Journal of Machine Learning Research 22 (2021) 1-51 Submitted 11/20; Revised 4/21; Published 4/21

Online stochastic gradient descent on non-convex losses from high-dimensional inference

Gerard Ben Arous BENAROUS@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA

Reza Gheissari GHEISSARI@BERKELEY.EDU
Departments of Statistics and EECS
University of California
Berkeley, CA, USA

Aukosh Jagannath A.JAGANNATH@UWATERLOO.CA
Departments of Statistics and Actuarial Science and Applied Mathematics
University of Waterloo
Waterloo, ON, Canada

Editor: Gabor Lugosi

Abstract

Stochastic gradient descent (SGD) is a popular algorithm for optimization problems arising in high-dimensional inference tasks. Here one produces an estimator of an unknown parameter from independent samples of data by iteratively optimizing a loss function. This loss function is random and often non-convex. We study the performance of the simplest version of SGD, namely online SGD, from a random start in the setting where the parameter space is high-dimensional.

We develop nearly sharp thresholds for the number of samples needed for consistent estimation as one varies the dimension. Our thresholds depend only on an intrinsic property of the population loss which we call the information exponent. In particular, our results do not assume uniform control on the loss itself, such as convexity or uniform derivative bounds. The thresholds we obtain are polynomial in the dimension and the precise exponent depends explicitly on the information exponent. As a consequence of our results, we find that except for the simplest tasks, almost all of the data is used simply in the initial search phase to obtain non-trivial correlation with the ground truth. Upon attaining non-trivial correlation, the descent is rapid and exhibits law of large numbers type behavior.

We illustrate our approach by applying it to a wide set of inference tasks such as phase retrieval, and parameter estimation for generalized linear models, online PCA, and spiked tensor models, as well as to supervised learning for single-layer networks with general activation functions.

Keywords: stochastic gradient descent, parameter estimation, non-convex optimization, supervised learning, generalized linear models, tensor PCA

1. Introduction

Stochastic gradient descent (SGD) and its many variants are the algorithms of choice for many hard optimization problems encountered in machine learning and data science. Since its introduction in the 1950s (Robbins and Monro, 1951), SGD has been abundantly studied. While first analyzed in fixed dimensions (McLeish, 1976; Benveniste et al., 1990; Ben-David et al., 1995; Duflo, 1996; Bottou, 1999; Benaïm, 1999), its analysis in high-dimensional settings has recently become the subject of intense interest, both from a theoretical point of view (see, e.g., Needell et al., 2014; Ge et al., 2015; Li et al., 2016; Wang and Lu, 2016; Wang et al., 2017; Mandt et al., 2017; Harvey et al., 2019; Tan and Vershynin, 2019) and a practical one (see, e.g., LeCun et al., 1998; Bishop, 2006; Goodfellow et al., 2016).

The evolution of SGD is often heuristically viewed as having two phases (Bottou, 2003; Bottou and Le Cun, 2004; Mandt et al., 2017): an initial "search" phase and a final "descent" phase. In the search phase, one often thinks of the algorithm as wandering in a non-convex landscape. In the descent phase, however, one views the algorithm as being in an effective trust region and descending quickly to a local minimum. In this latter phase, one expects the algorithm to be a good approximation to the gradient descent for the population loss, whereas in the former the quality of this approximation should be quite low.

In the fixed dimension setting one can avoid an analysis of the search phase by neglecting an initial burn-in period, and assuming that the algorithm starts in the descent phase. This reasoning is sometimes used (Bottou, 1999) as a motivation for convexity or quasi-convexity assumptions in the analysis of stochastic gradient descent for which there now is a large literature (Bottou, 1999; Bottou and Le Cun, 2004; Needell et al., 2014; Needell and Ward, 2017; Harvey et al., 2019; Dieuleveut et al., 2020). Furthermore, such a burn-in time perspective is (in a sense) implicit in approaches based on the ODE method of Ljung (1977) (see, e.g., the notion of asymptotic pseudotrajectories in Benaïm, 1999).

In the high-dimensional setting, however, it is less clear that one can ignore the burn-in period, as the dependence of its length on the dimension is poorly understood. Furthermore, in many problems of interest, random initializations are strongly concentrated in a neighborhood of an uninformative critical point of the population loss. Nevertheless, there is a wealth of numerical experiments showing that SGD may perform well in these regimes. This suggests that in high dimensions, the algorithm is able to recover from a random start on computationally feasible timescales, even in regimes where the behavior from a random start is quite different from that started in a trust region.

There has been a tremendous effort in recent years to understand these issues and to develop general frameworks for understanding the convergence of SGD for non-convex loss functions in the high-dimensional setting. While many results still require convexity or quasi-convexity, several only require uniform control on derivatives of the empirical risk—such as L-smoothness (namely a uniform control on the Lipschitz constant of the gradient), and possibly uniform Lipschitzness—on a bounded domain.[1] These latter approaches typically study Monte Carlo analogues of SGD or develop a stochastic differential equation (SDE) approximation to it and control the rate of convergence of the corresponding processes to its invariant measure. Here one can obtain bounds that depend polynomially on the dimension and exponentially on quantities like L and the radius of the domain R. See

1. When working in unbounded domains it is common to add a "growth at infinity" assumption.

for example Raginsky et al. (2017); Zhang et al. (2017); Cheng et al. (2018, 2020); Ma et al. (2019) for a representative (but non-exhaustive) collection of works in this direction.[2]

In many basic problems of high-dimensional statistics, however, the assumptions of dimensionless bounds on both the domain and the derivative are unrealistic, even in average case, due to the concentration of measure phenomenon. For illustrative purposes, consider the simplest example of linear regression with Gaussian covariates. Here a simple calculation shows that the usual empirical risk—the least squares risk—is L-smooth and K-Lipschitz on the unit ball for L and K diverging linearly in N. Of course it is possible to re-scale the loss so that $L = O(1)$, but such a re-scaling will dramatically change the invariant measure of natural Monte Carlo or SDE approximations and render it uninformative; the aforementioned bounds on the convergence rate are then no longer relevant for the estimation task. See Section 2.3 for more discussion on this, where we also consider the case of non-Gaussian covariates and errors.

Given the discussion above, we are led to the following questions which motivate our work here:

1. Can one prove sample complexity bounds for SGD that are polynomial in the dimension for natural estimation tasks?

2. For such tasks, what fraction of data is used/time is spent in the search phase as opposed to the descent phase?

3. How do these answers change as one varies the loss (or activation function)? Can this be answered using quantities that depend only on the *population* loss?

To understand these questions, we restrict our study to the simple but important high-dimensional setting of rank-one parameter estimation, though we believe that it can be extended to more general settings. This setting covers many important and classical problems such as:

- supervised learning for a single-layer network, and phase retrieval: §2.1
- generalized linear models (GLMs) (e.g., linear and logistic regression): §2.2–2.3
- online PCA and spiked matrix and tensor models: §2.4–2.5
- two-component mixtures of Gaussians: §2.6.

We treat all of these examples in detail in Section 2. For these examples, dimension-free smoothness assumptions do not apply just as in the case of linear regression described above. Furthermore for most of them, the loss function is non-convex and one expects exponentially many critical points, especially in the high entropy region (Ben Arous et al., 2019; Ros et al., 2019; Maillard et al., 2020). In spite of these issues, we prove that online SGD succeeds at these estimation tasks with polynomially many samples in the dimension.

Our main contribution is a classification of these rank-one parameter estimation problems on the sphere. This classification is defined solely by a geometric quantity, called the

2. We note here that some of these bounds depend polynomially on inverses of spectral quantities, such as the spectral gap of the Langevin operator or Cheeger constants. In general settings, these quantities often grow exponentially in the dimension.

information exponent (see Definition 1.2), which captures the non-linearity of the *population* loss near the region of maximal entropy (here, the equator of the sphere). More precisely, the information exponent is the degree of the first non-zero term in the Taylor expansion of the population loss about the equator. We study the dependence of the amount of data needed (i.e., the *sample complexity*) for recovery of the parameter using online SGD on the information exponent. We prove thresholds for the sample complexity for online SGD to exit the search phase which are linear, quasi-linear, and polynomial in the dimension depending on whether the information exponent is 1, 2, or at least 3 respectively (Theorem 1.3). Furthermore, we prove sharpness of these thresholds up to a logarithmic factor in the dimension, showing that the three different regimes in this classification are indeed distinct (Theorem 1.4). Finally, we show that once the algorithm is in the descent phase, there is a law of large numbers for the trajectory of the distance to the ground truth about that for gradient descent on the population loss (Theorem 3.2). In particular, the final descent phase is at most linear in time for all such estimation problems and one asymptotically recovers the parameter.

As a consequence of our classification, we find that when the information exponent is at least 2, essentially all of the data is used in the initial "search phase": the ratio of the amount of data used in the descent phase (linear in the dimension) to the amount used in the search phase (quasilinear or polynomial in the dimension) vanishes as the dimension tends to infinity. Put simply, the main difficulty in these problems is leaving the high entropy region of the initialization; once one has left this region, the algorithm's performance is fast in *all* of these problems. This matches the heuristic picture described above. For an illustration of this discussion, see Figures 2.1–2.2 for numerical experiments in the supervised learning setting.

Our classification yields nearly tight sample complexity bounds for the performance of SGD from a random start in a broad range of classification tasks. In particular, the setting $k \leq 2$ covers a broad range of estimation tasks for which the problem of sharp sample complexity bounds have received a tremendous amount of attention such as phase retrieval, supervised learning with commonly used activation functions such as ReLu and sigmoid, Gaussian mixture models, and spiked matrix models. The regime $k \geq 3$ has, to our knowledge, received less attention in the literature with one notable exception, namely spiked tensor models. That being said, we find fascinating phenomena in these problems. For example, in the supervised learning setting, we find that minor modifications of the activation function can dramatically change the sample complexity needed for consistent estimation with SGD. See, e.g., Fig. 2.3.

Importantly, our approach does not require almost sure or high probability assumptions on the geometry of the loss landscape, such as convexity or strictness and separation of saddle points. Furthermore, we make no assumptions on the Hessian of the loss. Indeed, in many of our applications the Hessian of the loss where the algorithm is initialized will have many positive and negative eigenvalues simultaneously. Instead we make a scaling assumption on the sample-wise error for the loss. More precisely, we assume a scaling in N for the low moments of the norm of its gradient and the variance of the directional derivative in the direction of the parameter to be inferred. We note that the assumption of L-smoothness falls into our setting, but our assumption is less restrictive. For a precise definition of our assumption and a comparison see Definition 1.1 below. We expect that

the classification we introduce has broader implications for efficient estimation thresholds than the setting we consider here, such as finite rank estimation, or offline algorithms with possibilities of reuse, batching, and overparametrization. Furthermore, while we restrict to the sphere, we expect a similar classification to hold in full space with e.g., an ℓ^2 penalty, but leave this to future investigation.

1.1 Formalizing "search" vs. "descent": weak and strong recovery

One of our goals in this work is to understand the dimension dependence of the relative proportion of time spent by SGD in the search phase as opposed to the descent phase. As such, we need a formal definition of the search phase. To this end, we focus on the simple setting where the population loss is a (possibly non-linear) function of the distance to the parameter. Furthermore, to make matters particularly transparent, and for simplicity of working in a bounded domain, we will assume that the norm of the parameter is known. Note that in some settings, fixing the norm amounts to assuming a fixed variance (Johnstone, 2001; Sur and Candès, 2019).

More precisely, suppose that we are given a parametric family of distributions, $(P_x)_{x\in\mathbb{S}^{N-1}}$, of $\mathbb{R}^D$-valued random variables, parameterized by the unit sphere, $\mathbb{S}^{N-1}$, and $M = \alpha N$ i.i.d. samples (Y^ℓ) from one of these distributions, $\mathbb{P}_N = P_{\theta_N}$, which we call the *data distribution*. Our goal is to estimate θ_N given these samples, via online SGD with loss function $\mathcal{L}_N : \mathbb{S}^{N-1} \times \mathbb{R}^D \to \mathbb{R}$. We study here the case where the *population loss* is of the form

$$\Phi_N(x) := \mathbb{E}_N[\mathcal{L}_N] = \phi(m_N(x)) \qquad \text{where} \qquad m_N(x) = x \cdot \theta_N \,, \tag{1.1}$$

for some $\phi : [-1, 1] \to \mathbb{R}$ (here $\cdot$ denotes the Euclidean inner product on $\mathbb{R}^N$). We call $m_N(x)$ the *correlation* of x with θ_N. We often also refer to m_N as the latitude, and call the set $m_N(x) \approx 0$ the *equator* of the sphere.

In order to formalize the notion of exiting the "search phase", we recall here the notion of "weak recovery", i.e., achieving macroscopic correlation with θ_N. We say that a sequence of estimators $\hat{\theta}_N \in \mathbb{S}^{N-1}$ *weakly recovers* the parameter θ_N if for some $\eta > 0$,

$$\lim_{N\to\infty} P\Big(m_N(\hat{\theta}_N) \geq \eta\Big) = 1 \,.$$

As $\|m_N\|_\infty \leq 1$, this definition is equivalent to the existence of $\eta > 0$ such that

$$\varliminf_{N\to\infty} \mathbb{E}[m_N(\hat{\theta}_N)] \geq \eta \,;$$

this latter formulation was used in, e.g., Mondelli and Montanari (2018). To understand the scaling here, i.e., $\hat{\theta}_N \cdot \theta_N = \Theta(1)$, recall the basic fact from high-dimensional probability (Vershynin, 2019), that if $\hat{\theta}_N$ were drawn uniformly at random, then $\hat{\theta}_N \cdot \theta_N \simeq N^{-1/2}$ and that the probability of weak recovery with a random choice is exponentially small in N. In the context we consider here, attaining weak recovery corresponds to exiting the search phase. On the other hand, our final goal is to understand *consistent estimation* or *strong recovery* which in this setting amounts to showing that $m_N(\hat{\theta}_N) \to 1$ in probability (or equivalently in L^p-norm for $p \geq 1$).

1.2 Algorithm and assumptions

As our parameter space is spherical, we will consider a spherical version of online SGD which is defined as follows. Let X_t denote the output of the algorithm at time t, and let $\delta > 0$ denote a step size parameter. The sequence of outputs of the algorithm are then given by the following procedure:

$$\begin{cases} X_0 = & x_0 \\ \tilde{X}_t = & X_{t-1} - \frac{\delta}{N} \nabla \mathcal{L}_N(X_{t-1}; Y^t) \ , \\ X_t = & \frac{\tilde{X}_t}{||\tilde{X}_t||} \end{cases} \tag{1.2}$$

where the initial point x_0 is possibly random, $x_0 \sim \mu \in \mathcal{M}_1(\mathbb{S}^{N-1})$ and where ∇ denotes the spherical gradient, i.e., for a function $f : \mathbb{S}^{N-1} \to \mathbb{R}$,

$$\nabla f = Df - \frac{\partial f}{\partial r} \frac{\partial}{\partial r} \, ,$$

where D is the derivative in $\mathbb{R}^N$ and $\frac{\partial}{\partial r}$ is the partial derivative in the radial direction, again in $\mathbb{R}^N$. In the online setting, we terminate the algorithm after it has run for M steps (though in principle one could terminate earlier). We take, as our estimator, the output of this algorithm. In order for this algorithm to be well-defined, we assume that the loss is almost surely differentiable in the parameter for all $x \in \mathbb{S}^{N-1}$.

As we are studying a first-order method, it is also natural to expect that the output of gradient flow is a Fisher consistent estimator for all initial data with positive correlation, meaning that it is consistent when evaluated on the *population* loss. To this end, we say that **Assumption A** holds if ϕ is differentiable and ϕ' is strictly negative on $(0, 1)$. Observe that Assumption A holds if and only if gradient flow for the population loss eventually produces a consistent estimator, when started anywhere on the upper half-sphere $\{x : m_N(x) > 0\}$[3]. The reason we restrict this assumption to $(0, 1)$ and the upper half-sphere is to include problems where the parameter is only recovered up to a sign due to an inherent symmetry in the task. We emphasize here that Assumption A is a property only of the population loss, and does not imply convexity of the loss $\mathcal{L}_N$. (We note however, that this is an assumption on the (unknown) data distribution, and cannot be empirically verified.)

In order to investigate the performance of SGD in a regime that captures a broad range of high-dimensional inference tasks, we need to choose a scaling for the fluctuations of the loss. These fluctuations are captured by the *sample-wise error*, defined by

$$H_N^\ell(x) := \mathcal{L}_N(x; Y^\ell) - \Phi_N(x) \, .$$

We seek a scaling regime which does not suffer from the issues described in the introduction. To this end, we work under the following assumption which is satisfied by the loss functions of many natural high-dimensional problems.

Definition 1.1. *For a sequence of data distributions and losses $(\mathbb{P}_N, \mathcal{L}_N)$, we say that **Assumption B** holds if there exists $C_1, \iota > 0$ such that the following two moment bounds hold for all N:*

3. Observe that the gradient flow on Φ reduces to its projection on m_N, which solves an autonomous ODE; Assumption A holds if and only if the window $(0, 1)$ is in the absorbing set of a minimum of ϕ at $m_N = 1$.

(a) We have that

$$\sup_{x\in\mathbb{S}^{N-1}} \mathbb{E}\big[(\nabla H_N(x)\cdot\theta_N)^2\big] \le C_1\,, \tag{1.3}$$

(b) and that,

$$\sup_{x\in\mathbb{S}^{N-1}} \mathbb{E}[||\nabla H_N(x)||^{4+\iota}] < C_1 N^{\frac{4+\iota}{2}}. \tag{1.4}$$

This assumption captures the scaling regimes commonly used for high-dimensional analyses of statistical problems throughout the literature, see, e.g., Johnstone (2001); Wainwright (2009); Richard and Montanari (2014); Candès et al. (2015). For an in-depth discussion, see Section 2, where we show that a broad class of statistical models satisfy Assumption B. Observe that the scaling relation between (a) and (b) is tight when ∇H_N is an i.i.d. sub-Gaussian vector. On the other hand, note that if $\mathcal{L}_N$ is L-smooth on the unit sphere for some fixed $L = O(1)$, then Assumption B holds, with an $O(1)$ bound instead of an $O(N^{2+\iota/2})$ bound in (1.4). In particular, Assumption B applies more generally than $O(1)$-smoothness.

1.3 Main results

In this paper, we show that a key quantity governing the performance of online SGD is the following, which we call the information exponent for a population loss.

Definition 1.2. *We say that a population loss Φ_N has **information exponent** k if $\phi \in C^{k+1}([-1,1])$ and there exist $C, c > 0$ such that*

$$\begin{cases} \frac{d^\ell \phi}{dm^\ell}(0) = 0 & 1 \le \ell < k \\ \frac{d^k \phi}{dm^k}(0) \le -c < 0 & \\ ||\frac{d^{k+1}\phi}{dm^{k+1}}(m)||_\infty \le C & \end{cases}. \tag{1.5}$$

We compute the information exponent for a broad class of examples in Section 2. (If the first non-zero derivative is instead positive, then Assumption A cannot hold, as $\phi'(\varepsilon)$ will be positive for $\varepsilon > 0$ sufficiently small.)

Our first result upper bounds the sample complexity for consistent estimation. In the sequel, let

$$\alpha_c(N,k) = \begin{cases} 1 & k = 1 \\ \log N & k = 2 \\ N^{k-2} & k \ge 3 \end{cases} \tag{1.6}$$

and say that a sequence $x_n \ll y_n$ if $x_n/y_n \to 0$. For concreteness, we state our result in the case of random initialization, namely we take μ_N to be the uniform measure conditioned on the upper half sphere $\{m(x) \ge 0\}$. (This conditioning is without loss of generality up to a probability $1/2$ event and is introduced to avoid obvious symmetry issues: see Remark 1.8.) We then have the following.

