J}V\tilde{J}^T = \begin{pmatrix} 4k(k-1)\tilde{u}_1^2 R^{2k-4} + 4knR^{2k-2} & 4k(k-1)n^{1/2}\tilde{u}_1(R^2 - m)R^{2k-4} \\ 4k(k-1)n^{1/2}\tilde{u}_1(R^2 - m)R^{2k-4} & 4k(k-1)n(R^2 - m)^2 R^{2k-4} \end{pmatrix}. \quad (7.9)$$

Multiplying by  $\delta = 1/n$  and taking the limit as  $n \rightarrow \infty$ , the two entries of this matrix that survive are  $\Sigma_{11}$  and  $\Sigma_{22}$ , where  $\Sigma_{11} = 4k$  and  $\Sigma_{22} = 4k(k-1)$ . All in all, we obtain (3.5).

## 8. PROOFS FOR THE BINARY GAUSSIAN MIXTURE MODEL

Recall the cross-entropy loss for the binary GMM with SGD from (4.1), and recall the set of summary statistics  $\mathbf{u}_n$  from (4.2).

**Lemma 8.1.** *The distribution of  $L((v, W))$  depends only on  $\mathbf{u}_n$  from (4.2). In particular, we have that  $\Phi(x) = \phi(\mathbf{u}_n)$  for some  $\phi$ . Furthermore,  $\mathbf{u}_n$  satisfy the bounds in item (1) of Definition 2.1 if  $E_K$  is the ball of radius  $K$  in  $\mathbb{R}^{2N+2}$ .*

*Proof.* Let  $X_\mu \sim \mathcal{N}(\mu, I/\lambda)$  and  $X_{-\mu} \sim \mathcal{N}(-\mu, I/\lambda)$ . Then, notice that

$$L((v, W)) \stackrel{d}{=} \begin{cases} -v \cdot g(WX_\mu) + \log(1 + e^{v \cdot g(WX_\mu)}) + p(v, W) & \text{w. prob. } 1/2 \\ \log(1 + e^{v \cdot g(-WX_\mu)}) + p(v, W) & \text{w. prob. } 1/2 \end{cases}.$$

Next, notice that as a vector,  $(W_1 X_\mu, W_2 X_\mu)$  is distributed as  $(m_1 + Z_{1,\mu} m_1 + Z_{1,\perp}, m_2 + Z_{2,\mu} m_2 + Z_{2,\perp})$ , where  $Z_{1,\mu}, Z_{2,\mu}$  are i.i.d.  $\mathcal{N}(0, \lambda^{-1})$ , and  $Z_{1,\perp}, Z_{2,\perp}$  are jointly Gaussian with means zero and covariance

$$\lambda^{-1} \begin{bmatrix} R_{11}^\perp & R_{12}^\perp \\ R_{12}^\perp & R_{22}^\perp \end{bmatrix} \quad (8.1)$$

Similarly, the distribution of  $WX_{-\mu}$  also only depends on  $(m_i, R_{ij}^\perp)_{i,j}$ . Finally,

$$p(v, W) = \frac{\alpha}{2} (v_1^2 + v_2^2 + m_1^2 + R_{11}^\perp + m_2^2 + R_{22}^\perp)$$

Therefore, at any point  $(v, W)$ , the law of  $L((v, W))$ , and thus  $\Phi$ , is simply a function of  $\mathbf{u}_n(v, W)$ . To see that the summary statistics satisfy the bounds of item (1) in Definition 2.1, write  $\nabla = (\partial_{v_1}, \partial_{v_2}, \nabla_{W_1}, \nabla_{W_2})$ . Then

$$J = (\nabla u_\ell)_\ell = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu & 0 & W_2^\perp & 2W_1^\perp & 0 \\ 0 & 0 & 0 & \mu & W_1^\perp & 0 & 2W_2^\perp \end{bmatrix}^\top \quad (8.2)$$

For the higher derivatives, evidently we only have second derivatives in the last 3 variables each of which is given by a block diagonal matrix where only one block is non-zero and is given by an identity matrix. The third derivatives of all elements of  $\mathbf{u}_n$  are zero.  $\square$

We can now express the loss, the population loss, and their respective derivatives and they (their laws at a fixed point) will evidently only depend on the summary statistics. One arrives at the following expressions for  $\nabla L$  by direct calculation from (4.1).

$$\nabla_{v_i} L = (W_i \cdot X) \mathbf{1}_{W_i \cdot X \geq 0} (-y + \sigma(v \cdot g(WX))) + \alpha v_i \quad (8.3)$$

$$\nabla_{W_i} L = v_i X \mathbf{1}_{W_i \cdot X \geq 0} (-y + \sigma(v \cdot g(WX))) + \alpha W_i \quad (8.4)$$

In what follows, for an arbitrary vector  $w \in \mathbb{R}^N$ , we use the notation

$$\mathbf{A}_i = \mathbb{E}[X \mathbf{1}_{W_i \cdot X \geq 0} (-y + \sigma(v \cdot g(WX)))] \quad (8.5)$$

(Notice that if  $w \in \{\mu, W_i, W_i^\perp\}$ , then  $\mathbf{A}_i \cdot w$  is only a function of  $\mathbf{u}_n$  by the same reasoning as used in Lemma 8.1.) Then, we can also easily express

$$\nabla_{v_i} \Phi = W_i \cdot \mathbf{A}_i + \alpha v_i \quad \nabla_{W_i} \Phi = v_i \mathbf{A}_i + \alpha W_i \quad (8.6)$$

and for  $H = L - \Phi$ ,

$$\nabla_{v_i} H = W_i \cdot \left( X \mathbf{1}_{W_i \cdot X \geq 0} (-y + \sigma(v \cdot g(WX))) - \mathbf{A}_i \right), \quad (8.7)$$

$$\nabla_{W_i} H = v_i \left( X \mathbf{1}_{W_i \cdot X \geq 0} (-y + \sigma(v \cdot g(WX))) - \mathbf{A}_i \right). \quad (8.8)$$

Finally, the matrix  $V$  can be expressed as follows:

$$\begin{aligned} V_{v_i, v_j} &= \mathbb{E}[(W_i \cdot X)(W_j \cdot X) \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \sigma(v \cdot g(WX)))^2] - (W_i \cdot \mathbf{A}_i)(W_j \cdot \mathbf{A}_j) \\ V_{v_i, W_j} &= v_j \mathbb{E}[(W_i \cdot X) X \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \sigma(v \cdot g(WX)))^2] - v_j (W_i \cdot \mathbf{A}_i) \mathbf{A}_j \\ V_{W_i, W_j} &= v_i v_j \mathbb{E}[X^{\otimes 2} \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \sigma(v \cdot g(WX)))^2] - v_i v_j \mathbf{A}_i \otimes \mathbf{A}_j. \end{aligned} \quad (8.9)$$

Let us conclude this subsection with the following simple preliminary bounds that will be useful towards establishing the conditions of  $\delta_n$ -localizability from Definition 2.1, and the promised limiting equations. The proofs of these are straightforward using Gaussianity and are provided in Section 10 for completeness.

**Lemma 8.2.** *Fix  $w \in \mathbb{R}^n$ . We have  $\mathbb{E}[|X \cdot w|^8] \lesssim (w \cdot \mu)^8 + \|w\|^8 \lambda^{-4}$  and  $\|\mathbf{A}_i\| \leq C(\mathbf{u}_n)$ .*

**Lemma 8.3.** *For each  $i$ , for every  $R_{ii}^\perp < \infty$  and every  $m_i > 0$ , we have*

$$\lim_{\lambda \rightarrow \infty} \mathbb{P}(W_i \cdot X_\mu < 0) = 0. \quad (8.10)$$

For every  $v_i, R_{ij}^\perp$  and  $m_i \neq 0$  for  $i, j = 1, 2$ , we have

$$\lim_{\lambda \rightarrow \infty} \mathbb{E}[|\sigma(v \cdot g(WX_\mu)) - \sigma(v \cdot g(m))|] = 0. \quad (8.11)$$

**Fact 8.1.** *Fix  $\mu \in S^{N-1}(1)$ , and let  $g(x) = x \vee 0$  and  $X_\mu \sim \mathcal{N}(\mu, I/\lambda)$ . There is a function  $C : \mathbb{R}^2 \rightarrow \mathbb{R}_+$  such that for all  $\lambda > 0$ ,  $\theta \in \mathbb{R}$ , and  $(v_i, W_i) \in \mathbb{R} \times \mathbb{R}^N$ ,*

$$\mathbb{E}[\exp(\theta v_i g(W_i \cdot X_\mu))] \leq \exp\left(\theta v_i m_i + \frac{1}{2\lambda} \theta^2 v_i^2 R_{ii}^\perp\right).$$

**8.1. Verifying the conditions of Theorem 2.3 for fixed  $\lambda$ .** Throughout this section we will take  $\mu = e_1$ . By rotational invariance of the problem, this is without loss of generality, and only simplifies certain expressions. The  $\delta_n$ -localizability can be seen by application of the moment bounds listed above.

**Lemma 8.4.** *For  $\delta_n = O(1/N)$  and any fixed  $\lambda$ , the 2-layer GMM with observables  $\mathbf{u}_n$  is  $\delta_n$ -localizable for  $E_K$  being balls of radius  $K$  about the origin in  $\mathbb{R}^7$ .*

*Proof.* The condition on  $\mathbf{u}_n$  was satisfied per Lemma 8.1. Recalling  $\nabla \Phi$  from (8.6), one can verify that the norm of each of the four terms in  $\nabla \Phi$  is individually bounded, using the Cauchy–Schwarz inequality together with the bound of Lemma 8.2 on  $\|\mathbf{A}_i\|$ .