Theorem 1.3. *Suppose that Assumptions A and B hold and that the population loss has information exponent k. Let $M = \alpha_N N$ with α_N growing at most polynomially in N. If (α_N, δ_N) satisfy $\alpha_N^{-1} \ll \delta_N \ll \alpha_N^{-1/2}$ and*

- $(k=1:)$ $\alpha_N \gg \alpha_c(N,1)$
- $(k=2:)$ $\alpha_N \gg \alpha_c(N,2)\cdot \log N$
- $(k\geq 3:)$ $\alpha_N \gg \alpha_c(N,k)\cdot (\log N)^2$

then online SGD with step size parameter δ_N started from $X_0\sim\mu_N$, will have

$$m_N(X_M)\to 1\,, \qquad \text{in probability, and in } L^p \text{ for every } p\geq 1\,.$$

Our second result is the corresponding lower bound on the sample complexity required for exiting the search phase with a given information exponent.

Theorem 1.4. *Suppose that Assumptions A and B hold and that the population loss has information exponent $k\geq 1$. If (α_N,δ_N) are such that*

- **$(k=1,2)$:** $\alpha_N \ll \alpha_c(N,k)$ *and* $\delta_N = O(1)$
- **$(k\geq 3)$:** $\alpha_N \ll \alpha_c(N,k)$ *and* $\delta_N = O(\alpha_N^{-1/2})$,

then the online SGD with step size parameter δ_N, started from $X_0\sim\mu_N$, will have

$$\sup_{t\leq M}|m_N(X_t)|\to 0\,, \qquad \text{in probability, and in } L^p \text{ for every } p\geq 1\,.$$

Let us pause to comment on the interpretation of these results. The first result states that the sample complexity of consistent estimation for a problem with finite information exponent is always at most polynomial, and provides a precise scaling for this polynomial, $\alpha_c(N,k)$, as a function of the information exponent k. The second result says that the thresholds $\alpha_c(N,k)$ are optimal up to $O((\log N)^2)$. We expect, in fact, that the thresholds $\alpha_c(N,k)$—without additional logarithmic factors—are sharp; see Section 1.4 for more on this. Observe that the second result in particular implies the algorithm can *only* exit the search phase when the number of samples is at least $\alpha_c(N,k)N$. Finally, notice that while the recovery result Theorem 1.3 specified a window of feasible learning rates δ_N for these guarantees to apply, the refutation result covers a much wider range of δ_N.

Our arguments will also show that the ratio of the number of samples used in the descent phase to the number used in the search phase is $O(\alpha_c(N,k)^{-1})$ which is vanishing for $k\geq 2$. More precisely, this observation follows from Theorem 1.4 together with the following. Let τ_η^+ denote the first t such that $m_N(X_t)>\eta$ and let $\tau_{1-\eta}^+$ denote the first t such that $m_N(X_t)>1-\eta$.

Theorem 1.5. *Suppose that Assumptions A and B hold and that the population loss has information exponent $k\geq 2$. Let $M=\alpha_N N$ with (α_N,δ_N) as in Theorem 1.3. For any $\eta>0$ there is a constant $C=C(k,\eta)>0$ such that $\tau_\eta^+\gg\alpha_c(N,k)$ and $|\tau_{1-\eta}^+-\tau_\eta^+|\leq CN$ with probability $1-o(1)$. Furthermore, $X_t>1-2\eta$ for all $\tau_{1-\eta}^+\leq t\leq M$ with probability $1-o(1)$.*

In words Theorem 1.5 says that most of the data is used in the search phase (i.e., to attain some non-trivial correlation), and that from there descent to essentially full correlation is rapid, and takes $O(N)$ samples independently of the class of the problem.

Remark 1.6. In the case that $k = 1$, we show consistent estimation and its refutation only in the regimes $\alpha \to \infty$ and $\alpha \to 0$ respectively. That this is optimal can be seen in the simple example of estimating the mean of an i.i.d. Gaussian vector, where consistent estimation is information theoretically impossible with $\alpha = O(1)$.[4] If one only considers *weak* recovery, one can sharpen the thresholds in α to the $O(1)$ scale: see Theorem 3.3.

Remark 1.7. All of these results hold in the broader setting that $\phi = \phi_N$ varies in N provided the following generalizations of Assumption A and Definition 1.2 hold. We take Assumption A to be that ϕ_N' are equi-continuous and uniformly negative on $(0, 1)$ and Definition 1.2 to be that (1.5) holds uniformly over the sequence.

1.4 Discussion of the main results

Let us now discuss the intuition behind the information exponent and Theorems 1.3–1.4. Together, Theorems 1.3-1.4 show that the information exponent governs, in a sharp sense, the performance of online SGD when started from a uniform at random initializations. (In fact, as we will see in Theorems 3.1-3.2, this is a more general phenomenon for all initializations starting near the equator.)

To understand where these thresholds come from, consider the following simplification of the recovery problem given by

$$m_t = m_{t-1} - \frac{\delta}{N}\phi'(m_{t-1})||\nabla m_{t-1}||^2 \approx m_{t-1} + \frac{\delta}{N} c m_{t-1}^{k-1}\,, \tag{1.7}$$

for m_{t-1} small and some $c > 0$. This amounts to gradient descent for the population loss (omitting the projection as its effects are second order for small δ, see Sections 4.1–4.2.) This corresponds to the "best case scenario" since the observed losses will be corrupted versions of the population loss. One would expect this corruption to only increase the difficulty of recovery.

Analyzing the finite difference equation of (1.7), one finds that if the initial latitude is positive but microscopic, $m_0 \asymp N^{-\zeta}$ for any $\zeta > 0$, there are three regimes with distinct behaviors:

- $k < 2$: the time for m_t to weakly recover ($m_t \geq \eta$) is linear, i.e, order $\delta^{-1}N$.
- $k = 2$: the time for m_t to weakly recover is quasi-linear, i.e., order $\delta^{-1}N \log N$.
- $k > 2$: the time for m_t to weakly recover is polynomial: $\delta^{-1}N^{1+c_\zeta}$ for $c_\zeta = (k-2)\zeta > 0$.

In the online setting, the number of time steps is equal to the number of samples used. Consequently, *no matter what* $\zeta \in (0, 1)$ *is*, there is a transition between linear sample complexity ($\alpha = O(1)$) and polynomial sample complexity ($\alpha = N^\zeta$ for some $\zeta > 0$) regimes for the gradient descent on the population loss as one varies the information exponent, through the critical $k_c = 2$. The precise thresholds obtained in our results correspond

4. Suppose that we are given $M = \alpha N$ samples of an N-dimensional Gaussian vector with law $\mathcal{N}(v, Id)$, where v is a unit vector and α is a fixed constant. It is easy to see that for large N, the sample average only achieves a (normalized) inner product of $v \cdot \frac{\hat{v}}{||v||} \to \sqrt{\alpha(1-\alpha)^{-1}} < 1$ when $\alpha = O(1)$. Furthermore, by the Cramer-Rao bound, this is tight for unbiased estimators satisfying a second moment constraint.

to the choice of $\zeta = 1/2$, which is the scaling for uninformative initializations in high dimensions, e.g., the uniform measure on $\mathbb{S}^{N-1}$.

When one considers the true online SGD, there is an effect due to the sample-wise error for the loss, H_N, whereby to first approximation,

$$m_t \approx m_{t-1} + \frac{\delta}{N} a_k m_{t-1}^{k-1} - \frac{\delta}{N} \nabla H_N^t(X_{t-1}) \cdot \theta_N \,.$$

Due to the independence of ∇H_N^t and X_{t-1}, as we sum the contributions of the third term in time, we obtain a martingale which we call the *directional error martingale*,

$$M_t = \frac{\delta}{N} \sum_{j=1}^{t} \nabla H_N^j(X_{j-1}) \cdot \theta_N \,. \tag{1.8}$$

By Doob's inequality, its cumulative effect can be seen to typically be of order $\delta\sqrt{T}/N$. In order for this term's contribution to be negligible for time scales on the order of M, and allow for recovery, we ask that it is comparable to m_0, dictating that $\alpha\delta^2 = O(1)$. This relative scaling of α and δ is therefore fundamental to our arguments. Indeed if $\alpha\delta^2$ were diverging, then on timescales that are of order αN, the cumulative effect of projections becomes a dominant effect potentially drowning out the drift induced by the signal. We remark here that for the refutation result of Theorem 1.4 when $k \geq 3$, if one only want to assume that $\delta = O(1)$, our arguments would show a refutation result for all $\alpha \leq N^{(k-2)/2}$, implying that for any $O(1)$ choice of δ, the $k \geq 3$ regime still requires polynomial sample complexity.

We end this discussion by briefly comparing our approach to three related approaches that have been used to investigate similar questions in recent years. An in-depth problem-by-problem discussion of the related literature will appear in Section 2 below. The classical approach to such problems is the "ODE method" which dates back at least to the work of Ljung (1977) (see Benaïm, 1999, for a textbook introduction). Here one proves convergence of the trajectory with small step-size to the solution of the population dynamics. While this approach is most similar to our approach, it is designed for fixed dimensions. Indeed, in the scaling regime studied here, namely that corresponding to the initial search phase, one cannot neglect the effect of the directional error martingale, as its increments are *larger* than those of the drift. Another approach to estimation problems via gradient-based algorithms is to study a diffusion approximation to this problem. Diffusion approximations to stochastic approximation algorithms date at least back to the work of McLeish (1976) in finite dimensions; more recently they have been studied in high-dimensions for online algorithms, e.g., Li et al. (2016); Tan and Vershynin (2019). In our setting, rather than setting up a functional law of large numbers or central limit theorem for m_t, the precise nature and analyzability of which we expect depends on, e.g., the choice of activation function for GLMs, we use a system of difference inequalities similar to Ben Arous, Gheissari, and Jagannath (2020a,b).

On the other hand, several statistical physics motivated approaches have recently been introduced to study such questions in both online (Wang and Lu, 2016; Wang, Mattingly, and Lu, 2017) and offline (Mannelli et al., 2019a, 2020) settings. These results develop a scaling limit, in the regime where one first sends $N \to \infty$ then $T \to \infty$. In these settings,

however, the solution to the corresponding limiting evolution equation admits a trivial zero solution if $m_0 = 0$ and a meaningful solution if $m_0 > 0$. This limiting system is then analyzed for its behavior and recovery thresholds, for various values of $m_0 > 0$ small, but uniformly bounded away from zero. As a consequence of our work, we find that if one can produce such initial data then the problem is always solvable with linear sample complexity, and satisfies a scaling limit to the trajectory of the population dynamics. On the other hand, for uninformed initializations, where $m_0 = o(1)$ except with exponentially small probability, the bulk of the data is used just to reach order 1 latitudes.

Remark 1.8. Here and in the following, we have restricted attention to initializations supported on the upper half-sphere. We do this to focus on the key issues and deal with general losses and data distributions in a unified manner with minimal assumptions. For even information exponents, this restriction is without loss of generality by symmetry. For odd information exponents, when started from the lower half-sphere, the dynamics would rapidly approach the equator. We also restrict our attention to starts which have initial correlation on the usual CLT scale ($m(X_0) \sim N^{-1/2}$) which will hold for the uniform at random start. We expect that one can obtain similar results for all possible initializations if one imposes an anti-concentration assumption on the directional error martingale. Without such an assumption, an initialization on the equator $m_0 = 0$ can be trapped for all time.

2. Applications to some important inference problems

In this section, we illustrate how our methods can be used to quantify recovery thresholds for online SGD in inference tasks arising from various broad parametric families commonly studied in high-dimensional statistics, machine learning, and signal processing: supervised learning for single-layer networks with Gaussian features, generalized linear models, linear regression, online PCA, spiked matrix and tensor models, and Gaussian mixture models. Each of these has a vast literature to which we cannot hope to do justice; instead we focus on describing how each satisfy the criteria of Theorems 1.3–1.5, emphasizing how the different information exponent regimes appear in variations on these problems, and relating the recovery thresholds we obtain for online SGD to past work on related algorithmic thresholds for these parameter estimation problems in the high-dimensional regime. In many of our examples, we work only under mild moment assumptions on the data. As our goal here is to demonstrate the applicability of our classification, we do no try to optimize the number of moments assumed. For ease of notation, we henceforth suppress the dependence of quantities on N when it is clear from context.

2.1 Supervised learning for single-layer networks

Consider the following model of supervised learning with a single-layer network. Let $v_0 \in \mathbb{S}^{N-1}$ be a fixed unit vector and suppose that we are given some (possibly non-linear) *activation function* $f : \mathbb{R} \to \mathbb{R}$, some *feature vectors* $(a^\ell)_{\ell=1,\dots,M}$, and with these, M noisy *responses* of the form

$$y^\ell = f(a^\ell \cdot v_0) + \epsilon^\ell \,. \tag{2.1}$$

where $(\epsilon^\ell)^\ell$ are additive errors. Our goal is to estimate v_0 given this data. Our approach is by minimizing the least squares error.

This model and special cases thereof have been studied under many different names by a broad range of communities. It is sometimes called a teacher-student network, a single-index model, or seen as a special class of generalized linear models (see, e.g., Hastie et al., 2009; Bishop, 2006; Barbier et al., 2019). This model has received a tremendous amount of attention in recent years, largely in the regime that the features are taken to be i.i.d. standard Gaussians. Here various properties have been studied such as information theoretic thresholds (Barbier et al., 2019; Mondelli and Montanari, 2018), the geometry of its landscape (Maillard et al., 2020; Sun et al., 2018), as well as performance of various gradient-type and spectral methods (Lu and Li, to appear; Mondelli and Montanari, 2018; Luo et al., 2019).

To place this in the framework of Theorem 1.3, we consider as data the pairs $Y^\ell = (y^\ell, a^\ell)$ for $\ell = 1, ..., M$. Our goal is then to optimize a quadratic loss $\mathcal{L}(\cdot; Y)$ of the form

$$\mathcal{L}(x; Y) = \mathcal{L}(x; (y, a)) = \Big(y - f\Big(a \cdot x\Big)\Big)^2 .$$

Note that we may write the population loss as

$$\Phi(x) = \mathbb{E}\Big[\Big(f\Big(a \cdot x\Big) - f\Big(a \cdot v_0\Big)\Big)^2\Big] + \mathbb{E}[\epsilon^2] . \tag{2.2}$$

For general activation functions f and distributions over the features (a^ℓ), one could proceed to calculate the information exponents by Taylor expansion.

Let us focus our discussion on the most studied regime, namely where (a^ℓ) are i.i.d. standard Gaussian vectors in $\mathbb{R}^N$; for the (ϵ^ℓ) we only assume they are i.i.d. mean zero with finite $4 + \delta$-th moment for some $\delta > 0$. Here we find an explicit representation for the population loss which allows us to compute the information exponent for activation functions of at most exponential growth. With this we find that Theorems 1.3-1.5 apply to all such activation functions and, in particular, we find a wealth of interesting phenomena.

Recall that the Hermite polynomials, which we denote by $(h_k(x))_{k=0}^\infty$, are the (normalized) orthogonal polynomials of the Gaussian distribution $\varphi(dx) \propto \exp(-x^2/2)dx$. Define the k-th Hermite coefficient for an activation function f by

$$u_k(f) = \langle f, h_k \rangle_{L^2(\varphi)} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(z) h_k(z) e^{-z^2/2} dz . \tag{2.3}$$

As long as f' is of at most polynomial growth—i.e., there exist $A, B \geq 0$ and integer $q \geq 0$ such that $|f'(x)| \leq A|x|^q + B$ for all x—the population loss is differentiable and the above exists. The population loss then has the following exact form which we believe is of independent interest.

Proposition 2.1. *Suppose that the features (a^ℓ) are i.i.d. standard Gaussian vectors, and the errors (ϵ^ℓ) are i.i.d. mean zero, variance C_ϵ and with finite $4 + \delta$-th moment for some $\delta > 0$. Suppose that the activation function, f, is differentiable a.e. and that f' has at most polynomial growth. Then*

$$\Phi(x) = \phi_f(m(x)) \qquad \text{where} \qquad \phi_f(m) := 2 \sum_{j=0}^{\infty} (u_j(f))^2 (1 - m^j) + C_\epsilon , \tag{2.4}$$

and where u_j are as in (2.3)*. Furthermore, Assumptions A and B hold.*

With Proposition 2.1, it is easy to compute the information exponents corresponding to general activation functions. The following result is an immediate consequence of Proposition 2.1.

Corollary 2.2. *Suppose that the features (a^ℓ) are i.i.d. standard Gaussian vectors, and the errors (ϵ^ℓ) are i.i.d. mean zero with finite $4+\delta$-th moment for some $\delta > 0$. Suppose that the activation function, f, is differentiable a.e. and that f' has at most polynomial growth. Then,*

1. *The information exponent of f is 1 if and only if $u_1(f) \neq 0$.*
2. *The information exponent of f is 2 if and only if $u_1(f) = 0$ and $u_2(f) \neq 0$*
3. *The information exponent of f is at least 3 if and only if $u_1(f) = u_2(f) = 0$.*

We prove Proposition 2.1 and Corollary 2.2 in Appendix B.1.

Let us now turn to some concrete examples of activation functions and their classification. Many commonly used activation functions have corresponding population loss with information exponent 1, for example:

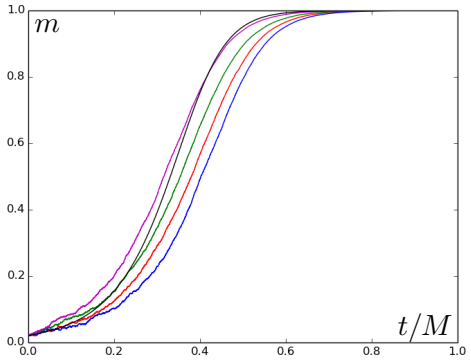
- Adaline: $f(x) = x$
- Sigmoid: $f(x) = \frac{e^x}{1+e^x}$
- ReLu: $f(x) = x \vee 0$.

The problems of smooth and non-convex phase retrieval ($f(x) = x^2$ and $f(x) = |x|$ respectively) are examples of models whose population loss has information exponent 2: see Section 2.1.1 below for a discussion. More generally, we note that the Hermite polynomial of degree k gives a simple example of an activation with information exponent k.

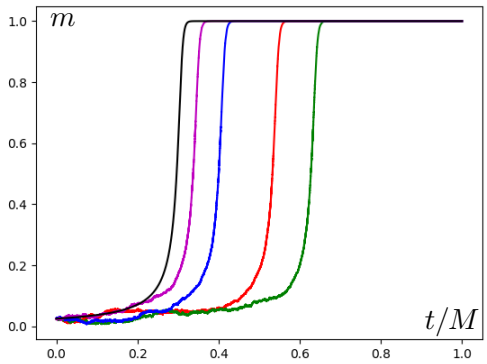
Let us now turn to discussing the implications of our results and in particular, the answers to the last two motivating questions from the introduction in this setting.

Our first observation is that seemingly innocuous changes to an activation function can yield dramatic changes to the sample complexity of the problem. For example, Adaline $f(x) = x$ and its cubic analogue $f(x) = x^3$ both have an information exponent of 1; however, a linear combination of the two, namely $f(x) = x^3 - 3x$ has an exponent of 3. Thus while the first two require linearly many samples to recover the unknown vector, the third requires $\Omega(N^{3/2} \log N)$ many samples just to exit the search phase. As an illustration of this, see Fig. 2.3 where we perform supervised learning via SGD with the same pattern vectors and same unknown v_0 but different activation functions, namely $f(x) = x^3$ and $f(x) = x^3 - 3x$.

On the other hand, note that in the descent phase of the algorithm, where the algorithm exhibits a law of large numbers by Theorem 3.2, one finds that the performance no longer depends on the activation function in a serious way. Thus from a "warm start", the choice of activation is less important. See Figs. 2.1–2.2 for the example of the performance of phase retrieval, $f(x) = x^2$, and $f(x) = x^3 - 3x$ with warm and random starts. Let us emphasize here, however, that as soon as the information exponent is at least 2, we know by Theorem 1.5, that almost all of the run-time is in the search phase.



(a) $f(x) = x^2$, with $N = 3000$ and $\alpha = 100$.



(b) $f(x) = x^3 - 3x$, $N = 3000$, and $\alpha = 30,000$.

Figure 2.1: (Random starts) In colors, 4-runs of single-layer supervised learning with activation functions $f(x) = x^2$ (phase retrieval) and $x^3 - 3x$ (3rd Hermite polynomial), initialized uniformly at random; in black, the corresponding population dynamics. Almost all of the data is used to attain a macroscopic correlation (the search phase); the trajectory through this phase does not concentrate (c.f. Fig. 2.2).

Remark 2.3 (Mis-specification). As another example of how the loss can dramatically affect performance, consider the problem of model mis-specification. Here we attempt to fit to the data using a possibly incorrect activation g. That is, our loss is $\mathcal{L} = (y - g(a \cdot x))^2$, but our data is $y = f(a \cdot x)$. One can compute the information exponent by noting that the loss satisfies

$$\phi(m) = -2 \sum_k u_k(f) u_k(g) m^k + ||f||^2 + ||g||^2 .$$

Here we find various phenomena (no recovery, only weak recovery, or strong recovery) depending on the alignment of the Hermite coefficients of f and g.

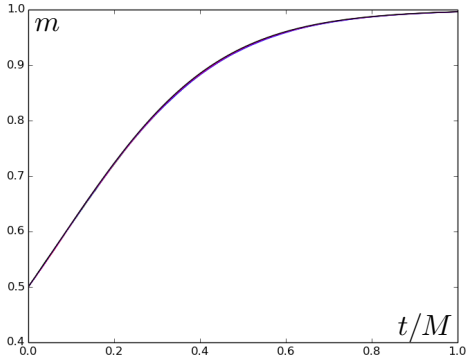
For concreteness, let us consider the case of

$$f(x) = u_2 h_2(x) + u_4 h_4(x) \qquad \text{and} \qquad g(x) = v_1 h_1(x) + v_2 h_2(x) + v_4 h_4(x) .$$

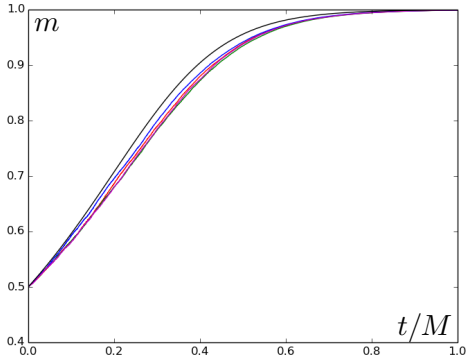
In this case if we let $a = 2u_2 v_2$ and $b = 2u_4 v_4$, we have $\phi(m) = -am^2 - bm^4 + C$. Here we have the following phenomenology.

Strong recovery. Suppose that, $a > 0$, and $a > -2b$, then this model satisfies Assumptions A and B and has information exponent 2. Thus strong recovery has quasilinear sample complexity by Theorem 1.3. On the other hand, if either f or g, has vanishing second Hermite coefficient, so that $a = 0$, while $b > 0$, the model has information exponent 4 and the sample complexity is now cubic (up to log factors).

Weak recovery. Consider for concreteness the case that, $a = 1$ and $b = -1$. Then Assumption B still holds. On the other hand, Assumption A only holds in a neighborhood of the origin. In particular, the global minimum of ϕ on $[0, 1]$ is attained at $1/\sqrt{2}$. Thus by Theorem 3.1 and a minor modification of Theorem 3.2, we see that SGD will weakly recover



(a) $f(x) = x^2$, with $N = 3000$ and $\alpha = 500$.



(b) $f(x) = x^3 - 3x$, $N = 3000$, and $\alpha = 500$.

Figure 2.2: (Warm start) In color, 4 runs of single-layer supervised learning, initialized from $m_0 = 0.5$, together with the corresponding population dynamics. Initialized here, and thus in the descent phase, the SGD rapidly converges (with linearly many data samples) to the ground truth, independently of the information exponent. Moreover, the trajectories are well-concentrated about the population dynamics' trajectory.

on quasi-linear time scales, will rapidly descent to latitude $m \approx 1/\sqrt{2}$, but then remain there on polynomial timescales, i.e., it will weakly recover but not strongly recover v_0.

No recovery. Suppose that and $a < 0$. This would correspond to mis-specifying the sign of the coefficients of f. In this case $\phi = 0$ is a local minimum of ϕ. As such, a modification of the arguments of Theorem 1.4 shows that SGD will not attain a macroscopic correlation in polynomial time. This is to be expected as the inference procedure we are applying is no longer Fisher consistent. On the other hand, consider the more extreme case of $v_1 = u_2 = 1$ and $u_4 = v_2 = v_4 = 0$. Here we try to fit the quadratic transformation of the data, $f(x) = x^2 - 1$, with a linear one, $g(x) = x$. In this case, a=b=0 so that $\phi(m) = 0$, i.e., the population loss is constant and equal to 0. In this case even the population loss is completely uninformative and the algorithm will simply wander around the sphere, never attaining non-trivial correlation.

2.1.1 Phase retrieval

One class of models of the form of (2.1) that has received a tremendous amount of attention in recent years is smooth and "non-smooth" phase retrieval. These correspond to the cases $f(x) = x^2$ and $f(x) = |x|$ respectively. Algorithmic recovery results with different algorithms, initializations, and variants of phase retrieval have been established in a wide array of related settings: Chen et al. (2019); Candès et al. (2015); Zhang and Liang (2016); Lu and Li (to appear); Tan and Vershynin (2018); Aubin et al. (2020); Jeong and Güntürk (2017); Sun et al. (2018).

For our results, in both cases we have that $u_0, u_2 > 0$ and $u_1 = 0$ so that their information exponents are both $k = 2$. As a consequence, Theorem 3.1 shows that if $\alpha/(\log N)^2 \to \infty$, then online SGD with step size $\delta \sim \log N/\alpha^2$ will solve the weak recovery problem

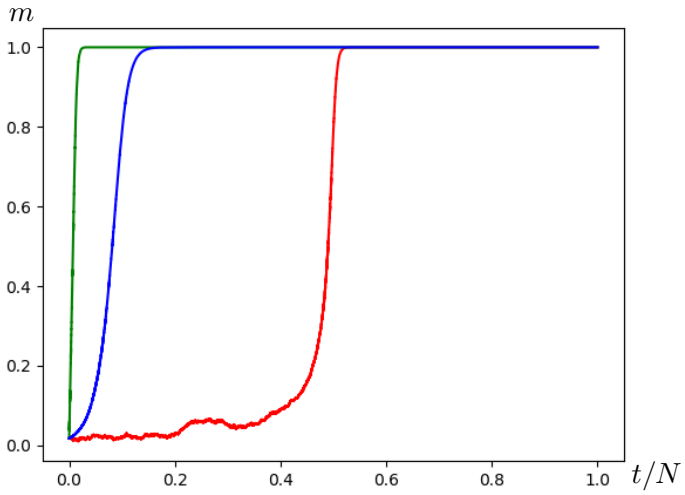


Figure 2.3: SGD trajectories (at $N = 3000$ and $\alpha = 30,000$) from a uniform at random start using the same data, but different activation functions $f(x) = x^3$ (green), $f(x) = x^2$ (blue) and $f(x) = x^3 - 3x$ (red), with information exponents $1, 2, 3$ respectively. The choice of activation function (changing the information exponent) can dramatically change the timescale for consistent estimation.

started from the uniform measure conditioned on the upper-half sphere. Going further, by symmetry of f, the same result holds started from the uniform measure on $\mathbb{S}^{N-1}$ itself, if we wish to weakly recover v_0 only up to a net sign.

In recent independent work, Tan and Vershynin (2019) have obtained sharper results for non-smooth phase retrieval. There they show that a similar recovery result holds as soon as α is order $\log N$, and uniformly over all possible initializations, in particular those with $m(x) = 0$ (c.f., Remark 1.8) Furthermore, their work applies to online learning in "full space", i.e., without restricting to the sphere. Some of the techniques there are similar in spirit to ours, but they involve a careful analysis of a 2D dynamical system which ends up being exactly solvable due to the choice of activation function, whereas we reduce to differential inequalities to handle general choices of activation function.

2.2 Generalized linear models

Generalized linear models (GLM's) are a commonly used family of statistical models which unify a broad class of regression tasks such as linear, logistic, and Poisson regression. For a textbook introduction, see McCullagh and Nelder (1989); Bishop (2006). We follow the presentation of McCullagh and Nelder (1989). (For ease of exposition, we focus only on the case of canonical link functions and constant dispersion.)

Given an observation $y \in \mathbb{R}$ and a covariate vector $a \in \mathbb{R}^N$, a generalized linear model consists of three components:

1. The *random component*: Conditionally on the covariate a, the observation y is distributed according to an element of the exponential family with canonical parameter θ,

$$p_\theta(y) = \exp\Big(\frac{y\theta - b(\theta)}{d} - c(y)\Big) \tag{2.5}$$

for some real-valued functions $b, c : \mathbb{R} \to \mathbb{R}$ and some constant d.

2. The *systematic component*: The covariate a and the unknown parameter $x \in \mathbb{R}^N$ form a linear predictor, $\eta = a \cdot x$.

3. The *link* between the random and the systematic components is given by an *activation function*, $f : \mathbb{R} \to \mathbb{R}$, which is monotone increasing and invertible, and satisfies

$$f(\eta) = \mathbb{E}[y \mid a].$$

For clarity, we work here in the case of a *canonical link*, where we assume that the canonical parameter and the linear predictor are equal, i.e., $\theta = \eta$. In this case, as by definition b is the cumulant generating function of $y|a$, we have that $f(\eta) = b'(\eta)$ The goal is then to infer the parameter x given the observation y. Note that standard regression tasks can be mapped to this setting by choosing b, c and d appropriately. For example, (in these examples we take $d = 1$.)

- Linear regression (linear activation function $f(x) = x$ and $c(t) = b(t) = -t^2/2$.)
- Logistic regression (sigmoid activation function $f(x) = (1 + e^{-x}))^{-1}$, and $b(t) = \log 1 + e^t$, $c(t) = 0$.)
- Poisson regression (exponential activation function $f(x) = e^x$, and $b(t) = t$, $c(t) = -\log t!$)

Consider a random instance of this estimation problem. Let $v_0 \in \mathbb{S}^{N-1}$ be a fixed, unknown parameter. Suppose that we are given M i.i.d. covariate vectors $(a^\ell)_{\ell=1}^M$ and M corresponding observations (y^ℓ). Our goal is to estimate v_0 given (y^ℓ) and (a^ℓ).

To place this in the framework of our results, we augment the observations by the covariates and consider the pairs $Y^\ell = (y^\ell, a^\ell)$. The canonical approach to these problems is maximum likelihood estimation, which in our setting amounts to minimizing the loss

$$\mathcal{L}(x; Y) = \mathcal{L}(x; y, a) = -y\, a \cdot x + b(a \cdot x).$$

For general activation functions f and distributions over the covariates (a^ℓ), one proceeds to calculate the information exponents by Taylor expanding the population loss $\Phi(x) = \mathbb{E}\Phi(x; Y)$.

To make things concrete, let us focus on the setting where the covariates are i.i.d. Gaussian vectors in $\mathbb{R}^N$. Here GLM's have been found to have a rich phenomenology in high dimensions, see, e.g., Sur and Candès (2019) for the case of logistic regression. Here for a function f, let $u_1(f) = \mathbb{E}[Zf(Z)]$ where $Z \sim a \cdot e_1$. We say that a function f is of at most exponential growth if there are constants $A, B > 0$ such that $|f(x)| \le A \exp(B|x|)$. We then have the following.

Proposition 2.4. *Let f be an invertible, strictly increasing activation function of at most exponential growth for a GLM with standard Gaussian covariates $(a^\ell)_{\ell=1}^M$. Suppose that f is differentiable a.e. and that the exponential family (2.5) has finite $8+\iota$-th moment for some $\iota > 0$. Then the population loss, $\Phi(x)$, satisfies*

$$\Phi(x) = -u_1(f)m + C \tag{2.6}$$

for some constant $C \in \mathbb{R}$. Furthermore, $u_1(f) > 0$ so that the information exponent is $k = 1$ and Assumptions A and B hold.

By Proposition 2.4, maximum likelihood estimation for generalized linear models with invertible, increasing activations as above will always have information exponent $k = 1$. The proof of Proposition 2.4 is deferred to Appendix B.2.

2.3 Linear regression with random covariates

Let us return to the example of linear regression with random covariates discussed in the introduction. Suppose that we are given i.i.d. data points of the form $\{(y^\ell, a^\ell)\}_{\ell=1}^M$, where $y^\ell = a^\ell \cdot v + \epsilon^\ell$, for some unknown unit vector v and additive noise ϵ^ℓ. Consider the squared-error loss $\mathcal{L}(x; y^\ell, a^\ell) = (y^\ell - a^\ell \cdot x)^2$. In this case, provided the covariates and errors are centered and have finite second moment, the population loss is of the form

$$\Phi(x) = ||x - v||^2 + 1 = 2 - 2m(x) + C, \tag{2.7}$$

for x on the unit sphere.

Proposition 2.5. *Consider linear regression with centered i.i.d. covariates $(a^\ell)_\ell$ whose entries have finite 10-th moment and centered i.i.d. errors $(\epsilon^\ell)_\ell$ with finite 5-th moment. Then Assumptions A and B hold and the population loss has information exponent 1.*

Proposition 2.5 is proved in Appendix B.3. We emphasize here that we do not assume that the covariate vectors have independent entries, nor do we assume Gaussianity of either the covariates or the errors.

Remark 2.6. We now return to the issue raised in the introduction regarding the scaling of the risk and its derivatives. For simplicity let us focus on the case where the designs and noise are i.i.d. standard Gaussian. The canonical empirical risk in this setting is the least squares risk $\hat{R}(x) = ||Y - Ax||_2^2$. To apply the SDE approximations referenced above to the SGD we consider, one should work with this risk. By direct differentiation, we see that the operator norm of $\nabla^2 \hat{R}$ scales like that of AA^T which, recalling classical properties of Wishart matricies (Anderson et al., 2010), scales like M so that the empirical risk is L-smooth for L that is of order M. A similar argument applies to the Lipschitz constant. On the other hand, one might rescale and consider instead $\tilde{R}(x) = \frac{1}{M}||Y - Ax||_2^2$. The same reasoning, however, shows that $\|\tilde{R}\|_\infty \leq C$ with high probability. Thus the resulting invariant measure $\pi(dx) \propto e^{\tilde{R}(x)}dx$ is a uniformly bounded tilt of the uniform measure ($cdx \leq d\pi \leq Cdx$ for some $C, c > 0$ independent of N) and yields an exponentially small mass for any ϵ-neighborhood of v for $\epsilon < \pi/2$ by concentration of measure.

2.4 Online PCA

From a statistical perspective, a classical application of online SGD to a matrix model is online PCA (sometimes called streaming PCA) whose analysis goes back to the work of Krasulina (1969) and Oja and Karhunen (1985). The long history and vast literature on this problem makes it impossible to provide anywhere near a complete discussion of previous work, but we show that the problem fits into our classification. Here one is given i.i.d. samples (Y^ℓ) from some distribution in an online fashion and the goal is to find the principal directions. Our results immediately apply in the commonly studied rank 1 setting where we assume that there is only one principal direction: namely, suppose that (Y^ℓ) are i.i.d. centered random vectors in $\mathbb{R}^N$ with covariance $\mathbb{E}[YY^T] = S$, where

$$S = I + \lambda v_0 v_0^T$$

where v is a fixed unit vector, $\lambda > 0$ is a fixed signal-to-noise ratio, and one iteratively optimizes the loss

$$\mathcal{L}(x; Y) = -(x, YY^T x)$$

over $\mathbb{S}^{N-1}$. We view this as a stochastic approximation to $-(x, \mathbb{E}[YY^T]x)$. In the case that $Y^\ell \sim N(0, I + \lambda v_0 v_0^T)$, the offline version of this problem immediately connects to the study of spiked Wishart matrices investigated in many works.

Observe that if we let $S = I + \lambda v_0 v_0^T$, we have

$$\Phi(x) = -(x, \mathbb{E}[YY^T]x) = -(x, Sx) = -\lambda(v_0, x)^2 - 1, \tag{2.8}$$

since x is a unit vector. Evidently this problem has information exponent 2 so that the thresholds match those in the spiked matrix models of Section 2.5. In this setting, we have the following which holds for general (Y^ℓ) under mild moment assumptions.

Proposition 2.7. *Suppose that $\lambda > 0$ is fixed and that (Y^ℓ) are i.i.d. centered random vectors whose entries have bounded 10-th moment: $\sup_i \mathbb{E}[Y_i^{10}] < C$ for some $C > 0$. Then Assumptions A and B hold, and the model has information exponent 2.*

The proof of Proposition 2.7 is elementary and can be found in Appendix B.4.

2.5 Spiked matrix and tensor models

Another important class of examples are the spiked matrix and tensor models (also referred to as tensor PCA) (Johnstone, 2001; Péché, 2006; Richard and Montanari, 2014). Here we are given $M = \alpha N$ i.i.d. observations (Y^ℓ) of a rank 1 p-tensor on $\mathbb{R}^N$ corrupted by additive noise:

$$Y^\ell = J^\ell + \lambda v_0^{\otimes p} \tag{2.9}$$

where (J^ℓ) are i.i.d. copies of p-tensors with i.i.d. entries of mean zero and variance one, and λ is a signal-to-noise parameter. The goal in these problems is to infer the vector $v_0 \in \mathbb{R}^N : \|v_0\| = 1$. A standard approach to inferring v_0 is by optimizing the following ℓ^2 loss:

$$\mathcal{L}(x; Y) := \big(Y, x^{\otimes p}\big) = (J, x^{\otimes p}) + \lambda\big(x \cdot v_0\big)^p. \tag{2.10}$$

When J is Gaussian, and we restrict to $\mathbb{S}^{N-1}$, this is simply maximum likelihood estimation. When J is non-Gaussian, this can be thought of as computing the best rank 1 approximation to J.

We obtain the following result regarding the information exponent of these models whose proof is deferred to Appendix B.5.

Proposition 2.8. *Consider the spiked tensor model where $\lambda = 1$ and J is an i.i.d. p-tensor with loss* (2.10)*. Suppose that entries of J have mean zero and finite 6th moment. Then Assumptions A and B hold and the population loss has information exponent p.*

From this we see that the spiked matrxi model falls in our quasi-linear class, whereas the spiked tensor model is in the polynomial class. We now discuss the relation between our sample complexity thresholds for online SGD and those found for other algorithms in preceding work. For the sake of that comparison, we recall here that the information theoretic threshold for this estimation problem, with the above scaling is at $\lambda\sqrt{\alpha}$ of order one (Péché, 2006; Montanari et al., 2015; Perry et al., 2018; Lesieur et al., 2017; Ben Arous et al., 2019; Ros et al., 2019; Jagannath et al., 2020).

2.5.1 Spiked tensor models

The tensor case has recently seen a surge of interest as it is expected to be an example of an inference problem which has a diverging statistical-to-algorithmic gap. Strong evidence for this comes from the analysis of spectral and sum-of-squares-type methods where sharp thresholds of the form $\lambda\sqrt{\alpha} \sim N^{\frac{k-2}{4}}$ have been obtained for efficient estimation (Richard and Montanari, 2014; Kim et al., 2017; Hopkins et al., 2015; Wein et al., 2019; Hopkins et al., 2016).

On the other hand, it was conjectured Richard and Montanari (2014) that power iteration and approximate message passing should have a threshold of $\lambda\sqrt{\alpha} = O(N^{\frac{k-2}{2}})$. It has been speculated that the latter threshold may be common to first-order methods without any global or spectral initialization step.