Next, consider bounding  $\mathbb{E}[\|\nabla H\|^8]$  by  $\sum_{i=1,2} \mathbb{E}[\|\nabla_{v_i} H\|^8] + \mathbb{E}[\|\nabla_{W_i} H\|^8]$ , and recall the expressions for  $\nabla H$  from (8.7)–(8.8). Using the trivial bound  $|\sigma(x)| \leq 1$ , and the inequality  $(a+b)^8 \leq C(a^8 + b^8)$ , for  $i \in \{1, 2\}$ , the first term is at most  $C(\mathbb{E}[|X \cdot W_i|^8] + \|W_i\|^8 \|\mathbf{A}_i\|^8)$  which is bounded by a constant depending continuously on  $\mathbf{u}_n$  per Lemma 8.2. If we let  $Z$  be a standard Gaussian, the quantity  $\mathbb{E}[\|\nabla_{W_i} H\|^8]$  is controlled by

$$C \left( v_i^8 \mathbb{E}[\|X \mathbf{1}_{W_i \cdot X \geq 0} \sigma(-v \cdot g(WX))\|^8] + v_i^8 \|\mathbf{A}_i\|^8 \right) \leq C |v_i|^8 \left( 1 + \frac{\mathbb{E}[|Z|^8]}{\lambda^4} \right).$$

Using the well-known bound that  $\mathbb{E}[|Z|^8] \leq N^4$ , and the fact that  $\delta = O(1/N)$ , we see that this is at most  $C\delta^{-4}$ . We next verify the claimed bound that

$$\delta_n^2 \sup_i \sup_{x \in \mathbf{u}_n^{-1}(E_K)} \mathbb{E}[\langle \nabla H, \nabla u_i \rangle^4] \leq C(K). \quad (8.12)$$

When  $u_i$  is  $v_i$ , this is simply a fourth moment bound on  $\nabla_{v_i} H$ , which follows from the 8'th moment by Jensen's inequality. When  $u_i$  is  $m_i$ , or  $R_{ij}^\perp$ , the bound follows from

$$\mathbb{E}[\langle \nabla_{W_i} H, w \rangle^4] \leq C|v_i|^4 (\mathbb{E}[|X \cdot w|^4] + \|w\|^4 \|\mathbf{A}_i\|^4),$$

for choices of  $w$  being either  $\mu$  in which case  $\|w\| = 1$  or  $W_i^\perp$  in which case  $\|w\| = R_{ii}^\perp$ . For each  $K$ , this is at most some constant  $C(K)$  using the two bounds of Lemma 8.2.

Finally, consider the quantity  $\mathbb{E}[\langle \nabla^2 u, \nabla H \otimes \nabla H - V \rangle^2]$ . This is only non-zero for  $u \in \{R_{ij}^\perp\}$  for which  $\nabla^2 u$  is a block-identity matrix, having operator norm at most 2 in all cases. Therefore, this quantity is at most  $4\mathbb{E}[\|\nabla H\|^4]$  which is at most  $N^2$  by the above proved second item in the definition of localizability. This is therefore  $O(\delta_n^{-2}) = o(\delta_n^{-3})$  as needed.  $\square$

**Proof of Proposition 4.1.** The convergence of the population drift to  $\mathbf{f}$  from Proposition 4.1 follows by taking the inner products of  $\nabla L$  from (8.6) with the rows of  $J$  from (8.2), and noticing that  $\mathbf{A}_i^\mu$  from (4.3) is exactly  $\mathbf{A}_i \cdot \mu$  and  $\mathbf{A}_i^\perp$  from (4.3) is exactly  $\mathbf{A}_i \cdot W_j^\perp$ .

Next consider the convergence of the correctors to the claimed  $\mathbf{g}$ . The variables  $u \in \{v_1, v_2, m_1, m_2\}$  are linear so  $\mathcal{L}_n u = 0$  and for these,  $\mathbf{g}_u = 0$ . For  $u = R_{ij}^\perp$  for  $i, j \in \{1, 2\}$ , the relevant entries in  $V$  are those corresponding to  $W_i^\perp$  and  $W_j^\perp$ . For ease of notation, in what follows let  $\pi = \sigma(v \cdot g(WX))$ .

For ease of calculation taking  $\mu = e_1$ , we have  $\mathcal{L}_n R_{ij}^\perp = \sum_{k \neq 1} V_{W_{ik}, W_{jk}}$ , which by (8.9), and the choice of  $\delta_n = c_\delta/N$ , is given by

$$\begin{aligned} \delta_n \mathcal{L}_n R_{ij}^\perp &= \frac{c_\delta}{N} \sum_{k \neq 1} v_i v_j \left( \mathbb{E}[(X \cdot e_k)^2 \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \pi)^2] - (\mathbf{A}_i \cdot e_k)(\mathbf{A}_j \cdot e_k) \right) \\ &= \frac{c_\delta}{N} v_i v_j \left( \mathbb{E}[\|X^\perp\|^2 \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \pi)^2] - \langle \mathbf{A}_i - \mathbf{A}_i^\mu \mu, \mathbf{A}_j - \mathbf{A}_j^\mu \mu \rangle \right). \end{aligned} \quad (8.13)$$

Consider the two terms separately. First, rewrite  $\frac{1}{N} \mathbb{E}[\|X^\perp\|^2 \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \pi)^2]$  as

$$\mathbb{E}\left[\left(\frac{1}{N} \|X^\perp\|^2 - \lambda^{-1}\right) \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \pi)^2\right] + \lambda^{-1} \mathbf{B}_{ij}.$$

Of course the second term is exactly what we want to be  $g_u$ , so we will show the first term here goes to zero. By Cauchy–Schwarz, if  $Z \sim \mathcal{N}(0, I - e_1^{\otimes 2})$ , the first term above is at most  $\lambda^{-1} \mathbb{E}[(\frac{\|Z\|^2}{N} - 1)^2]^{1/2} \leq \frac{2}{\lambda \sqrt{N}}$ , where we used the fact that for a standard Gaussian,  $g \sim \mathcal{N}(0, 1)$ , we have  $\mathbb{E}[(g^2 - 1)^2] = 2$ . It remains to show the inner product term in (8.13) goes to zero as  $n \rightarrow \infty$ . For this term, rewrite

$$\frac{1}{N} \langle \mathbf{A}_i - \mathbf{A}_i^\mu \mu, \mathbf{A}_j - \mathbf{A}_j^\mu \mu \rangle = \frac{1}{N} \mathbb{E}[(X_1^\perp \cdot X_2^\perp) \mathbf{1}_{W_i \cdot X_1 \geq 0} \mathbf{1}_{W_j \cdot X_2 \geq 0} (-y + \pi_1)(-y + \pi_2)],$$

where  $X_1, X_2$  are i.i.d. copies of  $X$ , and  $\pi_1, \pi_2$  are the corresponding  $\sigma(v \cdot g(WX_1))$  and  $\sigma(v \cdot g(WX_2))$ . By Cauchy–Schwarz, if  $Z, Z'$  are i.i.d.  $\mathcal{N}(0, I - e_1^{\otimes 2})$ , this is at most  $\frac{1}{\lambda N} \mathbb{E}[(Z \cdot Z')^2]^{1/2} \leq \frac{1}{\lambda \sqrt{N}}$ . This term therefore also vanishes as  $n \rightarrow \infty$ , yielding the desired limit for the corrector,

$$g_{R_{ij}^\perp} = \frac{c_\delta v_i v_j}{\lambda} \mathbb{E}[\mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \pi)^2] = \frac{c_\delta v_i v_j}{\lambda} \mathbf{B}_{ij}.$$

which we emphasize is only a function of  $\mathbf{u}_n$ . We lastly need to show that the diffusion matrix  $\Sigma_n$  goes to zero as  $n \rightarrow \infty$  when  $\delta_n = O(1/n)$ . This is straightforward to see by considering any element of  $JVJ^T$  and using Cauchy–Schwarz together with the two bounds of Lemma 8.2 to bound it in absolute value by some  $C(K)$  independent of  $n$ . Then when multiplying by any  $\delta_n = o(1)$ , this entire matrix will evidently vanish.  $\square$

**8.2. The small-noise limit of the effective dynamics.** One can now take a  $\lambda \rightarrow \infty$  limit to arrive at the ODE system of Proposition 4.2.

**Proof of Proposition 4.2.** We begin with considering  $\lim_{\lambda \rightarrow \infty} \mathbf{A}_i^\mu$ : its limiting value will depend on the signs of both  $m_1$  and  $m_2$ . We can express  $\mathbf{A}_i^\mu$  from (4.3) as

$$\begin{aligned}\mathbb{E}[(X \cdot \mu) \mathbf{1}_{W_i \cdot X \geq 0}(-y + \sigma(v \cdot g(WX)))] &= \frac{1}{2} \mathbb{E} \left[ (X_\mu \cdot \mu) \mathbf{1}_{W_i \cdot X_\mu \geq 0}(-1 + \sigma(v \cdot g(WX_\mu))) \right] \\ &\quad + \frac{1}{2} \mathbb{E} \left[ (-X_\mu \cdot \mu) \mathbf{1}_{W_i \cdot X_\mu \leq 0} \sigma(v \cdot g(-WX_\mu)) \right].\end{aligned}$$

We claim that the two terms on the right-hand side converge to  $-\frac{1}{2} \mathbf{1}_{m_i > 0} \sigma(-v \cdot g(m))$  and  $-\frac{1}{2} \mathbf{1}_{m_i < 0} \sigma(v \cdot g(-m))$  respectively. This follows by e.g., writing the difference as

$$\begin{aligned}\mathbb{E} \left[ (X_\mu \cdot \mu) \mathbf{1}_{W_i \cdot X_\mu \geq 0} \sigma(-v \cdot g(WX_\mu)) \right] - \mathbf{1}_{m_i \geq 0} \sigma(-v \cdot g(m)) &= \mathbb{E} \left[ (X_\mu \cdot \mu - 1) \mathbf{1}_{W_i \cdot X_\mu \geq 0} \sigma(-v \cdot g(WX_\mu)) \right] \\ &\quad + \mathbb{E} \left[ (\mathbf{1}_{W_i \cdot X_\mu \geq 0} - \mathbf{1}_{m_i \geq 0}) \sigma(-v \cdot g(WX_\mu)) \right] \\ &\quad + \mathbf{1}_{m_i \geq 0} \mathbb{E} \left[ \sigma(-v \cdot g(WX_\mu)) - \sigma(-v \cdot g(m)) \right].\end{aligned}\tag{8.14}$$

Call these three terms  $I$ ,  $II$ , and  $III$ . For  $I$ , we use the fact that  $\mathbb{E}[|X_\mu \cdot \mu - 1|]$  goes to zero as  $\lambda \rightarrow \infty$ ;  $II$  is evidently bounded by  $\mathbb{P}(W_i \cdot X_\mu < 0)$  when  $m_i > 0$  or its symmetric counterpart when  $m_i < 0$ —both vanishing as  $\lambda \rightarrow \infty$  per (8.10) in Lemma 8.3; finally,  $III$  goes to zero as  $\lambda \rightarrow \infty$  by (8.11) in Lemma 8.3.