In Ben Arous et al. (2020a), efficient recovery was proved for $\lambda\sqrt{\alpha}$ growing faster than $N^{(k-2)/2}$ for gradient descent and Langevin dynamics, using a differential inequalities based argument originating from Ben Arous et al. (2020a) and evidence was provided for hardness when it is smaller. (See also Mannelli et al. (2019b,a, 2020); Biroli et al. (2020) for related, non-rigorous analyses.) The threshold we find of $\alpha \sim N^{k-2}$ in Theorem 1.3, exactly matches this, and shows that the online SGD attains the (conjecturally) optimal thresholds for first order methods. Note, however, that the above works required the random tensor J to be Gaussian, whereas Proposition 2.8 covers non-Gaussian J as well.

Remark 2.9. Similarly to the online PCA example, when $p = 2$ our results imply a threshold for online SGD of at least $\alpha \sim \log N$ and at most $\alpha \sim (\log N)^2$ for spiked matrix models of the form (2.9). The reader may note that one expects to be able to solve spiked matrix estimation tasks with $\alpha = O(1)$, e.g., in the offline setting using gradient descent. The $\log N$ factor in our results is due to the on-line nature of this setting and should be compared to the run-time of gradient descent in the offline setting with a random start, which will indeed take $\log N$ time just to weakly recover.

2.5.2 Spiked tensor-tensor and matrix-tensor models: min-stability of information exponents

Recently, there have been several results regarding the so-called spiked matrix-tensor and spiked tensor-tensor models, where one is given a pair of tensors of the form

$$Y = \left(J + \lambda v_0^{\otimes p}, \tilde{J} + \lambda v_0^{\otimes k} \right) ,$$

where J and $\tilde{J}$ are independent p and k-tensors respectively. Here one is interested in inferring v_0 via the sum of the losses. Evidently these problems will have information exponent $\min\{p, k\}$.

In the case $p > k = 2$, scaling limits (as $N \to \infty$ and then $m_0 \downarrow 0^+$) of approximate message passing through state evolution equations and gradient descent through the Crisanti–Horner–Sommers–Cugliandolo–Kurchan equations have been studied in Mannelli et al. (2019a,b, 2020). In our results, we investigate the regime $m_0 \sim N^{-1/2}$ corresponding to a random start, and find that online SGD recovers for $\alpha \gtrsim (\log N)^2$ and fails for $\alpha = o(\log N)$. We expect that $\alpha = \Theta(\log N)$ is in fact the true threshold, since (offline) gradient decent requires $N \log N$ time to optimize the population loss from a random start.

2.6 Two-component mixture of Gaussians: an easy-to-critical transition

Consider the case of maximum likelihood estimation for a mixture of Gaussians with two components. We assume for simplicity that the variances are identical and that the mixture weights are known. We consider the "spherical" case and assume that the clusters are antipodal, i.e.,

$$Y \sim p\mathcal{N}(v_0, Id) + (1-p)\mathcal{N}(-v_0, Id) \qquad \text{for} \qquad v_0 \in \mathbb{R}^N : ||v_0|| = 1 \, . \tag{2.11}$$

Without loss of generality, take $v_0 = e_1$. The log-likelihood in this case is of the form

$$\tilde{f}(x, Y) = \log \left(p \cdot \exp\Big(-\frac{1}{2}||Y - x||^2 \Big) + (1-p) \cdot \exp\Big(-\frac{1}{2}||Y + x||^2 \Big) \right).$$

To place this in to our framework, it will be helpful to re-parameterize the mixture weights as $p \propto e^h$ for some $h \in \mathbb{R}$. In this setting, we see that maximum likelihood estimation is equivalent to minimizing the loss

$$\mathcal{L}(x; Y) = -\log\cosh(Y \cdot x + h).$$

Proposition 2.10. *Consider a two-component Gaussian mixture model as in* (2.11)*. If we take as loss the (negated) log-likelihood, then Assumptions A and B hold. Furthermore, if* $p \neq 1/2$*, it has information exponent* 1*, and if* $p = 1/2$*, it has information exponent* 2*.*

The proof of Proposition 2.10 is deferred to Appendix B.6.

While there is a huge literature in the Gaussian mixture model setting, we mention a few recent results related to our work. As a consequence of these results we obtain an $O(N \log^2 N)$ sample complexity upper bound for online SGD. It is known that, from the perspective of learning the density in TV distance, this is optimal, see Kalai et al. (2010); Suresh et al. (2014); Ashtiani et al. (2018). With different assumptions on the sample complexity, Mei et al. (2018) used a critical point analysis to understand the behavior of true gradient descent.

3. Analysis of two stages of performance

We now turn the proofs of our results. We begin by stating our main technical results for the search and descent stages of performance of SGD, together giving Theorem 1.3. These two results show that dynamics attains order one correlation and exits the search phase on a timescale of $\tilde{O}(\alpha_c(N,k)N)$ and, from there, well-approximates the population dynamics until it quickly attains $1-o(1)$ correlation. Without loss of generality and for notational convenience we will take $\theta_N = e_1$ for the remainder of this paper, where e_1 denotes the usual Euclidean basis vector.

3.1 Search phase

Our main results for the search phase shows that the timescale for attaining weak recovery when started at any point with $m_0 = \Omega(N^{-1/2})$ is of order $\tilde{O}(\alpha_c(N,k)N)$. For this weak recovery result, we actually do not need the full strength of Assumption A. To that end, for $0 < \varrho \leq 1$, let us introduce **Assumption $\mathbf{A}_\varrho$** that ϕ is differentiable, and ϕ' is strictly negative on $(0,\varrho)$. Evidently Assumption A implies Assumption A_ϱ for every ϱ.

Theorem 3.1. *Suppose there exists $\varrho > 0$ such that Assumptions A_ϱ and B hold and that the population loss has information exponent k. Let (α_N, δ_N) be in Theorem 1.3. Then there exists $\eta > 0$ such that if $X_t = X_t^{N,\delta}$ is the online SGD with step size δ, we have for every $\gamma > 0$,*

$$\lim_{N\to\infty} \inf_{x_0 : m(x_0) \geq \gamma/\sqrt{N}} \mathbb{P}_{x_0}\left(\tau_\eta^+ < \alpha N\right) = 1\,.$$

where, we recall, τ_η^+ is the stopping time $\inf\{t : m_t > \eta\}$.

Before turning to the corresponding refutation theorem, let us pause to comment on the role of initialization in this result. We note here that this result is uniform over any initial data with $\{m(x) \geq \gamma/\sqrt{N}\}$ for $\gamma > 0$. For $m(x)$ on this scale the sample-wise error for the loss, H_N, dominates the population loss substantially. In particular, both the sample and empirical losses in that region are typically highly non-convex. This scaling, however, is the natural scaling for initializations in high-dimensional problems as the uniform measure on the upper half-sphere satisfies m_0 of order $N^{-1/2}$ with probability $1-o(1)$ (Ledoux and Talagrand, 2011): in this manner, the assumption on the initialization of Theorem 3.1 is weaker than that of Theorem 1.3. For a discussion of initializations having $m_0 = o(N^{-1/2})$ c.f. Remark 1.8, where we noted that the timescales to recovery would depend on a lower bound on the variance of the directional-error martingale, corresponding to an additional lower-bound assumption in (1.3) of Assumption A.

In an analogous manner, the initialization in Theorem 1.4 can be boosted to be uniform over initializations having $m_0 = O(N^{-1/2})$; indeed this is the version we will prove.

3.2 Descent phase

In the search phase, the recovery follows by showing that m_t obeys a differential inequality comparable to that satisfied by $\bar{m}_t$, the correlation of the population dynamics: SGD on the population loss, as defined below. This shows that whereas $m_t - \bar{m}_t$ may be quite large (the

directional-error martingale is on the same scale as m_t), the timescale of the hitting time τ_η^+ is the same as that of the population dynamics. In the descent phase, low-dimensional intuition applies and we establish that $\bar{m}_t$ is indeed a good approximation to the trajectory of m_t, leading to consistent estimation in a further linear time.

To be more precise, let $\overline{X}_t$ be the *population dynamics*, i.e., given by the following procedure

$$\overline{X}_t = \frac{1}{\|\overline{X}_{t-1} - \frac{\delta}{N}\nabla\Phi(\overline{X}_{t-1})\|}\Big(\overline{X}_{t-1} - \frac{\delta}{N}\nabla\Phi(\overline{X}_{t-1})\Big)\,.$$

Since $\Phi(x) = \phi(m)$, the population dynamics can be reduced to its 1D projection onto the correlation variable. Under Assumption A, we have that for every $\delta = O(1)$, and initializations $\overline{X}_0$ with $\underline{\lim}_N m(\overline{X}_0) > 0$,

$$\lim_{T\to\infty} \underline{\lim}_{N\to\infty} m(\overline{X}_{T\delta^{-1}N}) \to 1\,.$$

We also note that if Assumption A does not hold, then the above would *not* hold.

Theorem 3.2. *Suppose that Assumption A and B hold. Fix any $\eta > 0$ and let $X_0 = \overline{X}_0$ be any point such that $m(\overline{X}_0) = \eta$. For $\alpha_N = \omega(1)$, and every $\alpha\delta^2 = o(1)$, we have*

$$\sup_{\ell\le M} \big|m(X_\ell) - m(\overline{X}_\ell)\big| \to 0 \qquad \text{in } \mathbb{P}\text{-prob.} \tag{3.1}$$

In this way, upon attaining non-trivial correlation with the ground truth, and reaching the descent phase, the SGD behaves as it would in a low-dimensional trust region and the usual ODE method applies. Namely, the high-dimensionality of the landscape is no longer relevant. Moreover, the information exponent no longer plays a central role, and the descent always takes linear time.

3.3 Linear sample complexity when $k = 1$

Problems in the $k = 1$ class behave the same in their search and descent phases, and are therefore easy in the sense that their entire trajectory satisfies the law of large numbers of Theorem 3.2. As a result, we are able to sharpen the results above in the $k = 1$ case to analyze the situations where α_N is of order one (as opposed to diverging/going to zero arbitrarily slowly as was assumed in Theorem 1.3), at the cost of only attaining weak recovery. As mentioned in Remark 1.6, in the linear sample complexity regime, there are problems for which weak recovery is the best that is information theoretically possible.

Theorem 3.3. *Suppose that Assumptions A and B hold and that the population loss has information exponent $k = 1$. For every $\epsilon > 0$, the following holds:*

(a) There exists $\eta > 0$ and $\alpha_0 > 0$, such that every $\alpha > \alpha_0$, there is a choice of $\delta > 0$ such that $X_t = X_t^{N,\delta}$ initialized from the uniform measure on the upper half-sphere μ_N satisfies

$$\underline{\lim}_{N\to\infty} \mathbb{P}_{\mu_N}\big(m(X_M) \ge 1 - \eta\big) \ge 1 - \epsilon\,;$$

(b) For every $\eta > 0$ there exists $\alpha_0' > 0$ such that for every $\alpha < \alpha_0'$, and every $\delta = O(1)$,

$$\overline{\lim_{N\to\infty}} \mathbb{P}_{\mu_N}\big(\max_{t\le M} m(X_t) > \eta\big) \le \epsilon\,.$$

Theorem 3.3 will be proved in conjunction with the proofs of Theorems 1.3–1.4.

4. A Difference inequality for Online SGD

The key technical result underlying our arguments is to show that with high probability, the value of $m_t = m(X_t)$ is a super-solution to (a constant fraction of) the integral equation satisfied by the underlying population dynamics. To state this result, we introduce the following notation.

Throughout this section, we view $X_0 = x_0$ as a fixed initial data point and recall the notation $\mathbb{P}_{x_0}$ emphasizing the dependence of the output of the algorithm on the initial data. We prove all the results in the general setting where the population loss $\Phi_N = \phi_N(m_N(x))$ may be such that ϕ also depends on N (see Remark 1.7 for the relevant modifications to Assumption A and the information exponent). From now on we suppress the N dependence in these notations whenever clear from context.

Recall the definition of $m(x)$ from (1.1); we will use the following notations for spherical caps:

$$E_\eta = \big\{x \in \mathbb{S}^{N-1} : m(x) \ge \eta\big\}\,. \tag{4.1}$$

Our results in this section focus on arbitrary $X_0 = x_0$ in $E_{\gamma/\sqrt{N}} \setminus E_\eta$, for some η sufficiently small but positive. If instead $m(X_0) > \eta$, then Theorem 1.3 follows immediately from Theorem 3.2.

For every θ, define the hitting times

$$\tau_\theta^+ := \min\{t \ge 0 : m(X_t) \ge \theta\}\,, \qquad \text{and} \qquad \tau_\theta^- := \min\{t \ge 0 : m(X_t) \le \theta\}\,.$$

Fix $\iota > 0$ given by (1.4) and define

$$\bar{L} := \sup_x \mathbb{E}\big[|\frac{1}{\sqrt{N}}\nabla H(x)|^{4+\iota}\big] \vee \sup_x \mathbb{E}\big[|\frac{1}{\sqrt{N}}\nabla H(x)|^2\big] \vee 1\,. \tag{4.2}$$

Furthermore, let $a_k = c \cdot k$ and $a_{k+1} = C \cdot (k+1)$, where C, c are as in Definition 1.2. Our first goal is to obtain a difference inequality for the evolution of $m_t := m_N(X_t)$ that holds for all $t \le \tau^-_{\gamma/(2\sqrt{N})} \wedge \tau^+_\eta \wedge \bar{t}$ where $\bar{t} \le M$ is some guaranteed recovery time for the algorithm.

Proposition 4.1. *Suppose that Assumptions A_ϱ and B hold and that the population loss has information exponent k. Let $D = D_N$ and $\varepsilon = \varepsilon_N = O(1)$, and suppose $\alpha = \alpha_N$ is of at most polynomial growth in N, and $\delta = \delta_N$ is such that $\delta^2 \le \varepsilon$ and for some $K > 0$,*

$$\delta \le \bar{\delta}_N(k) := \begin{cases} \frac{a_1}{4K\bar{L}} & k = 1 \\ \frac{a_k \gamma^{k-2}}{K\bar{L}N^{\frac{k-2}{2}}\log N} & k \ge 2 \end{cases}\,. \tag{4.3}$$

Then for every $\gamma > 0$ and every $T \le M := \alpha N$ satisfying

$$T \le \frac{N\gamma^2}{D^2\delta^2} =: \bar{t}\,, \tag{4.4}$$

online SGD with step-size δ satisfies the following as $N \to \infty$ for some $\eta > 0$, uniformly over the choice of D, ε, K:

1. *If $k = 1$, there exists a constant $C = C(C_1, a_1, a_2) > 0$ such that*

$$\inf_{x_0 \in E_{\gamma/\sqrt{N}}} \inf_{t \le T} \mathbb{P}_{x_0}\left(m_t \ge \frac{m_0}{2} + \frac{\delta a_k}{8N} t \quad \forall t \le \tau^-_{\gamma/(2\sqrt{N})} \wedge \tau^+_\eta \right) \ge 1 - C\Big(\frac{1}{K} - \frac{1}{D^2}\Big) - o(1), \tag{4.5}$$

2. *If $k \ge 2$, there exists a constant $C(C_1, a_k, a_{k+1}) > 0$ such that*

$$\inf_{x_0 \in E_{\gamma/\sqrt{N}}} \mathbb{P}_{x_0}\Big(m_t \ge \frac{m_0}{2} + \frac{\delta a_k}{8N} \sum_{j=0}^{t-1} m_j^{k-1} \quad \forall t \le \tau^-_{\gamma/(2\sqrt{N})} \wedge \tau^+_\eta \wedge T \Big) \ge 1 - \frac{C}{D^2} - o(1)\,. \tag{4.6}$$

The proof of Proposition 4.1 will follow in three stages. In Section 4.1, we split the evolution of $m(X_t)$ in three parts: the drift induced by the population loss, a martingale induced by the gradient of the sample-wise error in the direction of e_1, and a second order (in δ) effect caused by the non-linear projection step in the algorithm. In Section 4.2–4.3, we show that the second order effect is bounded by the drift of the population loss. In Section 4.4 we control the martingale; while its contribution dominates the drift on short time scales, its total contribution by time $T \le M$ is smaller than the initial bias $\gamma/\sqrt{N}$. We end this section with the proof of Proposition 4.1.

4.1 Controlling radial effects

Let us begin by controlling the effect of the projection in (1.2) and obtain a difference equation for m_t. By the chain rule, $\nabla\Phi(x) = \phi'(m(x))\nabla m(x)$, where

$$\nabla m(x) = e_1 - (x \cdot e_1)x\,.$$

Since $\|m\|_{L^\infty(\mathbb{S}^{N-1})} \le 1$, and ϕ' is continuous, there exists $A > 0$, depending only on a_k, a_{k+1} (and not on N), such that

$$\sup_x |\nabla\Phi(x)|^2 = \sup_x |\phi'(m(x))\nabla m(x)|^2 \le A\,.$$

By Jensen's inequality and (1.4), $\bar{L} = O(1)$. Let

$$L_t = |\frac{1}{\sqrt{N}} \nabla H^t(X_{t-1})|^2\,, \tag{4.7}$$

and observe that

$$|\frac{1}{\sqrt{N}} \nabla \mathcal{L}(X_{t-1}; Y^t)|^2 \le 2\left(\frac{A}{N} + L_t\right).$$

Recalling (1.2), let $r_t = |\tilde{X}_t|$, and note that by orthogonality of $\nabla \mathcal{L}(X_{t-1}; Y^t)$ and X_{t-1},

$$r_t \le \sqrt{1 + (\delta |\nabla \mathcal{L}(X_{t-1}; Y^t)|/N)^2}.$$

Since $1 \le \sqrt{1+u} \le 1 + u$ for every $u \ge 0$, we obtain

$$1 \le r_t \le 1 + \delta^2 \left(\frac{A}{N^2} + \frac{L_t}{N} \right). \tag{4.8}$$

By definition of X_t , we then see that for every $t \ge 1$,

$$m_t = \frac{\tilde{X}_t \cdot e_1}{r_t} = \frac{1}{r_t} \left(m_{t-1} - \frac{\delta}{N} \nabla \Phi(X_{t-1}) \cdot e_1 - \frac{\delta}{N} \nabla H^t(X_{t-1}) \cdot e_1 \right). \tag{4.9}$$

Since for $u \ge 0$, we have $|\frac{1}{1+u} - 1| \le |u|$, we may combine these estimates to obtain

$$\begin{aligned} m_t \ge m_{t-1} &- \frac{\delta}{N} \nabla \Phi(X_{t-1}) \cdot e_1 - \frac{\delta}{N} \nabla H^t(X_{t-1}) \cdot e_1 \\ &- \delta^2 (\frac{A}{N^2} + \frac{L_t}{N}) |m_{t-1}| - \delta^3 (\frac{A}{N^2} + \frac{L_t}{N}) \Big(\Big| \frac{\nabla \Phi(X_{t-1}) \cdot e_1}{N} \Big| + \Big| \frac{\nabla H^t(X_{t-1}) \cdot e_1}{N} \Big| \Big). \end{aligned} \tag{4.10}$$

To better control the terms that are second order in δ, let us introduce a truncation of L_t. Fix a truncation value $\hat{L} > 0$, possibly depending on N, to be chosen later. We can rewrite (4.10) as

$$\begin{aligned} m_t \ge m_{t-1} &- \frac{\delta}{N} \nabla \Phi(X_{t-1}) \cdot e_1 - \frac{\delta}{N} \nabla H^t(X_{t-1}) \cdot e_1 - \delta^2 \Big(\frac{L_t \mathbf{1}_{\{L_t < \hat{L}\}}}{N} \Big) |m_{t-1}| \\ &- \delta^2 \Big(\frac{A}{N^2} + \frac{L_t \mathbf{1}_{\{L_t \ge \hat{L}\}}}{N} \Big) |m_{t-1}| - \delta^3 (\frac{A}{N^2} + \frac{L_t}{N}) \Big(\Big| \frac{\nabla \Phi(X_{t-1}) \cdot e_1}{N} \Big| + \Big| \frac{\nabla H^t(X_{t-1}) \cdot e_1}{N} \Big| \Big). \end{aligned} \tag{4.11}$$

We now turn to controlling the second order in δ correction terms, which we will see have to be treated separately for the one on the first line above, and those on the second line above.

4.2 Bounding the higher order corrections

We begin with a priori bounds on the higher order terms in (4.11). Observe that for every $\eta < \frac{1}{2}$, for every $x \in \{x : m(x) \in [0, \eta]\}$,

$$\nabla m(x) \cdot e_1 = (e_1 - (x \cdot e_1) x) \cdot e_1 = 1 - m(x)^2 \ge 1 - \eta^2 \ge \frac{1}{2}.$$

Then there exists $\eta_0(\varrho, a_k, a_{k+1}) > 0$ such that for all $\eta < \eta_0$, for all $x \in \{x : m(x) \in [0, \eta]\}$,

$$\frac{1}{4} a_k (m(x))^{k-1} \le -\nabla \Phi(x) \cdot e_1 \le \frac{3}{2} a_k (m(x))^{k-1}. \tag{4.12}$$

With this in hand, the next lemma gives a pairing of the second order term on the second line of (4.11), which we bound subsequently by comparison to the initial m_0, and the second order term on the first line of (4.11) which we bound by comparison to the first order (in δ) drift term.

Lemma 4.2. Let $\eta, \gamma > 0$ with $\eta < 1/2$. For every $K > 0$, and every $\delta \le \bar{\delta}_N(k)$ as in (4.3):

1. If $k = 1$, for every $x \in \{x : m(x) \in [0, \eta]\}$,

$$\frac{\delta^2}{N}|m(x)| \le \frac{a_k}{4K\bar{L}}\frac{\delta}{N}\,, \tag{4.13}$$

2. If $k \ge 2$, for every $x \in \{x : m(x) \in [\gamma/(2\sqrt{N}), \eta]\}$,

$$\frac{\delta^2}{N}|m(x)| \le \frac{\delta}{N}\frac{a_k|m(x)|^{k-1}}{K\bar{L}\log N}. \tag{4.14}$$

Proof. First let $k = 1$ and suppose $x \in \{x : m(x) \in [0, \eta]\}$. For every $K > 0$ and every N , if we take $\delta \le a_k/(4K\bar{L})$ then we have

$$\delta^2|m(x)| \le \delta\frac{a_k}{4K\bar{L}}\,.$$

Similarly, if $k \ge 2$, we see that if $x \in \{x : m(x) \in [\gamma/(2\sqrt{N}), \eta]$, we have by (4.3),

$$\delta^2|m(x)| \le \frac{a_k}{K\bar{L}\log N}\delta|m(x)|\Big(\frac{\gamma}{\sqrt{N}}\Big)^{k-2} \le \frac{a_k}{K\bar{L}\log N}\cdot\delta|m(x)|^{k-1}. \qquad \square$$

Lemma 4.3. Suppose that $\alpha\delta^2 \le \varepsilon$ for some $\varepsilon > 0$, let $\bar{L}$ be as in (4.2). Then there is a constant $C = C(\bar{L}, A, C_1, C_2) > 0$ such that the following hold uniformly over $x_0 \in \mathbb{S}^{N-1}$

$$\mathbb{P}_{x_0}\left(\sup_{t\le M}\delta^3\sum_{j=0}^{t-1}\Big(\frac{A}{N^2}+\frac{L_{j+1}}{N}\Big)\Big(\Big|\frac{\nabla\Phi(X_j)\cdot e_1}{N}\Big|+\Big|\frac{\nabla H^{j+1}(X_j)\cdot e_1}{N}\Big|\Big) > \frac{\gamma}{10\sqrt{N}}\right) \le \frac{C\varepsilon}{\gamma N^{3/2}}\,. \tag{4.15}$$

$$\mathbb{P}_{x_0}\Big(\sup_{t\le M}\delta^2\sum_{j=0}^{t-1}\Big(\frac{A}{N^2}+\frac{L_{j+1}\mathbf{1}_{\{L_{j+1}\ge\hat{L}\}}}{N}\Big)|m_j| > \frac{\gamma}{10\sqrt{N}}\Big) \le \frac{C\varepsilon\sqrt{N}}{\hat{L}^{1+\frac{\iota}{2}}\gamma}\,. \tag{4.16}$$

Proof. Both bounds follow from Markov's inequality. Since the summands are positive, the suprema over $t \le M$ are attained at $t = M$, so it suffices to consider that case. Fix $x_0 \in \mathbb{S}^{N-1}$. The bounds will be seen to be uniform in this choice. We begin with (4.15); for every $\lambda > 0$,

$$\begin{aligned}
&\mathbb{P}_{x_0}\left(\sum_{j=0}^{M-1}\delta^3\ \Big(\frac{A}{N^2}+\frac{L_{j+1}}{N}\Big)\Big(\Big|\frac{\nabla\Phi(X_j)\cdot e_1}{N}\Big|+\Big|\frac{\nabla H^{j+1}(X_j)\cdot e_1}{N}\Big|\Big) > \lambda\right)\\
&\qquad\le\frac{M\delta^3}{\lambda N^2}\sup_x\mathbb{E}\left[\Big(\frac{|\nabla H(x)|^2}{N}+\frac{A}{N}\Big)\cdot\Big(\big|\nabla\Phi(x)\cdot e_1\big|+\big|\nabla H(x)\cdot e_1\big|\Big)\right]\\
&\qquad\le\frac{2\alpha\delta^3}{\lambda N^2}\sqrt{\sup_x\mathbb{E}\Big[\Big|\frac{\nabla H(x)}{\sqrt{N}}\Big|^4\Big]+\frac{A^2}{N^2}}\sqrt{\sup_x\big|\nabla\Phi(x)\cdot e_1\big|^2+\sup_x\mathbb{E}\big[\big|\nabla H(x)\cdot e_1\big|^2\big]}\\
&\qquad\le\frac{2\alpha\delta^3}{\lambda N^2}\sqrt{\Big(\bar{L}+\frac{A^2}{N^2}\Big)\Big(A+C_1\Big)}\,,
\end{aligned}$$

where the second inequality follows by Cauchy-Schwarz, and the third from natural scaling of $(\mathbb{P}, \mathcal{L})$ and (4.12). If we take $\lambda = \gamma/(10\sqrt{N})$, and use the fact that $\alpha\delta^2 \le \varepsilon$, we obtain the first bound.

Now, let us turn to the second bound. Again, as the summands are positive, the supremum is attained at $t = M$, so that

$$\mathbb{P}_{x_0}\Big(\sup_{t\le M} \sum_{j=0}^{t-1} \delta^2\Big(\frac{A}{N^2} + \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}\ge \hat{L}\}}}{N}\Big)|m_j| > \lambda\Big) \le \mathbb{P}_{x_0}\Big(\delta^2 \sum_{j=0}^{M-1} \Big(\frac{A}{N^2} + \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}>\hat{L}\}}}{N}\Big) > \lambda\Big)$$
$$\le \mathbb{P}\Big(\delta^2 \sum_{j=0}^{M-1} \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}>\hat{L}\}}}{N} > \lambda - \frac{A\cdot\varepsilon}{N}\Big),$$

where we used here that since $\alpha\delta^2 \le \varepsilon$, we have $\delta^2 \sum_{j=0}^{M-1} \frac{A}{N^2} \le \frac{A\cdot\varepsilon}{N}$. By Markov's inequality,

$$\mathbb{P}_{x_0}\Big(\sum_{j=0}^{M-1} L_{j+1}\mathbf{1}_{\{L_{j+1}>\hat{L}\}} > \Lambda\Big) \le \frac{M}{\Lambda}\Big(\sup_j \mathbb{E}_{x_0}[L_{j+1}\mathbf{1}_{\{L_{j+1}>\hat{L}\}}]\Big)$$
$$\le \frac{M}{\Lambda}\sqrt{\sup_x \mathbb{E}[|\frac{1}{\sqrt{N}}\nabla H(x)|^4] \cdot \sup_x \mathbb{P}(|\frac{1}{\sqrt{N}}\nabla H(x)|^2 > \hat{L})}$$
$$\le \frac{\alpha N}{\Lambda}\frac{\bar{L}}{\hat{L}^{1+\iota/2}},$$

where in the last line we again used Markov's inequality along with the definition of $\bar{L}$ from (4.2). Choosing $\Lambda = \frac{N}{\delta^2}(\lambda - \frac{A\cdot\varepsilon}{N})$, with $\lambda = \gamma/(10\sqrt{N})$, we get that for N large, the left-hand side in (4.16) is bounded by

$$\frac{C\alpha\delta^2\sqrt{N}}{\hat{L}^{1+\iota/2}\gamma} \le \frac{C\varepsilon\sqrt{N}}{\hat{L}^{1+\iota/2}\gamma},$$

for some $C(A, \bar{L}) > 0$ as desired. □

We now sum (4.14) (or (4.13) if $k = 1$) over $t \ge 1$, and combine this with (4.12), (4.15), and (4.16) with the choice of $\hat{L} = N^{\frac{1}{2}-\frac{\iota}{4}}$. We see that for every $\gamma > 0$, $\eta < \eta_0$ and $K > 0$, if $k = 1$,

$$\lim_{N\to\infty} \inf_{x_0} \mathbb{P}_{x_0}\left(\begin{array}{ll} m_t \ge & \frac{4}{5}m_0 + \sum_{j=0}^{t-1} \frac{\delta a_{k-1}}{4N}\Big(1 - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}}\Big) \\ & -\frac{\delta}{N}\sum_{j=0}^{t-1} \nabla H^{j+1}(X_j)\cdot e_1 \end{array} \quad \forall t < \tau_0^- \wedge \tau_\eta^+ \right) = 1, \quad (4.17)$$

and if $k \ge 2$,

$$\lim_{N\to\infty} \inf_{x_0} \mathbb{P}_{x_0}\left(\begin{array}{ll} m_t \ge & \frac{4}{5}m_0 + \sum_{j=0}^{t-1} \frac{\delta a_{k-1}|m_j|^{k-1}}{4N}\Big(1 - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}\log N}\Big) \\ & -\frac{\delta}{N}\sum_{j=0}^{t-1} \nabla H^{j+1}(X_j)\cdot e_1 \end{array} \quad \forall t < \tau^-_{\gamma/(2\sqrt{N})} \wedge \tau_\eta^+ \right) = 1. \quad (4.18)$$

(Here we used the fact that $t < \tau_0^-$ to replace m_j with $|m_j|$ on the right hand sides). Notice that these limits hold uniformly over the choices of D and $\varepsilon = O(1)$.

4.3 Controlling the drift

We now turn to proving the following estimate on the drift term in (4.17)–(4.18). Let us begin by first recalling the following useful martingale inequality due to Freedman (1975), for situations where the almost sure bound on the martingale increments is much larger than the conditional variances: for a submartingale S_n with $\mathbb{E}[(S_n - S_{n-1})^2 \mid \mathcal{F}_{n-1}] \le K_2^{(n)}$ and $|S_n - S_{n-1}| \le K_1$ a.s., we have

$$\mathbb{P}\Big(S_t \le -\lambda\Big) \le \exp\Big(\frac{-\lambda^2}{\sum_{n\le t} K_2^{(n)} + \frac{1}{3}K_1\lambda}\Big)\,.$$

With this in hand, we can estimate the drift as follows.

Proposition 4.4. *If $k = 1$, for every $t \le M$, and every $K > 0$, we have*

$$\inf_{x_0} \mathbb{P}_{x_0}\Big(\sum_{j=0}^{t-1} \frac{\delta a_k}{4N}\Big(1 - \frac{L_{j+1}}{K\bar{L}}\Big) \ge \sum_{j=0}^{t-1} \frac{\delta a_k}{8N}\Big) \ge 1 - \frac{2}{K}\,.$$

Let $\hat{L} = N^{\frac{1}{2}-\frac{\iota}{4}}$. If $k \ge 2$, if $\alpha\delta^2 \le \varepsilon$ for some $\varepsilon = O(1)$, $\delta \le 1$, and α is of at most polynomial growth in N, then for every $\gamma > 0$,

$$\lim_{N\to\infty} \inf_{x_0} \mathbb{P}_{x_0}\Big(\sum_{j=0}^{t-1} \frac{\delta a_k |m_j|^{k-1}}{4N}\Big(1 - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}\log N}\Big) \ge -\frac{\gamma}{10\sqrt{N}} + \sum_{j=0}^{t-1} \frac{\delta a_k |m_j|^{k-1}}{8N}\,,\quad \forall t \le M\Big) = 1.$$

Moreover, this limit holds uniformly over choices of D and $\varepsilon = O(1)$.

Proof. We begin with the case of $k = 1$. Observe that for every $t \le M$, we have for every $x_0 \in \mathbb{S}^{N-1}$,

$$\mathbb{P}_{x_0}\Big(\sum_{j=0}^{t-1}\Big(\frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}} - \frac{1}{2}\Big) > 0\Big) \le \frac{2\sup_x \mathbb{E}[|\frac{1}{\sqrt{N}}\nabla H(x)|^2]}{K\bar{L}} \le \frac{2}{K}\,.$$

As this bound is independent of t and x_0, we obtain the desired. Thus with probability $1 - \frac{2}{K}$,

$$\sum_{j=0}^{t-1} \frac{L_j}{K\bar{L}} \le \frac{t}{2}\,, \qquad \text{and} \qquad \sum_{j=0}^{t-1} \frac{\delta a_k}{4N}\Big(1 - \frac{L_{j+1}}{K\bar{L}}\Big) \ge \sum_{j=0}^{t-1} \frac{\delta a_{k-1}}{8N}\,.$$

In the case $k \ge 2$, it suffices to prove the following: for $\hat{L} = N^{\frac{1}{2}-\frac{\iota}{4}}$ and α of at most polynomial growth with $\alpha\delta^2 \le \varepsilon$, we have for every $\gamma > 0$,

$$\lim_{N\to\infty} \sup_{x_0\in\mathbb{S}^{N-1}} \mathbb{P}_{x_0}\Big(\inf_{t\le M} \sum_{j=0}^{t-1} \frac{\delta a_k |m_j|^{k-1}}{4N}\Big(\frac{1}{2} - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}\log N}\Big) < -\frac{\gamma}{10\sqrt{N}}\Big) = 0\,, \qquad (4.19)$$

Fix any $x_0 \in \mathbb{S}_{N-1}$; everything that follows will be uniform in $x_0 \in \mathbb{S}^{N-1}$. To this end, observe that by a union bound, the desired probability in (4.19) is upper-bounded by

$$\mathbb{P}_{x_0}\Big(\inf_{t\le M} \sum_{j=0}^{t-1} \frac{\delta a_k |m_j|^{k-1}}{4N}\Big(\frac{1}{2} - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}<\hat{L}\}}}{K\bar{L}\log N}\Big) < -\frac{\gamma}{10\sqrt{N}}\Big)$$
$$\le M \sup_{t\le M} \mathbb{P}_{x_0}\Big(\sum_{j=0}^{t-1} \frac{\delta a_k |m_j|^{k-1}}{4N}\Big(\frac{1}{\log N} - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}\le\hat{L}\}}}{K\bar{L}\log N}\Big) < -\frac{\gamma}{10\sqrt{N}}\Big).$$

Recall the filtration $\mathcal{F}_j$. For every j, m_j is $\mathcal{F}_j$-measurable and

$$\mathbb{E}\Big[L_{j+1}\mathbf{1}_{\{L_{j+1}\le\hat{L}\}} \mid \mathcal{F}_j\Big] \le \sup_x \mathbb{E}\Big[\Big|\frac{1}{\sqrt{N}}\nabla H(x)\Big|^2\Big] \le \bar{L}$$

so that for $K \ge 1$, the sum

$$Z_t := \sum_{j=0}^{t-1} \frac{\delta a_k |m_j|^{k-1}}{4N}\Big(\frac{1}{\log N} - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}\le\hat{L}\}}}{K\bar{L}\log N}\Big)$$

is an $\mathcal{F}_t$-submartingale. Observe that $|Z_t - Z_{t-1}| \le \frac{\delta a_k}{8N}\Big(\frac{1}{\log N} \vee \frac{\hat{L}}{\bar{L}\log N}\Big)$ almost surely and that the conditional variances are bounded by

$$\mathbb{E}\Big[\Big(\frac{\delta a_k |m_j|}{8N}\Big)^2\Big(\frac{1+L_{j+1}}{\bar{L}\log N}\Big)^2\Big|\mathcal{F}_j\Big] \le 2\Big(\frac{\delta a_k}{8N\log N}\Big)^2\Big(1+\frac{\sup_x \mathbb{E}[|\frac{1}{\sqrt{N}}\nabla H(x)|^4]}{\bar{L}^2}\Big) \le \Big(\frac{\delta a_k}{4N\log N}\Big)^2$$

almost surely. Thus we may apply Freedman's inequality to obtain

$$\sup_{T\le M} \mathbb{P}_{x_0}\Bigg(\sum_{j=0}^{T-1} \frac{\delta a_k |m_j|^{k-1}}{4N}\Big(\frac{1}{2} - \frac{L_{j+1}\mathbf{1}_{\{L_{j+1}\le\hat{L}\}}}{K\bar{L}\log N}\Big) < -\frac{\gamma}{10\sqrt{N}}\Bigg)$$
$$\le \exp\left(\frac{-\gamma^2/(100\cdot N)}{M\Big(\frac{\delta a_k}{4N\log N}\Big)^2 + \frac{1}{3}\frac{\gamma}{\sqrt{N}}\frac{\delta a_k}{8N\log N}(1+\hat{L}/\bar{L})}\right).$$

Since $\alpha\delta^2 \le \varepsilon$, the first term in the denominator is $o(\frac{1}{N(\log \alpha N)})$ as α is not growing faster than polynomially in N. If we then take any $\hat{L}$ such that

$$\frac{\delta\gamma\hat{L}}{\sqrt{N}\log N} = o\Big(\frac{1}{\log \alpha N}\Big),$$

the entire bound is $o(\frac{1}{M})$, and a union bound over all $T \le M$ implies that the probability is $o(1)$ uniformly in the choice of x_0. The above criterion on $\hat{L}$ is satisfied with the choice $\hat{L} = N^{\frac{1}{2}-\frac{\iota}{4}}$ provided $\delta \le 1$ and that α is of at most polynomial growth in N, implying the desired (4.19). □

4.4 Controlling the directional error martingale

It remains to bound the effect of the directional error martingale from (1.8). Recall that for every initialization $X_0 = x_0 \in \mathbb{S}^{N-1}$, the point X_t is $\mathcal{F}_t$ measurable. This implies that for each t, the increment

$$M_t - M_{t-1} = -\frac{\delta}{N}\nabla H^t(X_{t-1})\cdot e_1\,,$$

has mean zero conditionally on $\mathcal{F}_{t-1}$—for every x, by definition of the sample error, $\mathbb{E}[H(x)] = 0$ and $\mathbb{E}[\nabla H(x)] = (0,...,0)$—so that M_t is indeed an $\mathcal{F}_t$-adapted martingale. To control the fluctuations of this martingale, we recall Doob's maximal inequality: if S_t is a submartingale, then for every $p \geq 1$,

$$\mathbb{P}\Big(\max_{t\leq T} S_t \geq \lambda\Big) \leq \frac{p\mathbb{E}[|S_T|^p]}{(p-1)\lambda^p}\,.$$

By (1.3), we then deduce the following.