Putting the above together, we find that

$$\lim_{\lambda \rightarrow \infty} \mathbf{A}_i^\mu = -\frac{1}{2} \mathbf{1}_{m_i > 0} \sigma(-v \cdot g(m)) - \frac{1}{2} \mathbf{1}_{m_i < 0} \sigma(v \cdot g(-m)),$$

at which point, we see that if  $m_1, m_2 \geq 0$ , this becomes  $\frac{1}{2} \sigma(-v \cdot m)$ , as it is if  $m_1, m_2 \leq 0$ . If  $m_1 \geq 0$  and  $m_2 \leq 0$ , then you get  $\lim_{\lambda} \mathbf{A}_1^\mu = -\frac{1}{2} \sigma(-v_1 m_1)$  and  $\lim_{\lambda} \mathbf{A}_2^\mu = -\frac{1}{2} \sigma(-v_2 m_2)$  and likewise if  $m_1 \leq 0$  and  $m_2 \geq 0$ .

Next consider the limit as  $\lambda \rightarrow \infty$  of  $\mathbf{A}_{ij}^\perp$  from (4.3), which we claim converges to 0. Write

$$\begin{aligned}\mathbf{A}_{ij}^\perp &= -\frac{1}{2} \mathbb{E} \left[ (X_\mu \cdot W_j^\perp) \mathbf{1}_{W_i \cdot X \geq 0} \sigma(-v \cdot g(WX_\mu)) \right] \\ &\quad - \frac{1}{2} \mathbb{E} \left[ (X_\mu \cdot W_j^\perp) \mathbf{1}_{W_i \cdot X_\mu < 0} \sigma(v \cdot g(-WX_\mu)) \right].\end{aligned}\tag{8.15}$$

These two terms are bounded similarly. The absolute value of the first of these is bounded by  $(1/2) \mathbb{E}[|X_\mu \cdot W_j^\perp|]$  which is at most  $(1/2) \sqrt{R_{jj}^\perp} \lambda^{-1/2}$  by (8.2). The second is analogously bounded. These evidently go to zero as  $\lambda \rightarrow \infty$ .

Finally, since  $|\mathbf{B}_{ij}| \leq 1$ , the quantity  $g_{R_{ij}^\perp} = c_\delta \frac{v_i v_j}{\lambda} \mathbf{B}_{ij}$  evidently goes to zero as  $\lambda \rightarrow \infty$ .  $\square$

*Remark 5.* The above argument used  $m_i \neq 0$  for the limit of  $\mathbf{A}_i^\mu$ . If one considers the cases when  $m_i = 0$ , the limiting drifts still apply. For this, it suffices to show that if  $m_i = 0$ , then  $\mathbf{A}_i^\mu$  converges to zero. Without loss of generality, suppose  $m_1 = 0$  and consider

$$\mathbf{A}_1 \cdot \mu = \mathbb{E} [Z_{1,\mu} \mathbf{1}_{Z_{1,\perp} \geq 0} \sigma(-v \cdot g(Z_{1,\perp}, m_2 Z_{2,\mu} + Z_{2,\perp}))].$$

This is zero independently of  $\lambda$  by independence of  $Z_{1,\mu}$  from the other Gaussians in the expectation.

Evidently, every fixed point must have  $R_{ij}^\perp = 0$ . Furthermore, if we let  $u_i = v_i - m_i$ , then

$$\dot{u}_i = \begin{cases} -\frac{u_i}{2}\sigma(-v \cdot m) - \alpha u_i & m_1 m_2 > 0 \\ -\frac{u_i}{2}\sigma(-v_i m_i) - \alpha u_i & \text{else} \end{cases},$$

and therefore every fixed point of the ODE system must have  $u_i = 0$ , which is to say  $v_i = m_i$ . Therefore, it suffices to characterize the fixed points in terms of  $(v_1, v_2)$  as claimed. This reduces to  $v_i \sigma(-\|v\|^2) = 2\alpha v_i v_1 v_2 > 0$  if  $v_1 v_2 > 0$  and  $v_i \sigma(-v_i^2) = 2\alpha v_i$  otherwise. Observe first that the point  $(v_1, v_2) = (0, 0)$  is a fixed point of this system. If  $(v_1, v_2) \neq 0$ , then dividing out by  $v_i$ , the above reduces to  $\sigma(-\|v\|^2) = 2\alpha$  if  $v_1 v_2 > 0$  and  $\sigma(-v_i^2) = 2\alpha$  otherwise. Recalling that  $C_\alpha = -\log(2\alpha) = \log(1 - 2\alpha) - \log(2\alpha)$  we obtain the claimed set of fixed points by inverting these equations (they only have a solution if  $\alpha < 1/4$ ).

In order to study the stability of the various fixed points, notice first that the ODE system of Proposition 4.2 is a gradient system for the  $\lambda = \infty$  population loss,

$$\Phi(v, m) = \frac{1}{2} \left( \log(1 + e^{-v \cdot g(m)}) + \log(1 + e^{v \cdot g(-m)}) \right) + \frac{\alpha}{2} \sum_{i=1,2} (v_i^2 + m_i^2 + R_{ii}^\perp).$$

Since it is a gradient system, with only the specified fixed points, the stability of a fixed point can be deduced by showing it is the minimizer of  $\Phi$ . In particular, the values of  $\Phi$  at its critical points are given by  $\Phi_0 = \log 2$  at  $v_1 = v_2 = 0$ ,  $\Phi_+ = \frac{1}{2}(\log 2 + \log(1 + e^{-C_\alpha}) + \alpha C_\alpha)$  when  $v_1 v_2 > 0$ , and  $\Phi_- = \log(1 + e^{-C_\alpha}) + 2\alpha C_\alpha$  when  $v_1 v_2 < 0$ . It is a simple calculus exercise to show that the smallest of these is  $\Phi_0$  when  $\alpha > 1/4$  and  $\Phi_-$  when  $\alpha < 1/4$ .

To show that each of the other critical points are all unstable, one can find a direction along which the dynamical system is locally repelled from it. For instance, we will show that the ring of fixed points with  $v_i = m_i$  and  $R_{ij}^\perp = 0$  with  $v_1 v_2 \leq 0$  is unstable, by showing a repelling direction arbitrarily close to the point  $v_1 = -\sqrt{C_\alpha}$ ,  $v_2 = 0$ . If  $v_1 = -\sqrt{C_\alpha}$  and  $v_2 = \epsilon > 0$ , then  $\dot{v}_2$  there reduces to  $\epsilon(\frac{\sigma(-\epsilon^2)}{2} - \alpha)$ , and as long as  $\alpha < 1/4$ , there exists  $\epsilon > 0$  such that  $\sigma(-\epsilon^2) > 2\alpha$  so  $\dot{v}_2 > 0$  for all  $\epsilon$  small enough.

**8.3. Rescaled effective dynamics around unstable fixed points.** In this section, we consider scaling limits of the rescaled effective dynamics in their noiseless limit, where the rescaling is about the unstable set of fixed points given by the quarter circle  $v_1^2 + v_2^2 = C_\alpha$  per item (2) of Proposition 4.2. Let  $\delta_n = c_s/N$ , and fix  $(a_1, a_2) \in \mathbb{R}_+^2$  with  $a_1^2 + a_2^2 = C_\alpha$ , and let  $\mathbf{u}_n$  be the variables of (4.2) with  $v_i, m_i$  replaced by  $\tilde{v}_i = \sqrt{N}(v_i - a_i)$  and  $\tilde{m}_i = \sqrt{N}(m_i - a_i)$ .

**Proof of Proposition 4.3.** We start by considering the drift process for these rescaled variables. Notice that the rescaling induces the transformation  $\tilde{J}$  multiplying  $J$  by  $\sqrt{N}$  in its entries corresponding to  $v_i, m_i$ . The fact that the rescaled variables satisfy the conditions of Theorem 2.3 follows as in Lemma 8.4 with the only distinction arising in the bound on (8.12), where previously we did not use the  $\delta_n^2$  factor—in the new coordinates, the factor of  $\sqrt{N}$  raised to the fourth power is cancelled out by  $\delta_n^2$  as long as  $\delta_n = O(1/N)$ .

For the population drift of the new variables, if the variables  $\tilde{v}_i, \tilde{m}_i$  are in a ball of radius  $K$  in  $\mathbb{R}^4$  (which we take to be our  $E_K$ ), the signs of  $m_i$  agree, and therefore

$$f_{\tilde{v}_i} = -\sqrt{N} \frac{v_i}{2} \sigma(-v \cdot m) + \alpha \sqrt{N} m_i \quad \text{and} \quad f_{\tilde{m}_i} = -\sqrt{N} \frac{m_i}{2} \sigma(-v \cdot m) + \alpha \sqrt{N} v_i.$$

We wish to claim that these expressions have consistent limits when  $\tilde{v}_i, \tilde{m}_i$  are localized to  $E_K$  for fixed  $K$ . notice that in  $m_i = a_i + N^{-1/2} \tilde{m}_i$  and  $v_i = a_i + N^{-1/2} \tilde{v}_i$ , and using  $\sum a_j^2 = C_\alpha$ ,

$$v \cdot m = C_\alpha + N^{-1/2} \sum_{j=1,2} a_j (\tilde{v}_j + \tilde{m}_j) + O(1/n).$$

Now Taylor expanding the sigmoid function, and using the definition of  $C_\alpha$ , we get

$$\begin{aligned}\sigma(-v \cdot m) &= \sigma(-C_\alpha) + (v \cdot m - C_\alpha)\sigma(-C_\alpha)(1 - \sigma(-C_\alpha)) + O(n^{-1}) \\ &= 2\alpha + N^{-1/2}a_j \left( \sum_{j=1,2} (\tilde{v}_j + \tilde{m}_j)(2\alpha)(1 - 2\alpha) \right) + O(n^{-1}).\end{aligned}$$

Plugging these into the earlier expressions for  $f_{\tilde{v}_i}$ , we see that

$$\begin{aligned}f_{\tilde{v}_i} &= -\frac{N^{1/2}a_i + \tilde{m}_i}{2} \left( 2\alpha + \frac{a_j}{N^{1/2}} \sum_{j=1,2} (\tilde{v}_j + \tilde{m}_j)(2\alpha)(1 - 2\alpha) + O\left(\frac{1}{n}\right) \right) + \alpha(n^{1/2}a_i + \tilde{v}_i) \\ &= -\alpha\tilde{m}_i + \alpha\tilde{v}_i - a_i(\alpha - 2\alpha^2) \sum_{j=1,2} a_j(\tilde{v}_j + \tilde{m}_j) + O(n^{-1/2}).\end{aligned}$$

Taking the limit as  $n \rightarrow \infty$ , this yields exactly the population drift claimed for the  $\tilde{v}_i$  variable. The calculation for  $f_{\tilde{m}_i}$  is analogous, and the equations for  $R_{ij}^\perp$  are evidently unchanged by the transformation of  $v_i, m_i$  to  $\tilde{v}_i, \tilde{m}_i$ . Furthermore, these variables are still linear so no corrector is introduced.