Lemma 4.5. If C_1 is as in Definition 1.1 for every $r > 0$, we have

$$\sup_{T\leq M}\sup_{x_0\in\mathbb{S}^{N-1}} \mathbb{P}_{x_0}\left(\max_{t\leq T}\frac{1}{\sqrt{T}}\Big|\sum_{j=0}^{t-1}\nabla H^{j+1}(X_j)\cdot e_1\Big| \geq r\right) \leq \frac{2C_1}{r^2}\,, \tag{4.20}$$

Proof. For convenience, let $\tilde{M}_t = NM_t/\delta$. Observe that $\tilde{M}_t$ is a martingale with variance

$$\sup_{x_0}\mathbb{E}_{x_0}[\tilde{M}_t^2] = \sup_{x_0}\mathbb{E}_{x_0}\Big[\Big(\sum_{j=0}^{t-1}\nabla H^{j+1}(X_j)\cdot e_1\Big)^2\Big] \leq t\sup_x \mathbb{E}[(\nabla H(x)\cdot e_1)^2] \leq tC_1\,.$$

Thus, by Doob's maximal inequality, we have the desired bound,

$$\sup_{x_0}\mathbb{P}_{x_0}\Big(\sup_{t\leq T}|\tilde{M}_t| > r\sqrt{T}\Big) \leq \frac{2\sup_{x_0}\mathbb{E}_{x_0}[(\tilde{M}_T)^2]}{r^2T} \leq \frac{2C_1}{r^2}\,. \qquad \square$$

By (4.20), for every $d > 0$,

$$\sup_{T\leq M}\sup_{x_0}\mathbb{P}_{x_0}\Big(\sup_{t\leq T}\Big|\frac{\delta}{N}\sum_{j=0}^{t-1}\nabla H^{j+1}(X_j)\cdot e_1\Big| \geq \frac{\delta d\sqrt{T}}{10N}\Big) \leq \frac{200C_1}{d^2}\,. \tag{4.21}$$

4.5 Proof of Proposition 4.1

We are now in position to combine the above three bounds and conclude Proposition 4.1. Consider first the case $k \geq 2$. By Proposition 4.4 and (4.18) we have that for every $\gamma > 0$ and $\eta < \eta_0(\varrho, a_k, a_{k+1})$,

$$\lim_{N\to\infty}\inf_{x_0\in E_{\gamma/\sqrt{N}}}\mathbb{P}_{x_0}\left(\begin{array}{ll} m_t \geq & \frac{7}{10}m_0 + \sum_{j=0}^{t-1}\frac{\delta a_{k-1}m_j^{k-1}}{8N} \\ & -\frac{\delta}{N}\sum_{j=0}^{t-1}\nabla H^{j+1}(X_j)\cdot e_1 \end{array} \quad \forall t \leq \tau^-_{\gamma/(2\sqrt{N})}\wedge\tau^+_\eta\wedge M\right) = 1\,.$$

(We used here that since $x_0 \in E_{\gamma/\sqrt{N}}$, and $t \le \tau^-_{\gamma/(2\sqrt{N})}$, we have $\frac{\gamma}{10\sqrt{N}} \le \frac{m_0}{10}$ and $m_j = |m_j|$ deterministically.) Furthermore, if $D = D_N$, $\delta \le \bar\delta(k)$, and $\bar t$ are as in Proposition 4.1, for every $x_0 \in E_{\gamma/\sqrt{N}}$, if $T \le \bar t$, then

$$\frac{\delta D_N \sqrt{T}}{10N} \le \frac{\gamma}{10\sqrt{N}} \le \frac{m_0}{10}.$$

Thus, applying the directional error martingale bound (4.21) with $d = D$, we obtain the desired bound (observing that the $o(1)$ terms are uniform in D, K and $\varepsilon = O(1)$).

Suppose now that $k = 1$. By Proposition 4.4 and (4.17) we have that for every $K > 0$, every $\delta \le \bar\delta(1)$, and every N sufficiently large,

$$\inf_{t\le M}\inf_{x_0\in E_{\gamma/\sqrt{N}}} \mathbb{P}\Big(m_t \ge \frac{7}{10}m_0 + \sum_{j=0}^{t-1}\frac{\delta a_{k-1}}{8N} - \frac{\delta}{N}\sum_{j=0}^{t-1}\nabla H^{j+1}(X_j)\cdot e_1\,,\ \ \forall t \le \tau_0^- \wedge \tau_\eta^+\Big) \ge 1 - \frac{3}{K} - o(1)\,.$$

Controlling the directional error martingale by the same argument via (4.21) with $D = D_N$, we obtain (4.5) as desired.

5. Attaining weak recovery

We now turn to proving Theorem 3.1 and the recovery part of Theorem 3.3; we invite the reader to recall the definition of $\bar t$ from (4.4). The goal of this section will be to prove that the dynamics will have weakly recovered in some (possibly random) time before $\bar t \wedge M$.

Proposition 5.1. *Under the assumptions of Theorem 3.1, there exists $\eta_0(\varrho, a_k, a_{k+1}) > 0$ such that for every $\eta < \eta_0$, for every $\gamma > 0$, we have*

$$\lim_{N\to\infty}\inf_{x_0\in E_{\gamma/\sqrt{N}}} \mathbb{P}_{x_0}\Big(\tau_\eta^+ \le \bar t \wedge M\Big) = 1\,.$$

If $k = 1$ and we only assume $\alpha\delta^2 \le \varepsilon = O(1)$ and δ is sufficiently small, but order one, then

$$\varliminf_{N\to\infty}\inf_{x_0\in E_{\gamma/\sqrt{N}}} \mathbb{P}_{x_0}\Big(\tau_\eta^+ \le \bar t \wedge M\Big) \ge 1 - C\Big(\delta + \frac{1}{D^2}\Big)\,,$$

for some constant $C(C_1, a_1, a_2) > 0$.

The proposition immediately implies Theorem 3.1.

5.1 Proof of Proposition 5.1

We analyze the difference inequalities of Proposition 4.1 when $k = 1$, $k = 2$, and $k \ge 3$ separately. Observe that in all cases, the right-hand side of the difference inequality is increasing, so that since $x_0 \in E_{\gamma/\sqrt{N}}$, we have $\tau^-_{\gamma/(2\sqrt{N})} \ge \tau_\eta^+ \wedge \bar t$ necessarily, and we may drop the requirement $t \le \tau^-_{\gamma/(2\sqrt{N})}$ in all cases of (4.5)–(4.6).

Linear regime: $k = 1$**.**

The right-hand side of the difference inequality in (4.5) is non-decreasing and greater than η at time

$$t_* := \lceil \frac{8\eta N}{\delta a_k} \rceil.$$

As such, $(\tau_\eta^+ \wedge \bar{t} \wedge M) \leq t_\star$ necessarily, and as long as $t_* < \bar{t} \wedge M$, we will have the desired

$$\lim_{N\to\infty} \inf_{x_0 \in E_{\gamma/\sqrt{N}}} \mathbb{P}_{x_0}\Big(\tau_\eta^+ \leq \bar{t} \wedge M\Big) = 1\,.$$

In order for $t_* < \bar{t} \wedge M$, we need, for a sequence $D_N \uparrow \infty$ arbitrarily slowly,

$$\lceil \frac{8\eta}{\delta a_k} \rceil \leq \frac{\gamma^2}{D_N^2 \delta^2} \wedge \alpha.$$

We see that this inequality is satisfied for the choices of α, δ as in the statement of Theorem 3.1, since $\delta = o(1)$ (as we can choose $D_N \to \infty$ arbitrarily slowly, so that $D_N^2 \delta = o(1)$) and $\delta\alpha \uparrow \infty$.

In the case of $\alpha = O(1)$, we can take D sufficiently large, and subsequently take δ sufficiently small (depending on the D) such that the above inequality is satisfied. □

Quasilinear regime: $k = 2$**.**

In this case, we may use the discrete Gröwnwall inequality: suppose that (m_t) is any sequence such that for some $a, b \geq 0$

$$m_t \geq a + \sum_{\ell=0}^{t-1} b m_\ell \qquad \text{then} \qquad m_t \geq a(1+b)^t\,. \tag{5.1}$$

Define the lower-bounding function,

$$g_2(t) := \frac{m_0}{2} \exp\Big(\frac{\delta a_k}{8N} t\Big)\,.$$

Applying the discrete Grönwall inequality to (4.6), we obtain that for every $\gamma > 0$,

$$\lim_{N\to\infty} \inf_{x_0 \in E_{\gamma/\sqrt{N}}} \mathbb{P}_{x_0}\Big(m_t \geq g_2(t) \quad \text{for all } t \leq \tau_\eta^+ \wedge \bar{t} \wedge M\Big) = 1$$

Since $g_2(t) \geq \eta$ for all

$$t \geq t_* := \Big\lceil \frac{8N}{\delta a_k} (\log \frac{2}{m_0} + \log \eta) \Big\rceil\,,$$

we see that as long as $t_* \leq \bar{t} \wedge M$, we will have $\lim_{N\to\infty} \inf_{x_0 \in E_{\gamma/\sqrt{N}}} \mathbb{P}_{x_0}(\tau_\eta^+ \leq \bar{t} \wedge M) = 1$. The criterion $t_* \leq \bar{t} \wedge M$ is satisfied using the facts that $\delta = o(\frac{1}{\log N})$ (as we can choose D_N to go to ∞ arbitrarily slowly) and $\frac{\alpha\delta}{\log N} \uparrow \infty$. □

Polynomial regime: $k \geq 3$.

Observe the following discrete analogue of the Bihari-LaSalle inequality: suppose that (m_t) is a sequence satisfying, for some $k > 2$ and $a, b > 0$

$$m_t \geq a + \sum_{\ell=0}^{t-1} b m_\ell^{k-1} \qquad \text{then} \qquad m_t \geq \frac{a}{(1 - ca^{k-2}t)^{\frac{1}{k-2}}}. \tag{5.2}$$

For the reader's convenience, we provide a proof in Appendix C.

Applying (5.2), we obtain for every $t \leq \tau_\eta^+ \wedge \bar{t} \wedge M$,

$$m_t \geq \frac{m_0}{\left(1 - \frac{\delta a_k}{8N}(k-2)m_0^{k-2}t\right)^{\frac{1}{k-2}}} =: g_k(t)\,.$$

In particular, $g_k(t) \geq \eta$ provided

$$\eta\left(1 - \frac{\delta a_k}{8N}(k-2)\frac{\gamma^{k-2}}{N^{\frac{k-2}{2}}}t\right) = o\Big(\frac{\gamma^{k-2}}{N^{\frac{k-2}{2}}}\Big)\,.$$

As such, if for some K sufficiently large,

$$t \geq t_* = \Big\lceil \frac{8N}{\delta a_k (k-2)\gamma^{(k-2)/2}} N^{\frac{k-2}{2}} \Big(1 - \frac{K}{N^{(k-2)/2}}\Big) \Big\rceil\,,$$

then as long as $t_* \leq \bar{t} \wedge M$, we will have $\lim_{N\to\infty} \inf_{x_0 \in E_{\gamma/\sqrt{N}}} \mathbb{P}_{x_0}\big(\tau_\eta^+ \leq \bar{t} \wedge M\big) = 1$. The criterion $t_* \leq \bar{t} \wedge M$ is satisfied using the facts that $\delta = o(N^{-\frac{k-2}{2}})$ (as we can choose D_N to go to ∞ arbitrarily slowly) and $\frac{\alpha\delta}{N^{(k-2)/2}} \uparrow \infty$. □

6. Strong recovery and the descent phase

We begin this section by proving a law of large numbers for the trajectory in the descent phase, Theorem 3.2. We then combine it with Theorem 3.1 to conclude the proof of Theorem 1.3.

Proof of Theorem 3.2. Recall $m_t = m(X_t)$ and r_t and analogously define $\overline{m}_t := m(\overline{X}_t)$,

$$\overline{m}_t = \frac{1}{\bar{r}_t}\Big(\overline{m}_{t-1} - \frac{\delta}{N}\nabla\Phi(\overline{X}_{t-1}) \cdot e_1\Big) \qquad \text{where} \qquad \bar{r}_t := \sqrt{1 + \delta^2 |\nabla\Phi(\overline{X}_{t-1})|^2/N^2}$$

As shown in (4.8), $|r_t - 1| \leq \delta^2(\frac{A}{N^2} + \frac{L_t}{N})$ where $L_t = |\nabla H^t(X_{t-1})/\sqrt{N}|^2$. By similar reasoning, $|\bar{r}_t - 1| \leq \frac{A\delta^2}{N^2}$. At the same time,

$$\begin{aligned}\Big|m_{t-1} - \frac{\delta}{N}\nabla\Phi(X_{t-1}) \cdot e_1 - \frac{\delta}{N}\nabla H^t(X_{t-1}) \cdot e_1\Big| \\ \leq 1 + \sup_x \frac{\delta}{N}|\nabla\Phi(x) \cdot e_1| + \frac{\delta}{N}|\nabla H^t(X_{t-1}) \cdot e_1| \leq 1 + \frac{\delta\sqrt{A}}{N} + \frac{\delta\sqrt{L_t}}{\sqrt{N}}\,.\end{aligned}$$

As such, we have the bound

$$\left| m_t - \left(m_{t-1} - \frac{\delta}{N} \nabla \Phi(m_{t-1}) \cdot e_1 - \delta \frac{1}{N} \nabla H^t(X_{t-1}) \cdot e_1 \right) \right| \leq \frac{2\delta^2 (1 \vee L_t)(1 \vee \frac{\delta \sqrt{L_t}}{\sqrt{N}})}{N} .$$

Iterating the above bound, we see that

$$\sup_{t \leq M} \left| m_t - \left(m_0 - \sum_{\ell=0}^{t-1} \frac{\delta}{N} \nabla \Phi(m_\ell) \cdot e_1 - \frac{\delta}{N} \sum_{\ell=0}^{t-1} \nabla H^{\ell+1}(X_\ell) \cdot e_1 \right) \right| \leq \sum_{\ell=0}^{M-1} \frac{2\delta^2 (1 \vee L_\ell)(1 \vee \frac{\delta \sqrt{L_\ell}}{\sqrt{N}})}{N} .$$

Consider the probability that the quantity on the right-hand side is large: for every $\varepsilon > 0$,

$$\mathbb{P}\Big(\sum_{\ell=0}^{M-1} \frac{\delta^2 (L_\ell + \frac{\delta \sqrt{L_\ell}}{\sqrt{N}} + \frac{\delta L_\ell^{3/2}}{\sqrt{N}})}{N} > \varepsilon \Big) \leq \varepsilon^{-1} \alpha \delta^2 \sup_\ell \mathbb{E}\Big[L_\ell + \frac{\delta \sqrt{L_\ell}}{\sqrt{N}} + \frac{\delta L_\ell^{3/2}}{\sqrt{N}} \Big] .$$

Recalling from (4.2) that $\sup_\ell \mathbb{E}[L_\ell] \vee \mathbb{E}[L_\ell^2] \leq \bar{L}$, we see that each of the expectations above are bounded by $\bar{L}$, whereby as long as $\alpha \delta^2 = o(1)$, this is $o(1)$ for all fixed ε.

As such, it suffices to consider the linearization

$$\mathsf{m}_t := m_0 - \sum_{\ell=0}^{t-1} \frac{\delta}{N} \nabla \Phi(m_\ell) \cdot e_1 - \frac{\delta}{N} \sum_{\ell}^{t-1} \nabla H^{\ell+1}(X_\ell) \cdot e_1$$

as for every ε, with probability $1 - o(1)$, we have $\sup_{\ell \leq M} |\mathsf{m}_\ell - m_\ell| \leq \varepsilon$.

By the same reasoning, as long as $\alpha \delta^2 = o(1)$, for every ε, for N large enough, we have deterministically $|\bar{\mathsf{m}}_\ell - \bar{m}_\ell| \leq \varepsilon$, where $\bar{\mathsf{m}}$ is the linearized population dynamics,

$$\bar{\mathsf{m}}_t = m_0 - \sum_{\ell=0}^{t-1} \frac{\delta}{N} \nabla \Phi(m_\ell) .$$

With the above in hand, it clearly suffices to show that

$$\sup_{\ell \leq M} |\bar{\mathsf{m}}_\ell - \mathsf{m}_\ell| \to 0 \qquad \text{in } \mathbb{P}\text{-prob} . \tag{6.1}$$

Towards that, let us control the effect of the directional error martingale for all times. Recall from Doob's maximal inequality as in (4.21) that for every $\lambda > 0$,

$$\mathbb{P}\Big(\sup_{t \leq M} \frac{\delta}{N} \Big| \sum_{\ell=0}^{t-1} \nabla H^{\ell+1}(X_\ell) \cdot e_1 \Big| > \lambda \Big) \leq \frac{C_1 \alpha \delta^2}{\lambda^2 N} = o(N^{-1}) . \tag{6.2}$$

To show (6.1), consider the probability that the supremum is greater than some $\gamma > 0$, and split the supremum into a one over $\ell \leq T\delta^{-1}N$ and one over $T\delta^{-1}N \leq \ell \leq M$ for a T to be chosen sufficiently large. For the former, fix any T, and recall that $\nabla \Phi(m) \cdot e_1 = \phi'(m)(1 - m^2)$. As ϕ'_N are uniformly C^1 on $[0, 1]$, there exists K such that uniformly over N,

$$\sup_{x,y \in [0,1]} |\phi'(x)(1 - x^2) - \phi'(y)(1 - y^2)| \leq K|y - x| .$$

We obtain from this that for every $\varepsilon > 0$, on the event $\{\sup_{\ell \le M} |m_\ell - \mathsf{m}_\ell| \vee |\bar{m}_\ell - \bar{\mathsf{m}}_\ell| < \varepsilon\}$,

$$\begin{aligned}|\bar{\mathsf{m}}_t - \mathsf{m}_t| &\le \sum_{\ell=0}^{t-1} \frac{\delta}{N} |\nabla\Phi(\bar{m}_\ell) \cdot e_1 - \nabla\Phi(m_\ell) \cdot e_1| + \frac{\delta}{N} \Big| \sum_{\ell=0}^{t-1} \nabla H^{\ell+1}(X_\ell) \cdot e_1 \Big| \\ &\le \sum_{\ell=0}^{t-1} \Big[\frac{\delta K}{N} |\bar{\mathsf{m}}_\ell - \mathsf{m}_\ell| + \frac{2\delta K \varepsilon}{N} \Big] + \frac{\delta}{N} \Big| \sum_{\ell=0}^{t-1} \nabla H^{\ell+1}(X_\ell) \cdot e_1 \Big| .\end{aligned}$$

Combined with (6.2), as long as $\alpha\delta^2 = o(1)$, with probability $1 - o(1)$, we have for all $t \le T\delta^{-1}N$,

$$|\bar{\mathsf{m}}_t - \mathsf{m}_t| \le (2TK + 1)\varepsilon + \frac{\delta K}{N} \sum_{\ell=0}^{t-1} |\bar{\mathsf{m}}_\ell - \mathsf{m}_\ell| ,$$

which by the discrete Gronwall inequality, implies that for every $\varepsilon > 0$, with probability $1 - o(1)$,

$$\sup_{t \le T\delta^{-1}N} |\mathsf{m}_t - \bar{\mathsf{m}}_t| \le (\frac{2MK}{N}\delta + 1)\varepsilon e^{KT} .$$

For each $\gamma > 0$, there exists $\varepsilon(K, T) > 0$ such that the above is at most $\gamma/5$.