We now turn to computing the limiting diffusion matrix  $\Sigma$  in the new variables  $\tilde{v}_i, \tilde{m}_i$ . We first use the following expression for the matrix  $V$  when  $\lambda = \infty$ , by taking the  $\lambda = \infty$  in (8.9):

$$V_{v_i, v_j} = \frac{m_i m_j}{4} \cdot \begin{cases} \sigma(-v \cdot m)^2 & m_1 m_2 > 0 \\ \sigma(-v_i m_i) \sigma(-v_j m_j) & \text{else} \end{cases},$$

with similar expressions for  $V_{v_i, W_j}$  and  $V_{W_i, W_j}$ . Rewriting in  $\tilde{v}$  and  $\tilde{m}$ , we see that in  $E_K$ ,

$$V_{v_i, v_j} = \alpha^2 a_i a_j + O(n^{-1/2}), \quad V_{v_i, W_j} = \mu(\alpha^2 a_i a_j + O(n^{-1/2})),$$

$$V_{W_i, W_j} = \mu^{\otimes 2}(\alpha^2 a_i a_j + O(n^{-1/2})).$$

Now multiplying this on both sides by  $\tilde{J}$ , for the  $\tilde{\mathbf{u}}_n$  variables, the two factors of  $\sqrt{N}$  from  $\tilde{J}$  cancel out with the choice of  $\delta_n = 1/N$ , and in the  $n \rightarrow \infty$  limit, leave  $\tilde{\Sigma}_{v_i v_j} = \tilde{\Sigma}_{m_i m_j} = \tilde{\Sigma}_{v_i m_j} = \alpha^2 a_i a_j$  as claimed.  $\square$

## 9. PROOFS FOR THE XOR GAUSSIAN MIXTURE MODEL

Fix two orthogonal vectors  $\mu, \nu \in \mathbb{R}^N$  and recall the cross-entropy loss with penalty  $p(v, W) = \frac{\alpha}{2}(\|v\|^2 + \|W\|^2)$ . For the XOR GMM with SGD, the cross-entropy loss is given by

$$L(v, W) = -yv \cdot g(WX) + \log(1 + e^{v \cdot g(WX)}) + p(v, W) \tag{9.1}$$

where if the class label  $y = 1$ , then  $X$  is a symmetric binary Gaussian mixture with means  $\pm\mu$ , and if  $y = 0$ , then  $X$  is a symmetric Gaussian mixture with means  $\pm\nu$ . This has the same form as the loss for the 2-layer binary GMM, and we will find many similarities in the below between them. Indeed, the only difference is in the distribution of  $X$  conditionally on the class label  $y$  as described, and the fact that  $v$  is now in  $\mathbb{R}^K$  and  $W = (W_i)_{i=1,\dots,K}$  is now a  $K \times N$  matrix. In what follows we take  $n = KN + K$ . As such, all the formulae of (8.3)–(8.9) also hold for the XOR GMM, but with the law of  $(y, X)$  now understood differently.

*Remark 6.* We could also have added a bias at each layer, however the Bayes classifier in this problem is an “X” centered at the origin so we can safely take the biases to be 0.

**9.1. Summary statistics and localizability.** Recall the set of summary statistics  $\mathbf{u}_n$  from (5.1). The next lemma shows that  $\mathbf{u}_n$  form a good set of summary statistics.

**Lemma 9.1.** *The distribution of  $L((v, W))$  depends only on  $\mathbf{u}_n$  from (5.1). In particular, we have that  $\Phi(x) = \phi(\mathbf{u}_n)$  for some  $\phi$ . Furthermore,  $\mathbf{u}_n$  satisfy the bounds in item (1) of Definition 2.1 with an exhaustion by balls of  $\mathbb{R}^{KN+K}$ .*

*Proof.* Let  $X_w = \mathcal{N}(w, I/\lambda)$  for  $w \in \{\mu, -\mu, \nu, -\nu\}$ . Notice that the law of  $L$  at a fixed point  $(v, W) \in \mathbb{R}^{K+KN}$  can be written as

$$L((v, W)) \stackrel{d}{=} \begin{cases} -v \cdot g(WX_\mu) + \log(1 + e^{v \cdot g(WX_\mu)}) + p(v, W) & \text{w. prob. } 1/4 \\ -v \cdot g(WX_{-\mu}) + \log(1 + e^{v \cdot g(WX_{-\mu})}) + p(v, W) & \text{w. prob. } 1/4 \\ \log(1 + e^{v \cdot g(WX_\nu)}) + p(v, W) & \text{w. prob. } 1/4 \\ \log(1 + e^{v \cdot g(WX_{-\nu})}) + p(v, W) & \text{w. prob. } 1/4 \end{cases} \quad (9.2)$$

Next, notice that as a vector

$$WX_\iota = (m_i + Z_{i,\iota} m_i^\iota + Z_{i\perp})_{i=1,\dots,K} \quad \text{for } \iota \in \{\mu, \nu\},$$

where  $Z_{i,\iota}$  are i.i.d.  $\mathcal{N}(0, \lambda^{-1})$  and  $(Z_{i\perp})$  are jointly Gaussian with covariance matrix

$$\text{Cov}(Z_{i\perp}, Z_{j\perp}) = \lambda^{-1} R_{ij}^\perp.$$

Similarly, the law of  $WX_{-\iota}$  depends only on  $(m_i^\iota, R_{ij}^\perp)$ . Finally,

$$p(v, W) = \frac{\alpha}{2} \sum_{i=1,\dots,K} (v_i^2 + R_{ii}^\perp).$$

Therefore, at a fixed point  $(v, W)$  the law of  $L(v, W)$  is only a function of  $\mathbf{u}_n(v, W)$ .

To see that the summary statistics satisfy the bounds of item (1) in Definition 2.1, note that the non-zero entries of  $J = (\nabla u_\ell)_\ell$  are as follows.

$$\partial_{v_i} v_i = 1, \quad \nabla_{W_i} m_i^\mu = \mu, \quad \nabla_{W_i} m_i^\nu = \nu, \quad \nabla_{W_i} R_{jk}^\perp = W_j^\perp \delta_{ij} + W_k^\perp \delta_{ik}, \quad (9.3)$$

where  $\delta_{ij}$  is 1 if  $i = j$  and 0 otherwise. For higher derivatives, we only have second derivatives in the  $R_{jk}^\perp$  variables, each of which is given by a block diagonal matrix where only one block is non-zero and it is twice an identity matrix. Thus the operator norm of these second derivatives is 2. The third derivatives of all elements of  $\mathbf{u}_n$  are zero.  $\square$

In the following, let

$$\mathbf{A}_i = \mathbb{E}[X \mathbf{1}_{W_i \cdot X \geq 0}(-y + \sigma(v \cdot g(WX)))].$$

By the same reasoning as in Lemma 9.1, if  $w \in \{\mu, \nu, W_i, W_i^\perp\}$ , then  $w \cdot \mathbf{A}_i$  is only a function of  $\mathbf{u}_n$ . We then also have the conclusions of Lemma 8.2 for  $X$  distributed according to the XOR GMM by simply decomposing it into two mixtures, and we will therefore appeal to this lemma meaning its analogue for the XOR GMM.

**Lemma 9.2.** *For  $\delta = O(1/N)$  and any fixed  $\lambda$ , the 2-layer XOR GMM with observables  $\mathbf{u}_n$  is  $\delta_n$ -localizable for  $E_K$  being balls of radius  $K$  about the origin in  $\mathbb{R}^{4K+{K \choose 2}}$ .*

*Proof.* The condition on  $\mathbf{u}_n$  was satisfied per Lemma 9.1. Recalling  $\nabla \Phi$  from (8.6), one can verify that the norm of each of the four terms in  $\nabla \Phi$  is individually bounded, using the Cauchy–Schwarz inequality together with the bound of Lemma 8.2 on  $\|\mathbf{A}_i\|$ , naturally adapted to XOR. The remaining estimates are also analogous to the proof of Lemma 8.4 with the analogue of Lemma 8.2 applied.  $\square$

## 9.2. Effective dynamics for the XOR GMM.

*Proof of Proposition 5.1.* The convergence of the population drift to  $\mathbf{f}$  from Proposition 4.1 follows by taking the inner products of  $\nabla L$  from (8.6) with the rows of  $J$  from (9.3), and noticing that  $\mathbf{A}_i^\mu$  is exactly  $\mathbf{A}_i \cdot \mu$ ,  $\mathbf{A}_i^\nu$  is exactly  $\nu \cdot \mathbf{A}_i$ , and  $\mathbf{A}_{ij}^\perp$  is exactly  $\mathbf{A}_i \cdot W_j^\perp$ .

We next consider the population correctors. The fact that  $g_{v_i} = g_{m_i^\mu} = g_{m_i^\nu} = 0$  follows from the fact that the Hessians of  $v_i, m_i^\mu, m_i^\nu$  are zero. For the corrector  $g_{R_{ij}^\perp}$  for  $1 \leq i \leq j \leq K$ , the relevant entries of  $V$  are those corresponding to  $W_i^\perp$  and  $W_j^\perp$ . For ease of notation, in what follows let  $\pi = \sigma(v \cdot g(WX))$ .

Similar to the calculation of (8.13),

$$\begin{aligned} \delta_n \mathcal{L}_n R_{ij}^\perp &= \frac{c_\delta}{N} v_i v_j \left( \mathbb{E}[\|X^\perp\|^2 \mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (\pi - y)^2] \right. \\ &\quad \left. - \langle \mathbf{A}_i - \mathbf{A}_i^\mu \mu - \mathbf{A}_i^\nu \nu, \mathbf{A}_j - \mathbf{A}_j^\mu \mu - \mathbf{A}_j^\nu \nu \rangle \right). \end{aligned}$$

By the same arguments on the concentration of the norm of Gaussian vectors as used in the binary GMM case, then we deduce from this that

$$g_{R_{ij}^\perp} = \frac{c_\delta v_i v_j}{\lambda} \mathbb{E}[\mathbf{1}_{W_i \cdot X \geq 0} \mathbf{1}_{W_j \cdot X \geq 0} (-y + \pi)^2] = \frac{c_\delta v_i v_j}{\lambda} \mathbf{B}_{ij}.$$

Finally, let us establish that the limiting diffusion matrix is all-zero whenever  $\delta_n = o(1)$ . This follows exactly as it did in the proof of Proposition 4.1.  $\square$

**9.3. Small noise limit of the effective dynamics.** The aim of this section is to establish the following small-noise  $\lambda \rightarrow \infty$  limit of the effective dynamics ODE of Proposition 5.1. This will again be quite similar to the analogous proofs for the binary GMM in Section 8, and when these similarities are clear we will omit details.

**Proposition 9.1.** *In the  $\lambda \rightarrow \infty$  limit, the ODE from Proposition 5.1 converges to*

$$\begin{aligned} \dot{v}_i &= \frac{m_i^\mu}{4} \left( \mathbf{1}_{m_i^\mu \geq 0} \sigma(-v \cdot g(m^\mu)) - \mathbf{1}_{m_i^\mu < 0} \sigma(-v \cdot g(-m^\mu)) \right) \\ &\quad - \frac{m_i^\nu}{4} \left( \mathbf{1}_{m_i^\nu \geq 0} \sigma(v \cdot g(m^\nu)) - \mathbf{1}_{m_i^\nu < 0} \sigma(v \cdot g(-m^\nu)) \right) - \alpha v_i, \\ \dot{m}_i^\mu &= \frac{v_i}{4} \left( \mathbf{1}_{m_i^\mu \geq 0} \sigma(-v \cdot g(m^\mu)) - \mathbf{1}_{m_i^\mu < 0} \sigma(-v \cdot g(-m^\mu)) \right) - \alpha m_i^\mu, \\ \dot{m}_i^\nu &= -\frac{v_i}{4} \left( \mathbf{1}_{m_i^\nu \geq 0} \sigma(-v \cdot g(m^\nu)) - \mathbf{1}_{m_i^\nu < 0} \sigma(-v \cdot g(-m^\nu)) \right) - \alpha m_i^\nu, \end{aligned}$$

and  $\dot{R}_{ij}^\perp = -2\alpha R_{ij}^\perp$  for  $1 \leq i \leq j \leq K$ .

*Proof.* Let us begin with convergence of  $\mathbf{A}_i^\mu$ . We claim that it converges to

$$\lim_{\lambda \rightarrow \infty} \mathbf{A}_i^\mu = -\frac{1}{4} \mathbf{1}_{m_i^\mu > 0} \sigma(-v \cdot g(m^\mu)) - \frac{1}{4} \mathbf{1}_{m_i^\mu < 0} \sigma(v \cdot g(-m)).$$

In order to see this, expand

$$\begin{aligned} \mathbf{A}_i &= \frac{1}{4} \mathbb{E}[-X_\mu \mathbf{1}_{W_i \cdot X_\mu \geq 0} (\sigma(-v \cdot g(WX_\mu)))] - \frac{1}{4} \mathbb{E}[X_{-\mu} \mathbf{1}_{W_i \cdot X_{-\mu} \geq 0} (\sigma(-v \cdot g(WX_{-\mu})))] \\ &\quad + \frac{1}{4} \mathbb{E}[X_\nu \mathbf{1}_{W_i \cdot X_\nu \geq 0} (\sigma(v \cdot g(WX_\nu)))] + \frac{1}{4} \mathbb{E}[X_{-\nu} \mathbf{1}_{W_i \cdot X_{-\nu} \geq 0} (\sigma(v \cdot g(WX_{-\nu})))]. \end{aligned}$$

The point will be that when taking the inner product with  $\mu$ , the first two terms here contribute to the limit and the latter two vanish, while when taking the inner product with  $\nu$ , the first two terms vanish in the  $\lambda \rightarrow \infty$  limit while the latter two contribute.

Consider e.g., the first of the four terms above, and inner product with  $\mu$ . In this case, consider

$$\mathbb{E}[(X_\mu \cdot \mu) \mathbf{1}_{W_i \cdot X_\mu \geq 0} \sigma(-v \cdot g(WX_\mu))] - \mathbf{1}_{m_i^\mu \geq 0} \sigma(-v \cdot g(m^\mu)),$$

which is precisely the quantity that was exactly shown to go to zero as  $\lambda \rightarrow \infty$  in (8.14). To see that the third and fourth terms above go to zero when taking their inner product with  $\mu$ , observe that they become

$$|\mathbb{E}[(X_\nu \cdot \mu) \mathbf{1}_{W_i \cdot X_\nu \geq 0} \sigma(v \cdot g(WX_\nu))]| \leq \mathbb{E}[|X_\nu \cdot \mu|],$$

which by orthogonality of  $\mu$  and  $\nu$  is at most  $\lambda^{-1/2}$  by the reasoning of Lemma 8.2, therefore vanishing as  $\lambda \rightarrow \infty$ . Together with its analogue for  $X_{-\nu}$ , this implies the claim for the convergence of  $\mathbf{A}_i^\mu$ , as well as its analogous limit of  $\mathbf{A}_i^\nu$ .

We next consider the limit as  $\lambda \rightarrow \infty$  of  $\mathbf{A}_{ij}^\perp$ , which we claim goes to 0. Using the expansion of  $\mathbf{A}_i$  from earlier in this proof, we can consider  $\mathbf{A}_{ij}^\perp = \mathbf{A}_i \cdot W_j^\perp$  as four terms having the form of the terms in (8.15), which were there showed to go to zero as  $\lambda \rightarrow \infty$ . Since  $W_j^\perp$  here is orthogonal both to  $\mu$  and  $\nu$ , the same proof applies.

Finally, in order to see that the limit as  $\lambda \rightarrow \infty$  of  $g_{R_{ij}^\perp} = c_\delta \frac{v_i v_j}{\lambda} \mathbf{B}_{ij}$  is zero, which follows from the fact that  $|\mathbf{B}_{ij}| \leq 1$ .  $\square$

**Proposition 9.2.** *The fixed points of the ODE system of Proposition 9.1 are classified as follows. If  $\alpha > 1/8$ , then the only fixed point is at  $\mathbf{u}_n = \mathbf{0}$ .*

*If  $0 < \alpha < 1/8$ , then let  $(I_0, I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-)$  be any disjoint (possibly empty) subsets whose union is  $\{1, \dots, K\}$ . Corresponding to that tuple  $(I_0, I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-)$ , is a set of fixed points that have  $R_{ij}^\perp = 0$  for all  $i, j$ , and have*

- (1)  $m_i^\mu = m_i^\nu = v_i = 0$  for  $i \in I_0$ ,
- (2)  $m_i^\mu = v_i > 0$  such that  $\sum_{i \in I_\mu^+} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\nu = 0$  for all  $i \in I_\mu^+$ ,
- (3)  $-m_i^\mu = v_i > 0$  such that  $\sum_{i \in I_\mu^-} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\nu = 0$  for all  $i \in I_\mu^-$ ,
- (4)  $m_i^\nu = v_i < 0$  such that  $\sum_{i \in I_\nu^+} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\mu = 0$  for all  $i \in I_\nu^+$ ,
- (5)  $-m_i^\nu = v_i < 0$  such that  $\sum_{i \in I_\nu^-} v_i^2 = \text{logit}(-4\alpha)$  and  $m_i^\mu = 0$  for all  $i \in I_\nu^-$ .

*In the  $K = 4$  case, these form 39 connected sets of fixed points, and of which  $4! = 24$  are fixed points that are stable, corresponding to the possible permutations in which each of  $I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-$  are singletons.*

*Proof.* Evidently, any fixed point must have  $R_{ij}^\perp = 0$  for all  $i, j$ . Furthermore, the point  $v_i = m_i^\mu = m_i^\nu = 0$  for  $i = 1, \dots, K$  evidently forms a fixed point of the system. Now suppose there is some fixed point with  $v_i = 0$  for some  $i$ ; in that case, it must be that  $m_i^\mu = 0$  and  $m_i^\nu = 0$ . Therefore, we can select a subset  $I_0$  of  $\{1, \dots, K\}$  such that  $v_i = m_i^\mu = m_i^\nu = 0$  for  $i \in I_0$ .

For any such choice of  $I_0$ , consider next,  $i \notin I_0$ . We first claim that if  $v_i > 0$  at a fixed point, then  $m_i^\mu \in \{\pm v_i\}$  and  $m_i^\nu = 0$ , whereas if  $v_i < 0$  then  $m_i^\nu \in \{\pm v_i\}$  and  $m_i^\mu = 0$ . To see this, notice that at any fixed point,

$$\begin{aligned} 4\alpha m_i^\mu &= v_i \left( \mathbf{1}_{m_i^\mu \geq 0} \sigma(-v \cdot g(m^\mu)) - \mathbf{1}_{m_i^\mu < 0} \sigma(-v \cdot g(-m^\mu)) \right), \\ 4\alpha m_i^\nu &= -v_i \left( \mathbf{1}_{m_i^\nu \geq 0} \sigma(-v \cdot g(m^\nu)) - \mathbf{1}_{m_i^\nu < 0} \sigma(-v \cdot g(-m^\nu)) \right). \end{aligned}$$

Since  $\sigma$  is non-negative, if  $v_i > 0$ , the sign of the right-hand side of the first equation is the same as the sign of  $m_i^\mu$  so it can have a non-zero solution, while the sign of the right-hand side of the second equation is the opposite of the sign of  $m_i^\nu$ , so any such fixed point must have  $m_i^\nu = 0$ . To see that  $m_i^\mu = \pm v_i$  at such a fixed point, now set  $m_i^\nu = 0$  and take the fixed point equations for  $v_i$  and  $m_i^\mu$ ,

dividing one by  $v_i$  and the other by  $m_i^\mu$  to see that

$$4\alpha \frac{v_i}{m_i^\mu} = 4\alpha \frac{m_i^\mu}{v_i}, \quad \text{or} \quad v_i^2 = (m_i^\mu)^2,$$

as claimed. The fixed points having  $v_i < 0$  are solved symmetrically.

Our classification now reduces to understanding the possible values taken by  $(v_1, \dots, v_K)$  given their signs (when non-zero). Fix a partition  $(I_0, I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-)$  of  $\{1, \dots, K\}$  and consider the set of fixed points having  $m_i^\mu = m_i^\nu = v_i = 0$  for  $i \in I_0$ ,  $m_i^\mu = v_i > 0$  on  $I_\mu^+$  and so on as designated by Proposition 9.2; by the above any fixed point is of this form. It remains to check that the values of  $v_i$  on each of these sets are as described by the proposition.