Now let $T = T(\gamma)$ be such that

$$\sup_{T\delta^{-1}N \le t \le N} |\bar{\mathsf{m}}_t - 1| < \frac{\gamma}{5} ;$$

this T exists and is $O(1)$ by Assumption A. In that case, using again Assumption A to note that $\nabla\Phi(m) \ge 0$ while $m \ge 0$, we get

$$\sup_{T\delta^{-1}N \le t \le M} |\mathsf{m}_t - \bar{\mathsf{m}}_t| < \frac{2\gamma}{5} + |\mathsf{m}_{T\delta^{-1}N} - \bar{\mathsf{m}}_{T\delta^{-1}N}| + \frac{\delta}{N} \Big| \sum_{\ell=T\delta^{-1}N}^{M} \nabla H^{\ell+1}(X_\ell) \cdot e_1 \Big|$$

By the first part of (6.1), and the bound of (6.2) applied to $\gamma/5$, we see that the probability that the above is greater than γ is also $o(1)$. Together these yield (6.1). □

6.1 Proof of Theorem 1.3 and item (a) of Theorem 3.3

Let (α_N, δ_N) be as in Theorem 1.3, fix any $\varepsilon > 0$, and consider the $\mu_N \times \mathbb{P}$-probability that $|m(X_M) - 1| > \varepsilon$; we need to show this goes to zero as $N \to \infty$. By the Poincaré Lemma, for every $\zeta > 0$, there exists γ such that for all N sufficiently large,

$$\mu_N(m_0 < \gamma/\sqrt{N}) < \zeta/3 .$$

Let us now suppose that X_0 is such that $m_0 \ge \gamma/\sqrt{N}$. By Theorem 3.1, there exists $\eta_0 > 0$ such that for any such X_0, for N sufficiently large, we have

$$\mathbb{P}(\tau_\eta^+ < M_1) < \zeta/3$$

for $M_1 = M/2$ (notice that the criteria of the theorem apply equally whether we take α or $\alpha/2$).

Observe that by the Markov property, conditionally on the stopping time τ_η^+ and the value $X_{\tau_\eta^+}$. we have the distributional equality

$$\mathbb{P}_{x_0}(X_{\tau_\eta^+ + s} \in \cdot \mid \tau_\eta^+, X_{\tau_\eta^+}) \stackrel{d}{=} \mathbb{P}_{X_s}(X_s \in \cdot)$$

As such, for X_0 satisfying $m(X_0) \geq \gamma/\sqrt{N}$, we have

$$\mathbb{P}_{X_0}(m(X_M) \leq 1-\varepsilon) \leq \sup_{M_1 \leq s \leq M} \sup_{y_0 : m(y_0) \geq \eta} \mathbb{P}_{y_0}(m(X_s) \leq 1-\varepsilon) + \frac{2\zeta}{3}\,.$$

Using again the fact that the criteria in Theorem 3.2 on α, δ apply equally well if we replace α with $\alpha/2$, we see that for N sufficiently large, the right-hand side above is bounded by

$$\sup_{M_1 \leq M_2 \leq M} \zeta + \mathbf{1}\{m(\bar{X}_{M_2}) < 1 - \frac{\varepsilon}{2}\}\,.$$

Using the Assumption A, the indicator function above is 0 as long as M_2 is a sufficiently large constant depending on ε, which it necessarily is since it is at least $M/2$ and $M = \omega(1)$ by the assumptions of Theorem 1.3.

The $k = 1$ case with linear sample complexity of item (a) of Theorem 3.3 follows naturally by noticing that α could have been taken to be a large enough constant in the above at the expense of ε and ζ being small but order one. □

7. Online SGD does not recover with smaller sample complexity

Here, we prove an accompanying refutation theorem, showing that if α is smaller than in Theorem 3.1 (up to factors of $\log N$), there is not enough time for the online SGD to weakly recover in one pass through the M samples.

Proof of Theorem 1.4 and item (b) of Theorem 3.3. We will in fact prove the following stronger refutation: for every $\eta > 0$, and every $\Gamma > 0$,

$$\sup_{x : m(x) < \Gamma N^{-1/2}} \mathbb{P}_x\Big(\sup_{t \leq M} m(X_t) > \eta\Big) = o(1)\,. \tag{7.1}$$

This implies Theorem 1.4 because for every ε there exists a Γ such that $\mu_N(m(x) > \Gamma N^{-1/2}) < \varepsilon$ for all N sufficiently large, by the Poincaré lemma.

Recall that the radius $r_t = |\tilde{X}_t|$ satisfies $r_t \geq 1$ deterministically, so that by (4.9), as long as $m_t > 0$, we have

$$m_t \leq m_{t-1} - \frac{\delta}{N}\nabla\Phi(X_{t-1}) \cdot e_1 - \frac{\delta}{N}\nabla H^t(X_{t-1}) \cdot e_1\,.$$

From this, we can first observe the following crude bound on the maximal one-step change $m_t - m_{t-1}$. If $m_{t-1} < 0$ but $m_t > 0$, then we have the above inequality without the m_{t-1},

and furthermore we can put absolute values on each of those terms. Recall that for every $r > 0$,

$$\sup_{x \in \mathbb{S}^{N-1}} \mathbb{P}\Big(\Big| \frac{\delta}{N} \nabla H(x) \cdot e_1 \Big| > r \Big) \leq \frac{\delta^2 C_1}{N^2 r^2}$$

from which, for every sequence $d_N > 0$ going to infinity arbitrarily slowly, and every $t > 0$,

$$\sup_{x : m(x) \leq 0} \mathbb{P}\Big(m_t > d_N N^{-1/2} \mid X_{t-1} = x \Big) \leq \frac{C_1 \alpha \delta^2}{M d_N^2} = O(\frac{\alpha \delta^2}{M d_N^2})$$

By the Markov property of the online SGD, by a union bound over the M samples and using the fact that $\alpha \delta^2 = o(1)$, we may, with probability $1 - o(\alpha \delta^2 / (d_N^2))$, work under the event that $|\delta(\nabla H^{j+1}(X_j) \cdot e_1)/N| < d_N / \sqrt{N}$ for every j.

With that in mind, summing up the finite difference inequality over all time, we see that for every $t \leq \tau_0^- \wedge \tau_\eta^+$, we have

$$m_t \leq m_0 + \frac{\delta}{N} \sum_{j=0}^{t-1} \frac{3a_k}{2N} m_j^{k-1} - \frac{\delta}{N} \sum_{j=0}^{t-1} \nabla H^{j+1}(X_j) \cdot e_1 \, .$$

Recall from Doob's maximal inequality, that

$$\sup_{x_0} \mathbb{P}_{x_0}\Big(\sup_{1 \leq s \leq M} \Big| \frac{\delta}{N} \sum_{j=0}^{s-1} \nabla H^{j+1}(X_j) \cdot e \Big| > r \Big) \leq \frac{2\alpha \delta^2 C_1}{N r^2} \, . \tag{7.2}$$

Now define a sequence of excursion times as follows: for $i \geq 1$, we let $\tau_0 = 0$ and

$$\tau_{2i-1} := \inf\{ t \geq \tau_{2i-2} : m(X_t) \leq 0 \} \qquad \tau_{2i} := \inf\{ t \geq \tau_{2i-1} : m(X_t) > 0 \}$$

Taking $r = d_N N^{-1/2}$ in (7.2), it follows that

$$\inf_{x : m(x) < \Gamma N^{-1/2}} \mathbb{P}\Big(m_t \leq m_{\tau_{2i}} + \frac{d_N}{\sqrt{N}} + \frac{\delta}{\sqrt{N}} \sum_{j=0}^{t-1} \frac{3a_k}{2\sqrt{N}} m_j^{k-1}, \quad \forall i, \forall t \leq [\tau_{2i}, \tau_{2i+1} \wedge \tau_\eta^+] \Big)$$
$$= 1 - O(\frac{\alpha \delta^2}{d_N^2}) \, .$$

We next claim that the inequality above implies, deterministically, that through every excursion (i.e., for all i), we have $\tau_\eta^+ > \tau_{2i+1} \wedge M$. First of all, recall that we have restricted to the part of the space on which $m_{\tau_{2i}}$ is always less than $d_N N^{-1/2}$. Then, if we have the inequality in the probability above, by the discrete Gronwall inequality (5.1) when $k = 2$ and the discrete Bihari–LaSalle inequality (5.2) when $k > 2$, we have for some $c = c(k) > 0$,

$$m_t \leq \begin{cases} 2 d_N N^{-1/2} + \frac{2\delta a_k}{N} t & k = 1 \\ 2 d_N N^{-1/2} \exp\left(\frac{2\delta a_k}{N} t \right) & k = 2 \\ 2 d_N N^{-1/2} \big(1 - c \delta a_k d_N^{k-2} N^{-\frac{k-2}{2} - 1} t \big)^{-1/(k-2)} & k \geq 3 \end{cases}$$

For N sufficiently large, the right-hand side above is smaller than η for all $t \le \tilde{t}$ where

$$\tilde{t} = \begin{cases} \epsilon\eta\delta^{-1}N & k = 1 \\ \frac{\epsilon}{\delta} N \log N & k = 2 \\ d_N^{-\epsilon}\delta^{-1}N^{1+\frac{k-2}{2}} & k \ge 3 \end{cases}$$

for some $\epsilon > 0$ sufficiently small depending on k, a_k, a_{k+1}, C_1 (as long as d_N is growing slower than $N^{\frac{1}{2}-\zeta}$ for some $\zeta > 0$ say). Recall the restrictions on α and let b_N be a sequence going to infinity arbitrarily slowly.

1. If $k = 1$, we can choose d_N to be any diverging sequence. Since $\alpha = o(1)$ and $\delta = O(1)$, the above probabilities were all $1 - o(1)$, and for every fixed $\eta > 0$, we have $\tilde{t} \ge M$.

2. If $k = 2$, we can choose d_N diverging as a power in N, say $N^{1/4}$ such that for all $\alpha\delta^2 = o(N^{1/2})$, and in particular $\delta = O(1)$, the above probabilities $1 - O(\alpha\delta^2/d_N^2)$ were all $1 - o(1)$ *and* $\tilde{t} \ge M$.

3. If $k > 2$, we can choose d_N to be a sequence diverging sufficiently slowly. Then for all $\alpha\delta^2 = O(1)$, the above probabilities were all $1 - o(1)$ *and* $\tilde{t} \ge M$ (where to see this, we combined the inequalities $\sqrt{\alpha} = o(N^{(k-2)/2})$ and $\sqrt{\alpha} = O(\delta^{-1})$).

In both cases, we conclude that necessarily $\tau_\eta^+ > \tau_{2i+1}$, and therefore for every $\eta > 0$,

$$\inf_{x : m(x) < \Gamma N^{-1/2}} \mathbb{P}\Big(m_t < \eta \quad \text{for all } t \in \bigcup_i [\tau_{2i-2}, \tau_{2i-1}] \Big) = 1 - o(1)$$

At the same time, deterministically, for all $t \in \bigcup_i [\tau_{2i-1}, \tau_{2i}]$ we have $m(X_t) \le 0 < \eta$, implying (7.1).

In order to conclude part (b) of Theorem 3.3 when $k = 1$ we reason as above, taking d_N and α to be sufficiently large constants together, then subsequently taking δ to be a sufficiently small constant so that the probabilities of order $\alpha\delta^2/d_N^2$ above can be made arbitrarily small. □

Acknowledgments

The authors thank the anonymous referees for their detailed comments and suggestions. The authors thank Y. M. Lu for interesting discussions. G.B.A. thanks A. Montanari, Y. Le Cun, and L. Bottou for interesting discussions at early stages of this project. A.J. thanks S. Sen for helpful comments on this work. R.G. thanks the Miller Institute for Basic Research in Science for their support. A.J. acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [RGPIN-2020-04597, DGECR-2020-00199].

Appendix A. Useful bounds on norms of random vectors

Before getting to the deferred proofs of Section 2, we give a few useful inequalities we will need. Recall that if $(Y_\ell)_{\ell=1}^m$ is a collection of m non-negative random variables with finite p-th moment, i.e., $\max_\ell \mathbb{E}Y_\ell^p = K$, and $(p_\ell)_{\ell=1}^m$ is such that $\sum p_\ell = p$, then

$$\mathbb{E}\big[Y_1^{p_1}\cdots Y_m^{p_m}\big] \le K. \tag{A.1}$$

This follows by noting that by Young's inequality and Jensen's inequality we have that

$$\mathbb{E}\Big[(Y_1^{\frac{p_1}{p}}\cdots Y_m^{\frac{p_m}{p}})^p\Big] \le \mathbb{E}\Big[\Big(\sum_{\ell=1}^m \frac{p_\ell}{p} Y_\ell\Big)^p\Big] \le \sum \frac{p_\ell}{p}\mathbb{E}Y_\ell^p \le C\,,$$

for some $C(K,p)$.

Using the above we can easily obtain the following bound on the moments of the norm of a random vector. Suppose that X is a centered random vector in $\mathbb{R}^n$ whose entries have uniformly bounded $2k$-th moment, i.e., $\sup_i \mathbb{E}[X_i^{2k}] < K$ for some $k \ge 1$. Then using (A.1) for $1 \le q \le k$, there is $C = C(K,q) > 0$ such that

$$\mathbb{E}[\|X\|_2^{2q}] = \mathbb{E}\Big[\Big(\sum X_i^2\Big)^q\Big] \le \mathbb{E}\Big[\sum_{i_1,\dots,i_q} X_{i_1}^2\cdots X_{i_q}^2\Big] \le CN^q\,. \tag{A.2}$$

Lemma A.1. Suppose that X is a centered random vector in $\mathbb{R}^n$ whose entries have uniformly bounded $2k$-th moment, i.e., $\sup_i \mathbb{E}[X_i^{2k}] < K$ for some $k \ge 1$. Then for $1 \le q \le 2k$ there is $C = C(K,q) > 0$ such that

$$\mathbb{E}[|X\cdot w|^q] \le CK^{\frac{q}{2k}}||w||_2^q \quad \forall w \in \mathbb{R}^n. \tag{A.3}$$

Proof. First note the following symmetrization inequality (see e.g., Exercise 6.4.5 of Vershynin (2019)): if (ϵ_i) are i.i.d. Rademacher random variables, then since X is centered, we have

$$\mathbb{E}\big[|X\cdot w|^q\big] \le 2^q\mathbb{E}\Big[|\sum_i \epsilon_i X_i w_i|^q\Big]\,.$$

Thus it suffices to assume that the entries of X are jointly symmetric in the sense that that the law of X is invariant under the sign change of any given entry. Furthermore, by Jensen's inequality it suffices to consider the case that $q = 2k$.

We begin with a direct computation:

$$\mathbb{E}[|X\cdot w|^{2k}] = \sum_{i_1,\dots,i_{2k}} w_{i_1}\cdots w_{i_{2k}}\cdot\mathbb{E}[X_{i_1}\cdots X_{i_{2k}}].$$

As the entries of X are symmetric, any summand corresponding to an index that appears an odd number of times must be zero. In particular, we have the following. Let $\mathcal{P} = \{p_1,\cdots,p_{2k}\}$ be a partition of the integer $2k$. We say $\mathcal{P}$ is *even* if the p_ℓ are even. By (A.1), all of these moments are of the form $\mathbb{E}X_{j_1}^{p_1}\cdots X_{j_m}^{p_m} \le K$ for some even partition $\mathcal{P}$. As such, the above sum is bounded by

$$\sum_{\mathcal{P}\ even} c(\mathcal{P})K\prod_{p\in\mathcal{P}}||w||_p^p \le \sum_{\mathcal{P}\ even} c(\mathcal{P})K||w||_2^{2k},$$

where here $c(\mathcal{P})$ is the number of groupings of $2k$ items in to m groups of sizes $p_1, \dots, p_m$. Here we used that since $\mathcal{P}$ is even, $p \geq 2$ for any $p \in \mathcal{P}$ so that $||w||_p \leq ||w||_2$, and that $\sum p_\ell = 2k$ as $\mathcal{P}$ is a partition. In particular, as $\max c(\mathcal{P})$, and the number of even partitions of $2k$, each depend only on k we get

$$\mathbb{E}[|X \cdot w|^{2k}] \leq C \cdot K \cdot ||w||_2^{2k},$$

for some $C = C(K, k) > 0$ as desired. □

Appendix B. Deferred proofs from Section 2

In this section, we verify that the various examples of Section 2 satisfy Assumptions A–B.

B.1 Proof of Proposition 2.1

We begin with the to the proof of (2.4). Fix f as in the statement of the theorem. Since f' is of at most polynomial growth, so is f. In particular, $f \in L^2(\varphi)$ where φ is the standard Gaussian measure on $\mathbb{R}$. Recall (2.2), and let $C_\epsilon = \mathbb{E}[\epsilon^2]$. By rotational invariance of the Gaussian ensemble, we may take $v_0 = e_1$ there. Furthermore, the x-dependence of the population loss depends only on x through $x \cdot e_1$, so that

$$\Phi(x) = \mathbb{E}\left[\left(f\left(a_1 m(x) + a_2\sqrt{1-(m(x))^2}\right) - f(a_1)\right)^2\right] + C_\epsilon \tag{B.1}$$

where $a_1, a_2 \sim \mathcal{N}(0,1)$ are independent.

To compute this expectation, recall the following. For $s \in [-1,1]$, consider the Noise operator $T_s : L^2(\varphi) \to L^2(\varphi)$

$$T_s f(x) = \mathbb{E}[f(xs + \sqrt{1-s^2}a_2)].$$

Recall that the Hermite polynomials satisfy $T_s h_k = s^k h_k$ (Ledoux and Talagrand, 2011). (Usually, this is stated only for $s \geq 0$. To see this for $s < 0$, simply note that $T_s h_k(x) = T_{|s|}h_k(-x) = (-1)^k T_{|s|} h_k(x)$.) Consequently we have that

$$\Phi(x) = ||f||^2_{L^2(\varphi)} - 2\langle f, T_m f\rangle_{L^2(\varphi)} + C_\epsilon = 2\sum_j u_j^2 - 2\sum_j u_j^2 m^j + C_\epsilon = \phi_f(m)$$

as desired.

Assumption A is immediate from (2.4). It remains to show that the pair satisfies Assumption B. To this end, recall that since f' is of at most polynomial growth, we have $f \in H^1(\varphi)$, where H^1 is the Sobolev space with norm

$$||f||^2_{H^1(\varphi)} := \int f(z)^2 + f'(z)^2 d\varphi(z) = \sum_{j\geq 0}\left(1+j^2\right)u_j^2 < \infty.$$

We now turn checking (1.3)-(1.4). First note that for every x,

$$\nabla\Phi(x) = -2\sum j u_j^2 m^{j-1} \cdot \nabla m.$$

Thus, for every x, $|\nabla\Phi(x)| \le 2\sum j u_k^2 < 2||f||_{H^1}^2 < \infty$, where we used here that $|m| \le 1$. Similarly for every x, $\nabla\Phi(x)\cdot e_1 = \nabla\Phi(x)\cdot\nabla m = -2\sum j u_j^2 m^{k-1}$ so that $|\nabla\Phi\cdot e_1|^2 < 4||f||_{H^1}^4 < \infty$.

Thus it suffices to check (1.3)-(1.4) for $\mathcal{L}$ itself. Here we have that if we let $\pi_x(v) = v - (v\cdot x)x$ be the projection on to the tangent space at x, we have

$$\nabla\mathcal{L}(x;a,\epsilon) = 2(f(a\cdot x)+\epsilon - f(a_1))f'(a\cdot x)\pi_x a\,.$$

Thus, by Hölder's inequality and the fact that $|x+y|^p \le 2^p(|x|^p+|y|^p)$ for $p \ge 1$, for any $q \ge 1$, if we take $\gamma > 0$ and $r = \frac{2q(q+\gamma)}{\gamma}$, we have that

$$\mathbb{E}|\nabla\mathcal{L}(x)|^q \le 2^q\Big(\mathbb{E}[(f(a\cdot x)-f(a_1))^{q+\gamma}]^{\frac{q}{q+\gamma}} + \mathbb{E}[|\epsilon|^{q+\gamma}]^{\frac{q}{q+\gamma}}\Big)\cdot\mathbb{E}[|f'(a\cdot x)|^r]^{\frac{q}{r}}\cdot\mathbb{E}[||a||_2^r]^{q/r}.$$

The expectations involving f, f' are bounded since f and f' are of at most polynomial growth and $a\cdot x$ is Gaussian for every x; the expectation involving ϵ is bounded as long as ϵ has finite p-th moment for some $p > q$. The last term is bounded by (A.2) since a is a standard Gaussian vector. Taking $q = 4+\iota$ yields (1.4) for ι small enough, after recalling our assumption that ϵ has finite $4+\delta$-th moment.