In order to see this, fix e.g.,  $i \in I_\mu^+$ . Then,  $m_i^\mu = v_i$  and  $m_i^\nu = 0$ , and so the fixed point equations reduce to

$$4\alpha v_i = v_i \sigma(-v \cdot g(m^\mu)), \quad \text{or} \quad 4\alpha = \sigma\left(-\sum_{j \in I_\mu^+} v_j^2\right),$$

since the only coordinates where  $g(m^\mu)$  will be non-zero are  $j \in I_\mu^+$ , where  $m_j^\mu = v_j$ . Inverting the sigmoid function, this implies exactly the claimed  $\sum_{j \in I_\mu^+} v_j^2 = \text{logit}(-4\alpha)$ . The cases of  $I_\mu^-, I_\nu^+, I_\nu^-$  are analogous, concluding the proof.

The count of the number of connected components of fixed points this forms is sensitive to  $K$ , so for concreteness let us perform it when  $K = 4$ . We first notice that the fixed point at  $(0, \dots, 0)$  is disconnected from all others. Fixed points corresponding to some  $(I_0, \dots, I_\nu^-)$  are part of the same connected component of fixed points if one goes from one to the other by moving an element of  $I_\iota^\eta$  (for some  $\iota \in \{\mu, \nu\}$  and  $\eta \in \{\pm\}$ ) to  $I_0$  without making  $I_\iota^\eta$  empty, or by moving an element of  $I_0$  to a non-empty  $I_\iota^\eta$ .

We turn now to studying the stability of these various sets of fixed points. Observe that in the  $\lambda \rightarrow \infty$  limit, the dynamical system of Proposition 9.1 is a gradient system for the population loss

$$\begin{aligned} \Phi = & \frac{1}{4} \left( \log(1 + e^{-v \cdot g(m^\mu)}) + \dots + \log(1 + e^{-v \cdot g(-m^\nu)}) \right) \\ & + \frac{\alpha}{2} \sum_i (v_i^2 + (m_i^\mu)^2 + (m_i^\nu)^2 + R_{ii}^\perp). \end{aligned}$$

At a fixed point (which necessarily has  $v_i = m_i$ ,  $R_{ii}^\perp = 0$ , and is characterized by the partition of  $\{1, \dots, 4\}$  into  $I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-$ , this reduces to

$$\Phi = \frac{1}{4} \left( \log(1 + e^{-\sum_{i \in I_\mu^+} v_i^2}) + \dots + \log(1 + e^{-\sum_{i \in I_\nu^-} v_i^2}) \right) + \alpha \sum_i v_i^2$$

At this point, noticing that  $\sum_{i \in I_\mu^+} v_i^2$  is equal to  $C_\alpha = -\text{logit}(4\alpha)$  if  $I_\mu^+$  is non-empty and 0 if it is empty, and similarly for  $I_\mu^-, I_\nu^+, I_\nu^-$ , this turns into a simple optimization problem over the number of non-empty  $I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-$ . Just as in the binary GMM case, it becomes evident that when  $\alpha > 1/8$ , this is minimized at  $v_i = 0$  for all  $i$  (i.e., they are all empty and  $I_0 = \{1, \dots, 4\}$ , whereas when  $\alpha < 1/8$  the above is minimized when every one of  $I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-$  are all non-empty. This yields the global minima of  $\Phi$  in these coordinates, and ensures the fixed points we claimed were stable are indeed stable.

To show the instability of any other connected set of fixed points, the reasoning goes just as in the binary GMM case: consider a small perturbation of the specified critical region in the direction of the stable fixed points and it can be seen by examining the drifts directly, that the dynamical system has a repelling direction.  $\square$

*Remark 7.* When  $K > 4$ , the counting of connected components of fixed points of course changes. However, what is still clear by an identical calculation is that the sets of fixed points minimizing  $\Phi$  will still be  $(0, \dots, 0)$  when  $\alpha > 1/8$  and will be all fixed points that have all four of  $I_\mu^+, \dots, I_\nu^-$  being non-empty if  $\alpha < 1/8$ . Notice that when  $\alpha < 1/8$  and  $K > 4$ , even the set of stable fixed points become connected to form a single stable manifold.

**9.4.  $\frac{3}{32}$ -probability of ballistic convergence to an optimal classifier.** We now reason that when  $K = 4$  the ballistic effective dynamics of Proposition 9.1 is such that under an uninformative Gaussian initialization, the probability of being in a basin of attraction of one of the 24 stable fixed points is  $3/32$ . Begin by noticing that if the first layer weights are initialized as  $W_i \sim \mathcal{N}(0, I_N/N)$  independently for  $i = 1, \dots, 4$  and the second layer weights  $v_i$  are independent standard Gaussians, then the projection onto the coordinate system  $(v_i, m_i^\mu, m_i^\nu, R_{ij})$  is given by

$$\lim(\mathbf{u}_n)_*\mu_n = \mathcal{N}(0, 1)^{\otimes 4} \otimes \delta_0^{\otimes 4} \otimes \delta_0^{\otimes 4} \otimes \delta_{I_4}$$

The  $\delta$ -functions at zero for  $m_i^\mu, m_i^\nu$  however cause some trouble because of the indicator functions on the sign of  $m_i^\mu$  and  $m_i^\nu$  in the equations of Proposition 9.1.

In order to handle this, we can instead consider the pre-limit as a mixture (over all the possible signings  $\epsilon_i^\mu, \epsilon_i^\nu \in \{-1, +1\}$  of  $m_i^\mu, m_i^\nu$ ) of initializations where  $m_i^\mu \sim \epsilon_i^\mu |Z|$  for  $Z$  being Gaussian of variance  $1/N$ . For any such signing  $\epsilon$ , we take the limit per Proposition 9.1 to obtain the limiting ODE's with the indicators taking their values corresponding to the signings  $\epsilon$ . Thus the limit with the Gaussian initialization can be thought of as the equal mixture over the same signings  $\epsilon$  of the various ODE's obtained from Proposition 9.1 with the various indicators taking values 0 or 1. With that in mind, can interpret the initial  $m_i^\mu(0), m_i^\nu(0)$  as random variables that take values  $0^+$  and  $0^-$  with probability  $1/2$  each, the superscript being the signing dictating which indicator should be 1.

Under the flow of Proposition 9.1, if  $v_i(0)$  is positive, then  $m_i^\nu$  stays fixed at zero, and if  $m_i^\mu(0) = 0^-$  then  $m_i^\mu$  becomes negative infinitesimally quickly, whereas if  $m_i^\mu(0) = 0^+$  then it becomes positive infinitesimally quickly. At any rate, the sign of  $v_i$  never changes to negative from such an initialization, and similarly if  $v_i(0)$  is negative, the sign of  $v_i$  will never change to positive. As such, in order to have a chance at being in the basin of attraction of one of the stable fixed points outlined in Proposition 9.2, it must be the case that two of  $(v_i(0))_i$  have positive sign and two of them have negative sign; evidently this has probability  $\binom{4}{2}/2^4 = 3/8$ .

Given that two of  $v_i(0)$  are positive, and two of them are negative—say without loss of generality that  $i = 1, 2$  are the coordinates in which it is positive, and  $i = 3, 4$  are the coordinates in which it is negative—then the dynamical system for  $(v_1, v_2, m_1^\mu, m_2^\mu)$  is exactly the ballistic limit of the two-layer GMM studied in Section 4, for which we found that the probability of converging to a good classifier is  $1/2$ . Similarly, the dynamical system for  $(v_2, v_4, m_3^\nu, m_4^\nu)$  independently gives a further probability  $1/2$  of converging to its good classifier. Together, these yield a probability of  $3/32$  of converging to one of the  $4!$  many optimal classifiers for the XOR GMM.

*Remark 8.* Generically, if  $K \geq 4$ , by a similar reasoning to the above, in order to fall in the basin of attraction of the stable fixed points, it must be the case that the initialization has some four indices each of which initially belong to  $I_\mu^+, I_\mu^-, I_\nu^+, I_\nu^-$ . This is the probability that  $v_i(0)$  are positive for at least two indices, and negative for at least two indices, and then among the indices at which  $v_i(0)$  is positive, there is at least one index where  $m_i^\mu$  is positive and one where it is negative, and similarly with  $v_i(0)$  negative and  $m_i^\nu$ . Doing this combinatorial calculation out, we find that the probability of being in a good initialization is exactly the expression in (5.4). This is easily seen to go to 1 exponentially fast as  $K \rightarrow \infty$  since the initial choice of  $v_i(0)$ 's will typically have around  $K/2$  positive and  $K/2$  negative coordinates, and with exponentially high probability those will have both positive and negative  $m_i^\mu$  and  $m_i^\nu$ .

**9.5. Diffusive limit on critical submanifolds.** We now consider scaling limits of the rescaled effective dynamics in their noiseless limit, where the rescaling is about the unstable set of fixed points given by the product of two quarter circles where  $I_\mu^+ = \{1, 2\}$  and  $I_\nu^+ = \{3, 4\}$  (if  $K > 4$ , examine the fixed point in which all coordinates after the first four are in  $I_0$ ). In what follows, fix  $(a_{1,\mu}, a_{2,\mu}) \in \mathbb{R}_+^2$  with  $a_{1,\mu}^2 + a_{2,\mu}^2 = C_\alpha$ , and  $a_{3,\nu}^2 + a_{4,\nu}^2 = C_\alpha$ , and let  $\mathbf{u}_n$  be the variables of (4.2) with  $v_i, m_i^\mu, m_i^\nu$  replaced by

$$\tilde{v}_i = \begin{cases} \sqrt{N}(v_i - a_{i,\mu}) & i = 1, 2 \\ -\sqrt{N}(v_i - a_{i,\nu}) & i = 3, 4 \end{cases}$$

and

$$\tilde{m}_i^\mu = \begin{cases} \sqrt{N}(m_i^\mu - a_{i,\mu}) & i = 1, 2 \\ 0 & i = 3, 4 \end{cases}, \quad \tilde{m}_i^\nu = \begin{cases} 0 & i = 1, 2 \\ \sqrt{N}(m_i^\nu - a_{i,\nu}) & i = 3, 4 \end{cases}.$$

By the choices of  $\tilde{m}_i^\mu = 0$  and  $\tilde{m}_i^\nu = 0$ , we mean that we formally mean that we remove those variables from  $\tilde{\mathbf{u}}_n$ , and for us now  $E_K$  will be the ball of radius  $K$  in the other coordinates, and the point  $\{0\}$  for  $(\tilde{m}_i^\mu)_{i=3,4}$  and  $(\tilde{m}_i^\nu)_{i=1,2}$ .

**Proof of Proposition 5.3.** The fact that the rescaled variables  $\tilde{\mathbf{u}}_n$  satisfy the conditions of Theorem 2.3 follows as in Lemma 9.2 with the only distinction arising in the bound on (8.12), where previously we did not use the  $\delta_n^2$  factor, but is still satisfied using  $\delta_n = O(1/n)$ .

We next consider the population drift of the new variables  $\tilde{v}_i, \tilde{m}_i^\mu$  and  $\tilde{m}_i^\nu$ . If we take these variables to be in  $E_K$ , and recall the population drifts etc. in the  $\lambda = \infty$  setting from Proposition 9.1, for  $i = 1, 2$ , we have  $f_{\tilde{v}_i}$  is the  $n \rightarrow \infty$  limit of

$$\sqrt{N} \frac{m_i^\mu}{4} \sigma(-v \cdot g(m^\mu)) - \sqrt{N} \alpha v_i$$

If we then use the expansion

$$v \cdot g(m^\mu) = C_\alpha + N^{-1/2} \sum_{j=1,2} a_{j,\mu} (\tilde{v}_j + \tilde{m}_j^\mu) + O(1/n)$$

from which we obtain

$$\sigma(-v \cdot g(m^\mu)) = \sigma(-C_\alpha) + \frac{1}{\sqrt{N}} \left( \sum_{j=1,2} a_{j,\mu} (\tilde{v}_j + \tilde{m}_j^\mu) \right) (4\alpha)(1 - 4\alpha) + O(\frac{1}{n})$$

Plugging these in, and taking the  $n \rightarrow \infty$  limit we find that for  $i = 1, 2$ ,

$$f_{\tilde{v}_i} = \alpha(\tilde{v}_i - \tilde{m}_i^\mu) - a_{i,\mu}(\alpha - 4\alpha^2) \sum_{k=1,2} a_{k,\mu} (\tilde{v}_k + \tilde{m}_k^\mu).$$

By a similar reasoning, for  $i = 3, 4$ , we have

$$f_{\tilde{v}_i} = \alpha(\tilde{v}_i - \tilde{m}_i^\nu) - a_{i,\nu}(\alpha - 4\alpha^2) \sum_{k=3,4} a_{k,\nu} (\tilde{v}_k + \tilde{m}_k^\nu).$$

The claimed equations for  $f_{\tilde{m}_i^\mu}$  when  $i = 1, 2$  and  $f_{\tilde{m}_i^\nu}$  when  $i = 3, 4$  hold by analogous reasoning, and the equations for  $f_{R_{ij}^\perp}$  are evidently unaffected by the change of variables to  $\tilde{\mathbf{u}}_n$ . Regarding the population correctors, they are also unaffected (all zero) since the variables that were changed in  $\tilde{\mathbf{u}}_n$  are all linear.

It remains to compute the volatility matrix in the coordinates  $v_i, \tilde{m}_i^\mu, \tilde{m}_i^\nu$ . We first use the following expression for the matrix  $V$  when  $\lambda = \infty$ , by taking  $\lambda = \infty$  in (8.9). If  $i, j \in \{1, 2\}$ , then

$$V_{v_i, v_j} = \begin{cases} \frac{3}{16} m_i^\mu m_j^\mu \sigma(-v \cdot m^\mu)^2 & i, j \in \{1, 2\} \\ \frac{3}{16} m_i^\nu m_j^\nu \sigma(v \cdot m^\nu)^2 & i, j \in \{3, 4\} \end{cases}$$

and if  $i \in \{1, 2\}$  and  $j \in \{3, 4\}$ , then

$$V_{v_i, v_j} = -\frac{1}{16} m_i^\mu m_j^\nu \sigma(-v \cdot m^\mu) \sigma(v \cdot m^\nu)$$

When considering  $\Sigma_{v_i, v_j}$  we multiply this by  $N$  coming from  $\tilde{J}$  and  $\tilde{J}^T$ , but also multiply by  $\delta = 1/N$ , so that taking the limit as  $n \rightarrow \infty$ , we get

$$\tilde{\Sigma}_{v_i, v_j} = \begin{cases} 3\alpha^2 a_{i,\mu} a_{j,\mu} & i, j \in \{1, 2\} \\ 3\alpha^2 a_{i,\nu} a_{j,\nu} & i, j \in \{3, 4\} \\ -3\alpha^2 a_{i,\mu} a_{j,\nu} & i \in \{1, 2\}, j \in \{3, 4\} \end{cases}.$$

By a similar reasoning, if  $i, j \in \{1, 2\}$ , then

$$\begin{aligned} V_{v_i, W_j} \cdot \mu &= \frac{3}{16} v_j m_i^\mu \sigma(-v \cdot m^\mu)^2 & i, j \in \{1, 2\} \\ V_{v_i, W_j} \cdot \nu &= \frac{3}{16} v_j m_i^\nu \sigma(v \cdot m^\nu)^2 & i, j \in \{3, 4\} \end{aligned}$$

and if  $i \in \{1, 2\}$  and  $j \in \{3, 4\}$ , then

$$V_{v_i, W_j} \cdot \nu = -\frac{1}{16} v_j m_i^\mu \sigma(-v \cdot m^\mu) \sigma(v \cdot m^\nu).$$

Taking the limit as  $n \rightarrow \infty$ , we again recover the claimed limiting diffusion matrix, and similar calculations yield the same for  $\Sigma_{\tilde{m}_i^\mu, \tilde{m}_j^\mu}$ ,  $\Sigma_{\tilde{m}_i^\nu, \tilde{m}_j^\nu}$  and  $\Sigma_{\tilde{m}_i^\mu, \tilde{m}_j^\nu}$ , concluding the proof.  $\square$

## 10. PROOFS OF TECHNICAL LEMMAS FOR GAUSSIAN MIXTURES

In this section, we establish the technical bounds on Gaussian moments in Lemmas 8.2–8.3.

**Proof of Lemma 8.2.** For the first bound, let  $Z \sim \mathcal{N}(0, I)$  and consider

$$\mathbb{E}[|X \cdot w|^8] = \frac{1}{2} \mathbb{E}[(w \cdot \mu + \lambda^{-1/2} w \cdot Z)^8] + \frac{1}{2} \mathbb{E}[(-w \cdot \mu + \lambda^{-1/2} w \cdot Z)^8].$$

The quantities in the expectations are at most some universal constant times  $(w \cdot \mu)^8 + \lambda^{-4}(w \cdot Z)^8$ . To bound the expectation of the second term here, notice that  $w \cdot Z$  is distributed as  $\mathcal{N}(0, \|w\|^2)$  implying the desired.

The bound on  $\mathbf{A}_i$  goes as follows. Evidently it suffices to let  $X_\mu = \mu + \lambda^{-1/2}Z$  for  $Z \sim \mathcal{N}(0, I)$ , and prove the bound on the norm of

$$\mathbb{E}[X_\mu \mathbf{1}_{W_i \cdot X_\mu \geq 0} (-1 + \sigma(g(WX_\mu)))] = \mathbb{E}[(\mu + \lambda^{-1/2}Z) \mathbf{1}_{W_i \cdot X_\mu \geq 0} (-1 + \sigma(g(WX_\mu)))].$$

Now decompose  $Z$  as  $Z_\mu \mu + Z_{1,\perp} W_1^\perp + Z_{2,\perp} W_2^\perp + Z_3$ , where  $Z_\mu \sim \mathcal{N}(0, 1)$  is independent of  $(Z_{1,\perp}, Z_{2,\perp})$  which is distributed as  $\mathcal{N}(0, A)$  with  $A$  given by (8.1), which is independent of  $Z_3$  distributed as a standard Gaussian vector orthogonal to the subspace spanned by  $(\mu, W_1^\perp, W_2^\perp)$ . By independence of  $Z_3$  from the indicator and the argument of the sigmoid, all those terms contribute nothing to the expectation, and therefore,

$$\|\mathbf{A}_i\|^2 \leq \sum_{w \in \{\mu, W_1^\perp, W_2^\perp\}} \mathbb{E}[(X \cdot w)^2 \mathbf{1}_{W_i \cdot X \geq 0} (-y + \sigma(g(WX)))] \leq (1 + R_{11}^\perp + R_{22}^\perp)(1 + \lambda^{-1}).$$

Here, we used the first inequality of the lemma. This yields the desired.  $\square$

**Proof of Lemma 8.3.** The proof of (8.10) is easily seen by rewriting the probability in question as

$$\mathbb{P}(W_i \cdot X_\mu < 0) = \mathbb{P}(\mathcal{N}(0, \lambda^{-1}) < -m_i(m_i^2 + R_{ii}^\perp)^{-1/2}) = e^{-m_i^2 \lambda/2(m_i^2 + R_{ii}^\perp)},$$

so that as long as  $m_i > 0$  this goes to zero as  $\lambda \rightarrow \infty$ .

We turn to (8.11). Consider

$$\begin{aligned}\mathbb{E}[|\sigma(v \cdot g(WX_\mu)) - \sigma(v \cdot g(m))|] &\leq \mathbb{E}[|e^{v \cdot g(WX_\mu)} - e^{v \cdot g(m)}|] \\ &\leq \mathbb{E}[|e^{v_1 g(W_1 \cdot X_\mu)} e^{v_2 g(W_2 \cdot X_\mu)} - e^{v_1 g(m_1)} e^{v_2 g(m_2)}|].\end{aligned}$$

This in turn is bounded by

$$\mathbb{E}[e^{v_2 g(W_2 X_\mu)} |e^{v_1 g(W_1 X_\mu)} - e^{v_1 g(m_1)}|] + e^{v_1 g(m_1)} \mathbb{E}[|e^{v_2 g(W_2 X_\mu)} - e^{v_2 g(m_2)}|]. \quad (10.1)$$

Applying Cauchy–Schwarz to the first term, it suffices to establish the following bounds

$$\mathbb{E}[e^{2v_i g(W_i X_\mu)}] \leq C, \quad \text{and} \quad \lim_{\lambda \rightarrow \infty} \mathbb{E}[(e^{v_i g(W_i X_\mu)} - e^{v_i g(m_i)})^2] = 0.$$

To demonstrate the first of these inequalities, notice that

$$\mathbb{E}[e^{2v_i g(W_i X_\mu)}] \leq \mathbb{E}[e^{2v_i |W_i X_\mu|}] \leq C.$$

uniformly over  $\lambda$ , per Fact 8.1. For the second desired bound, expand  $e^{v_i g(W_i \cdot X_\mu)} - e^{v_i g(m_i)}$  as

$$(e^{v_i(W_i \cdot X_\mu)} \mathbf{1}_{W_i \cdot X_\mu \geq 0} - e^{v_i(W_i \cdot X_\mu)} \mathbf{1}_{m_i \geq 0}) + (e^{v_i(W_i \cdot X_\mu)} \mathbf{1}_{m_i \geq 0} - e^{v_i m_i} \mathbf{1}_{m_i \geq 0}).$$

It suffices to show the expectation of the square of each of these goes to zero as  $\lambda \rightarrow \infty$ . First,

$$\mathbb{E}[(e^{v_i(W_i \cdot X_\mu)} \mathbf{1}_{W_i \cdot X_\mu \geq 0} - e^{v_i(W_i \cdot X_\mu)} \mathbf{1}_{m_i \geq 0})^2] \leq (1 \vee e^{v_i(W_i \cdot X_\mu)}) \mathbb{E}[\mathbf{1}_{W_i \cdot X_\mu \geq 0} - \mathbf{1}_{m_i \geq 0}].$$

If  $m_i \neq 0$ , the expectation on the right goes to zero by (8.10). Second,

$$\mathbb{E}[(e^{v_i(W_i \cdot X_\mu)} \mathbf{1}_{m_i \geq 0} - e^{v_i m_i} \mathbf{1}_{m_i \geq 0})^2] \leq \mathbb{E}[(e^{v_i(W_i \cdot X_\mu)} - e^{v_i m_i})^2 \mathbf{1}_{m_i \geq 0}].$$

When  $m_i < 0$ , this is evidently zero; when  $m_i > 0$ , if  $G_\lambda \sim \mathcal{N}(0, I/\lambda)$ , this is

$$e^{2v_i m_i} \mathbb{E}[(e^{v_i(W_i \cdot G_\lambda)} - 1)^2].$$

which goes to zero as  $O(\lambda^{-1})$  when  $\lambda \rightarrow \infty$ , by the explicit formula for the moment generating function of the Gaussian  $W_i \cdot G_\lambda$ , whose variance is  $(m_i^2 + R_{ii}^\perp)\lambda^{-1}$ .  $\square$

**Acknowledgements.** The authors thank the anonymous referees for their useful comments and suggestions. The authors thank F. Krzakala, L. Zdeborova, and B. Loureiro for interesting conversations and suggestions, especially suggesting we investigate the role of overparametrization in the XOR GMM. The authors thank M. Sellke for pointing out the relationship to the lottery ticket hypothesis. The authors also thank M. Glasgow for a careful reading and helpful suggestions. R.G. acknowledges the support of NSF DMS-2246780 and the Miller Institute for Basic Research in Science. A.J. acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada Research Chairs programme. Cette recherche a été entreprise grâce, en partie, au soutien financier du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [RGPIN-2020-04597, DGECR-2020-00199], et du Programme des chaires de recherche du Canada.

## REFERENCES

- [1] Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suárez, and Yi Zhang. Learning threshold neurons via the "edge of stability", 2022.
- [2] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale CLT. In *Conference on Learning Theory*, pages 115–137. PMLR, 2019.
- [3] Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*. Wiley-Interscience [John Wiley & Sons], New York, 1962.
- [4] Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv preprint arXiv:1906.00193*, 2019.
- [5] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 948–1024. PMLR, 17–23 Jul 2022.
- [6] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.
- [7] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.
- [8] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Algorithmic thresholds for tensor PCA. *Annals of Probability*, 48(4):2052–2087, 2020.
- [9] Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Bounding flows for spherical spin glass dynamics. *Communications in Mathematical Physics*, 373(3):1011–1048, 2020.
- [10] Gérard Ben Arous, Song Mei, Andrea Montanari, and Mihai Nica. The landscape of the spiked tensor model. *Comm. Pure Appl. Math.*, 72(11):2282–2330, 2019.
- [11] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 1–68. Springer, Berlin, 1999.
- [12] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.
- [13] Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.
- [14] Léon Bottou. *On-Line Learning and Stochastic Approximations*. Cambridge University Press, USA, 1999.
- [15] Léon Bottou and Yan Le Cun. Large scale online learning. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 217–224. MIT Press, 2004.
- [16] Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral. The largest eigenvalues of finite rank deformation of large wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability*, pages 1–47, 2009.
- [17] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.
- [18] Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. Stochastic gradient and Langevin processes. In *International Conference on Machine Learning*, pages 1810–1819. PMLR, 2020.
- [19] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.
- [20] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.
- [21] A Crisanti, H Horner, and H-J Sommers. The spherical p-spin interaction spin-glass model. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, 1993.
- [22] Leticia F. Cugliandolo and Jorge Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Phys. Rev. Lett.*, 71:173–176, Jul 1993.
- [23] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023.
- [24] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Ann. Statist.*, 48(3):1348–1382, 06 2020.
- [25] Marie Duflo. *Algorithmes stochastiques*, volume 23 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1996.
- [26] Paul Dupuis and Harold J Kushner. Stochastic approximation and large deviations: Upper bounds and w.p.1 convergence. *SIAM Journal on Control and Optimization*, 27(5):1108–1135, 1989.

- [27] Ahmed El Alaoui, Florent Krzakala, and Michael Jordan. Fundamental limits of detection in the spiked Wigner model. *The Annals of Statistics*, 48(2):863–885, 2020.
- [28] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986. Characterization and convergence.
- [29] Delphine Féral and Sandrine Péché. The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in mathematical physics*, 272(1):185–228, 2007.
- [30] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.
- [31] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 797–842, Paris, France, 03–06 Jul 2015. PMLR.
- [32] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.
- [33] Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1579–1613, Phoenix, USA, 25–28 Jun 2019. PMLR.
- [34] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191. ACM, 2016.
- [35] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015.
- [36] Aukosh Jagannath, Patrick Lopatto, and Léo Miolane. Statistical thresholds for tensor PCA. *Ann. Appl. Probab.*, 30(4):1910–1933, 2020.
- [37] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.
- [38] Chiheon Kim, Afonso S Bandeira, and Michel X Goemans. Community detection in hypergraphs, spiked tensor models, and sum-of-squares. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 124–128. IEEE, 2017.
- [39] Harold J Kushner. Asymptotic behavior of stochastic approximation and large deviations. *IEEE transactions on automatic control*, 29(11):984–990, 1984.
- [40] Thibault Lesieur, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 511–515. IEEE, 2017.
- [41] Chris Junchi Li, Mengdi Wang, Han Liu, and Tong Zhang. Diffusion approximations for online principal component estimation and global convergence. *Advances in Neural Information Processing Systems*, 30, 2017.
- [42] Chris Junchi Li, Zhaoran Wang, and Han Liu. Online ICA: Understanding global dynamics of nonconvex optimization via diffusion processes. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4967–4975. Curran Associates, Inc., 2016.
- [43] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.
- [44] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). *Advances in Neural Information Processing Systems*, 34, 2021.
- [45] Tengyuan Liang, Subhabrata Sen, and Pragya Sur. High-dimensional asymptotics of Langevin dynamics in spiked matrix models. *arXiv preprint arXiv:2204.04476*, 2022.
- [46] Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, AC-22(4):551–575, 1977.
- [47] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate Bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.
- [48] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Marvels and pitfalls of the Langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020.
- [49] D. L. McLeish. Functional and random central limit theorems for the Robbins-Munro process. *Journal of Applied Probability*, 13(1), 1976.
- [50] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018.

- [51] Marvin Minsky and Seymour A Papert. *Perceptrons, Reissue of the 1988 Expanded Edition with a new foreword by Léon Bottou: An Introduction to Computational Geometry*. MIT press, 2017.
- [52] Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors. In *Advances in Neural Information Processing Systems*, pages 217–225, 2015.
- [53] Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, Cambridge, MA, USA, 2014. MIT Press.
- [54] Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5012–5021. PMLR, 09–15 Jun 2019.
- [55] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pages 3548–3626. PMLR, 2021.
- [56] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, pages 1617–1642, 2007.
- [57] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira. Statistical limits of spiked tensor models. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56(1):230–264, 2020.
- [58] Amelia Perry, Alexander S Wein, Afonso S Bandeira, and Ankur Moitra. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.
- [59] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.
- [60] Maria Refinetti, Sebastian Goldt, Florent Krzakala, and Lenka Zdeborová. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *International Conference on Machine Learning*, pages 8936–8947. PMLR, 2021.
- [61] Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.
- [62] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.
- [63] Grant M Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.
- [64] David Saad and Sara Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. *Advances in neural information processing systems*, 8, 1995.
- [65] David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995.
- [66] Levent Sagun, Utku Evci, V. Ugur Güney, Yann N. Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *CoRR*, abs/1706.04454, 2017.
- [67] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, and Lenka Zdeborová. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. *Advances in Neural Information Processing Systems*, 32, 2019.
- [68] Ohad Shamir. Convergence of stochastic gradient descent for PCA. In *International Conference on Machine Learning*, pages 257–265. PMLR, 2016.
- [69] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.
- [70] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. Reprint of the 1997 edition.
- [71] Yan Shuo Tan and Roman Vershynin. Phase retrieval via randomized Kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, 8(1):97–123, 04 2018.
- [72] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.
- [73] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *arXiv preprint arXiv:2202.00293*, 2022.
- [74] Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press (to appear), 2018.
- [75] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- [76] Chuang Wang, Jonathan Mattingly, and Yue Lu. Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA. *arXiv preprint arXiv:1712.04332*, 2017.

- [77] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient Langevin dynamics. volume 65 of *Proceedings of Machine Learning Research*, pages 1980–2022, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.
- [78] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023.

(Gérard Ben Arous) COURANT INSTITUTE, NEW YORK UNIVERSITY  
*Email address:* `benarous@cims.nyu.edu`

(Reza Gheissari) DEPARTMENT OF MATHEMATICS, NORTHWESTERN UNIVERSITY  
*Email address:* `gheissari@northwestern.edu`

(Aukosh Jagannath) DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE, DEPARTMENT OF APPLIED MATHEMATICS, AND CHERITON SCHOOL OF COMPUTER SCIENCE, UNIVERSITY OF WATERLOO  
*Email address:* `a.jagannath@uwaterloo.ca`