For (1.3), observe that since f and f' are of at most polynomial growth,

$$\mathbb{E}|\nabla\mathcal{L}\cdot e_1|^2 = \mathbb{E}\Big[|f(a_1 m + a_2\sqrt{1-m^2}) - f(a_1)+\epsilon|^2|f'(a_1 m + a_2\sqrt{1-m^2})|^2|a_1|^2\Big] \le C\,,$$

again using Hölder's inequality together with the moment assumption on ϵ. □

B.2 Proof of Proposition 2.4

Let us first prove (2.6). Since $\mathbb{E}[y|a] = f(a\cdot v_0)$, we have

$$\Phi(x) = \mathbb{E}[y(a\cdot x) - b(a\cdot x)] = \mathbb{E}[f(a\cdot v_0)(a\cdot x)] - c,$$

for some constant c (in particular $c = \mathbb{E}b(a_1)$), where in the second equality, we have used the tower property. Then as in the proof of Proposition 2.1, we see that upon taking $v_0 = e_1$ without loss of generality,

$$\mathbb{E}f(a\cdot v_0)a\cdot x = \mathbb{E}f(a_1)T_m h_1(a_1) = u_1(f)m\,.$$

Finally, since f is increasing, invertible, and differentiable, Gaussian integration-by-parts shows that

$$u_1(f) = \mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)] > 0.$$

Assumption A is now evident from the representation formula (2.6).

Let us now turn to proving that Assumption B holds. To this end, note that as in Proposition 2.1 the relevant inequalities hold for the population loss. Thus it suffices to show it for the true loss. Here we see that if we let $\pi_x(v) = v - (v\cdot x)x$ be the projection on to the tangent space at x, we have

$$\nabla_x\mathcal{L}(y;a,x) = (y - b'(a\cdot x))\pi_x a.$$

We then have by Cauchy-Schwarz

$$\mathbb{E}|\nabla_x \mathcal{L}(y; a, x) \cdot e_1|^2 \leq C\sqrt{\mathbb{E}y^4 + \mathbb{E}b'(a \cdot x)^4}$$

for some $C > 0$. The first term under the radical is finite by assumption and the second term under is finite since $b' = f$ is of at most exponential growth.

For the gradient bound, note again that by the Cauchy-Schwarz inequality,

$$\mathbb{E}|\nabla_x \mathcal{L}|^q = \mathbb{E}\Big[(y - b'(\tfrac{a \cdot x}{\sqrt{N}}))^q ||a||^q\Big] \leq \mathbb{E}[(y - b'(a \cdot x))^{2q}] \cdot C(2q) N^{q/2}$$

where $C(q)$ is as in (A.2). Choosing $q = 4 + \iota$ yields the desired bound by the same reasoning. □

B.3 Proof of Proposition 2.5

In the following C will denote a constant that may change from line to line. It is evident from (2.7) that Φ has information exponent 1 and satisfies Assumption A. It remains to check Assumption B.

Observe that since $\Phi(x) = -2m(x) + c$, where $m(x) = (v, x)$, we have that $\nabla\Phi \cdot v = (\nabla\Phi, \nabla m) = \phi'(m) = -2$ and that $||\nabla\Phi|| \leq C$. Thus it suffices to show the desired bounds for $\nabla\mathcal{L}$. To that end, notice that

$$\nabla\mathcal{L}(x; a, \epsilon) \cdot v = (y - a \cdot x)\, \pi_x a \cdot v = (a \cdot (v - x) + \epsilon)(a \cdot \pi_x v)$$

so that if we let $w = v - x$ and $\tilde{v} = \pi_x v$, then

$$\begin{aligned}\mathbb{E}[(\nabla\mathcal{L}(x) \cdot v)^2] &= \mathbb{E}\left[(a \cdot w)^2 (a \cdot \tilde{v})^2\right] + \mathbb{E}[\epsilon^2]\mathbb{E}[(a \cdot \tilde{v})^2] \\ &\leq \sqrt{\mathbb{E}|a \cdot w|^4 \cdot \mathbb{E}|a \cdot v|^4} + \mathbb{E}[\epsilon^2]\mathbb{E}[|a \cdot v|^2] \leq C\end{aligned}$$

where in the second line we used Cauchy-Schwarz and in the last inequality we used our moment assumption to apply (A.3) and the fact that $||\tilde{v}||, ||w|| \leq 2$.

For the norm bound, we take $4 + \iota = 5$. Then we have

$$\mathbb{E}||\nabla\mathcal{L}||_2^5 \leq \mathbb{E}[|a \cdot w + \epsilon|^5 \cdot ||a||_2^5] \leq 2^5(\mathbb{E}[|a \cdot w|^5 \cdot ||a||_2^5] + \mathbb{E}[|\epsilon|^5] \cdot \mathbb{E}[||a||_2^5])$$

As $||w||_2 \leq 2$, and a is an i.i.d. centered random vector whose entries have finite 10-th moments, by Cauchy–Schwarz and (A.2) and (A.3), we obtain

$$\mathbb{E}|a \cdot w|^5 \cdot \mathbb{E}||a||_2^5 \leq \sqrt{\mathbb{E}|a \cdot w|^{10}}\sqrt{\mathbb{E}||a||_2^{10}} \leq C||w||^5 N^{5/2} = O(N^{5/2})$$

Similarly since ϵ has finite 5-th moment, we see that $\mathbb{E}|\epsilon|^5 \cdot \mathbb{E}||a||_2^5 = O(N^{5/2})$. Combining these bounds yields the desired. □

B.4 Proof of Proposition 2.7

Setting $m(x) = (v_0, x)$, we see from (2.8) that Assumption A holds and the problem has information exponent 2. To verify the first part of Assumption B, observe that if we let $\tilde{v}_0 = \pi_x v_0$,

$$\nabla\mathcal{L}(x) \cdot v_0 = (Y \cdot x)(Y \cdot \tilde{v}_0)\,,$$

so that by Cauchy-Schwarz and (A.3), $\mathbb{E}(\nabla \mathcal{L}(x) \cdot v_0)^2 \leq C(\lambda)$. To verify the second part of Assumption B,

$$||\nabla \mathcal{L}(x)||^q \leq ||YY^T x||^q \leq ||Y||^q |(Y,x)|^q \,,$$

so that by Cauchy-Schwarz, (A.2), and (A.3),

$$\mathbb{E}||\nabla \mathcal{L}(x)||^q \leq \sqrt{\mathbb{E}||Y||^{2q}} \sqrt{\mathbb{E}|(Y,x)|^{2q}} \leq C' N^{q/2}$$

by assumption if we take $q = 4 + \iota = 5$. □

B.5 Proof of Proposition 2.8

Taking the expectation of (2.10), we have $\Phi(x) = -\lambda m(x)^p$, where $m(x) = x \cdot v_0$ so that Assumption A holds, and the problem has information exponent p. For Assumption B we argue as follows.

First note that $H(x) = (J, x^{\otimes p}) = J(x, \dots, x)$. If we let D denote the Euclidean derivative we have $||\nabla H|| \leq ||DH||$. From this, it follows from writing out DH that

$$\mathbb{E}[||DH(x)||^q] \leq C \mathbb{E}[||J(x^{\otimes p-1}, \cdot)||_2^q] \,,$$

for some $C = C(p,q)$, where

$$J(x^{\otimes p-1}, \cdot)_k = \sum_{i_1 \cdots i_{p-1}} J_{i_1 \cdots i_{p-1} k} x_{i_1} \cdots x_{i_{p-1}},$$

Let $I = (i_1, \dots, i_{p-1})$ denote a multi-index and $x_I = \prod_{i \in I} x_i$. Observe that $J(x^{\otimes p-1}, \cdot)$ is a centered i.i.d. vector with

$$\mathbb{E}\Big[J\left(x^{\otimes p-1}, \cdot\right)_k^6 \Big] = \mathbb{E}\Big[\sum_{I_1, \dots, I_6} J_{I_1 1} \cdots J_{I_6 1} x_{I_1} \cdots x_{I_6} \Big] , \leq C$$

for some $C > 0$ depending on the law of J, where here we have used that the entries of J have finite 6-th moment and that x is a unit vector. Taking $q = 4 + \iota = 6$ we see that since the entries of J have finite 6-th moment, we have by (A.2) that $\mathbb{E}||DH||^6 \leq CN^3$, for some $C > 0$. This yields the second half of Assumption B.

For the first half of Assumption B, note that if we let $\tilde{v} = \pi_x v$, then $\nabla H \cdot v = \sum_i \partial_i H(x) \tilde{v}_i$ is a sum of centered i.i.d. random variables with deterministic weights $\tilde{v}$ with uniformly bounded 6th moment, i.e., $\sup_i \mathbb{E}[\partial_i H(x)]^6 \leq C$, by the above argument. Thus by (A.3) we have $\mathbb{E}(\sum_i \partial_i H(x) \tilde{v}_i)^2 \leq C$ as desired since $||\tilde{v}|| \leq 1$. □

B.6 Proof of Proposition 2.10

Observe that $Y \overset{(d)}{=} Z + \epsilon \mu$ with $Z \sim \mathcal{N}(0, Id)$ and ϵ an independent Rademacher r.v. Writing $p = e^h/(e^{-h} + e^h)$ for some $h \in \mathbb{R}$ as above we have that $\Phi(x) = \phi(m(x))$ where

$$\begin{aligned} \phi(m) &= -\mathbb{E}\Big[\log\cosh\Big(Z_1 m + \sqrt{1-m^2} Z_2 + \epsilon m(x) + h \Big) \Big] \\ &= -\mathbb{E}\Big[\log\cosh\left(Z_1 + \epsilon m + h \right) \Big]. \end{aligned}$$

In the first line we used rotation invariance of the law of Z and that Z_1, Z_2 are the first two entries of Z, and in the second line we used that $Z_1 m + Z_2\sqrt{1-m^2} \overset{(d)}{=} Z_1$ since $m \in [-1,1]$.

Consequently

$$\phi'(m) = -\mathbb{E}\tanh\left(Z_1 + \epsilon m + h\right)\epsilon.$$

From this we see that

$$\phi'(0) = -\mathbb{E}\big[\tanh(Z_1+h)\epsilon\big] = \begin{cases} -\mathbb{E}\tanh(Z_1+h)\epsilon < 0 & p \neq 1/2 \\ 0 & p = 1/2. \end{cases}$$

and

$$\phi''(m) = -\mathbb{E}\big[\operatorname{sech}^2(Z_1 + \epsilon m + h)\big] < 0.$$

Combining these two results yields: (a) that the information exponent is 1 if $p \neq 1/2$ and 2 if $p = 1/2$ and (b) that $\phi'(m) > 0$ for $m > 0$ the desired. □

Appendix C. The discrete Bihari–LaSalle inequality

For the purposes of completeness, in this appendix, we provide a proof of the discrete version of the Bihari–LaSalle inequality (5.2). Fix a $k > 2$ and suppose that a_t satisfies

$$a_t = a + \sum_{\ell=0}^{t-1} c(a_j)^k.$$

for some $a, c > 0$, then inductively, we can deduce that $m_t \geq a_t$ for all $t \geq 0$. To see this, note that if we let

$$b_t = a + \sum_{j=0}^{t-1} c(m_j)^{k-1},$$

it suffices to show that $b_t \geq a_t$. Clearly $b_0 = a_0$. Suppose now that $b_j \geq a_j$; then

$$b_{j+1} = a + \sum_{\ell=0}^{j} c(m_\ell)^k = b_j + c(m_j)^k \geq b_j + c(b_j)^k \geq a_j + c(a_j)^k = a_{j+1}$$

where the first inequality follows by definition of b_j and the follows from the inductive hypothesis. Thus $m_t \geq b_t \geq a_t$. It remains to lower bound a_t.

To this end, note that by definition, a_t is non-decreasing for $a > 0$ and

$$c = \frac{a_t - a_{t-1}}{a_{t-1}^k} \leq \int_{a_{t-1}}^{a_t} \frac{1}{x^k} dx = \frac{1}{(k-1)}\Big[\frac{1}{a_{t-1}^{k-1}} - \frac{1}{a_t^{k-1}}\Big].$$

So by re-arrangement,

$$a_t \geq \Big(a_{t-1}^{-(k-1)} - (k-1)c\Big)^{-\frac{1}{k-1}} \quad \text{and} \quad a_t^{-(k-1)} \leq a_{t-1}^{-(k-1)} - (k-1)c.$$

Since this holds for each t, we see that

$$a_{t-1}^{-(k-1)} \leq a_0^{-(k-1)} - (k-1)c(t-1)$$

from which it follows that

$$a_t \geq \frac{1}{(a_0^{-(k-1)} - (k-1)ct)^{\frac{1}{k-1}}} = \frac{a}{(1-(k-1)ca^{k-1}t)^{\frac{1}{k-1}}}.$$

as desired.

References

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2010. ISBN 978-0-521-19452-5.

H. Ashtiani, S. Ben-David, N. Harvey, C. Liaw, A. Mehrabian, and Y. Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3412–3421. Curran Associates, Inc., 2018.

B. Aubin, B. Loureiro, A. Baker, F. Krzakala, and L. Zdeborová. Exact asymptotics for phase retrieval and compressed sensing with random generative priors. In J. Lu and R. Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 55–73, Princeton University, Princeton, NJ, USA, 20–24 Jul 2020. PMLR. URL `http://proceedings.mlr.press/v107/aubin20a.html`.

J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1802705116.

G. Ben Arous, S. Mei, A. Montanari, and M. Nica. The landscape of the spiked tensor model. *Comm. Pure Appl. Math.*, 72(11):2282–2330, 2019. ISSN 0010-3640. doi: 10.1002/cpa.21861.

G. Ben Arous, R. Gheissari, and A. Jagannath. Bounding flows for spherical spin glass dynamics. *Communications in Mathematical Physics*, 373(3):1011–1048, 2020a. doi: 10.1007/s00220-019-03649-4.

G. Ben Arous, R. Gheissari, and A. Jagannath. Algorithmic thresholds for tensor PCA. *Annals of Probability*, 48(4):2052–2087, 2020b.

S. Ben-David, E. Kushilevitz, and Y. Mansour. Online learning versus offline learning. In *Computational learning theory (Barcelona, 1995)*, volume 904 of *Lecture Notes in Comput. Sci.*, pages 38–52. Springer, Berlin, 1995. doi: 10.1007/3-540-59119-2_167.

M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 1–68. Springer, Berlin, 1999. doi: 10.1007/BFb0096509.

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. ISBN 3-540-52894-6. doi: 10.1007/978-3-642-75894-2. Translated from the French by Stephen S. Wilson.

G. Biroli, C. Cammarota, and F. Ricci-Tersenghi. How to iron out rough landscapes and get optimal performances: Averaged gradient descent and its application to tensor pca. *Journal of Physics A: Mathematical and Theoretical*, 2020.

C. M. Bishop. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0387-31073-2; 0-387-31073-8. doi: 10.1007/978-0-387-45528-0.

L. Bottou. *On-Line Learning and Stochastic Approximations*. Cambridge University Press, USA, 1999. ISBN 0521652634.

L. Bottou. Stochastic learning. In *Summer School on Machine Learning*, pages 146–168. Springer, 2003.

L. Bottou and Y. 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.

E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inform. Theory*, 61(4):1985–2007, 2015. ISSN 0018-9448. doi: 10.1109/TIT.2015.2399924.

Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1): 5–37, 2019. doi: 10.1007/s10107-019-01363-6.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018.

X. Cheng, D. Yin, P. Bartlett, and M. Jordan. Stochastic gradient and Langevin processes. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1810–1819. PMLR, 13–18 Jul 2020. URL `http://proceedings.mlr.press/v119/cheng20e.html`.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *Ann. Statist.*, 48(3):1348–1382, 06 2020. doi: 10.1214/19-AOS1850.

M. Duflo. *Algorithmes stochastiques*, volume 23 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1996. ISBN 3-540-60699-8.

D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

R. Ge, F. Huang, C. Jin, and Y. 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.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In A. Beygelzimer and D. 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.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

S. B. Hopkins, J. Shi, and D. Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006, 2015.

S. B. Hopkins, T. Schramm, J. Shi, and D. 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.

A. Jagannath, P. Lopatto, and L. Miolane. Statistical thresholds for tensor PCA. *Ann. Appl. Probab.*, 30(4):1910–1933, 2020. ISSN 1050-5164. doi: 10.1214/19-AAP1547.

H. Jeong and C. S. Güntürk. Convergence of the randomized kaczmarz method for phase retrieval. *ArXiv*, abs/1706.10291, 2017.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300506. doi: 10.1145/1806689.1806765.

C. Kim, A. S. Bandeira, and M. 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.

T. Krasulina. The method of stochastic approximation for the determination of the least eigenvalue of a symmetrical matrix. *USSR Computational Mathematics and Mathematical Physics*, 9(6):189 – 195, 1969. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(69)90135-9.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. ISBN 978-3-642-20211-7. Isoperimetry and processes, Reprint of the 1991 edition.

T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová. Statistical and computational phase transitions in spiked tensor estimation. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 511–515. IEEE, 2017.

C. J. Li, Z. Wang, and H. 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.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, AC-22(4):551–575, 1977. ISSN 0018-9286. doi: 10.1109/tac.1977.1101561.

Y. M. Lu and G. Li. Phase transitions of spectral initialization for high-dimensional nonconvex estimation. *Information and Inference: A Journal of the IMA*, to appear.

W. Luo, W. Alghamdi, and Y. M. Lu. Optimal Spectral Initialization for Signal Recovery With Applications to Phase Retrieval. *IEEE Transactions on Signal Processing*, 67(9): 2347–2356, May 2019. doi: 10.1109/TSP.2019.2904918.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

A. Maillard, G. Ben Arous, and G. Biroli. Landscape complexity for the empirical risk of generalized linear models. In J. Lu and R. Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 287–327, Princeton University, Princeton, NJ, USA, 20–24 Jul 2020. PMLR. URL `http://proceedings.mlr.press/v107/maillard20a.html`.

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

S. S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, and L. Zdeborová. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. In *Advances in Neural Information Processing Systems 32*, pages 8679–8689. Curran Associates, Inc., 2019a.

S. S. Mannelli, F. Krzakala, P. Urbani, and L. Zdeborova. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4333–4342, Long Beach, California, USA, 09–15 Jun 2019b. PMLR.

S. S. Mannelli, G. Biroli, C. Cammarota, F. Krzakala, P. Urbani, and L. Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020.

P. McCullagh and J. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989. ISBN 9780412317606.

D. L. McLeish. Functional and random central limit theorems for the robbins-munro process. *Journal of Applied Probability*, 13(1), 1976. doi: 10.2307/3212676.

S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *Ann. Statist.*, 46(6A):2747–2774, 12 2018. doi: 10.1214/17-AOS1637.

M. Mondelli and A. Montanari. Fundamental limits of weak recovery with applications to phase retrieval. In *Conference On Learning Theory*, pages 1445–1450. PMLR, 2018.

A. Montanari, D. Reichman, and O. 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.

D. Needell and R. Ward. Batched stochastic gradient descent with weighted sampling. In G. E. Fasshauer and L. L. Schumaker, editors, *Approximation Theory XV: San Antonio 2016*, pages 279–306, Cham, 2017. Springer International Publishing. ISBN 978-3-319-59912-0.

D. Needell, N. Srebro, and R. 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.

E. Oja and J. Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69 – 84, 1985. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(85)90131-3.

S. Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, Jan 2006. ISSN 1432-2064. doi: 10.1007/s00440-005-0466-z.

A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra. Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.*, 46(5):2416–2451, 2018. ISSN 0090-5364. doi: 10.1214/17-AOS1625.

M. Raginsky, A. Rakhlin, and M. 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.

E. Richard and A. Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22: 400–407, 1951. ISSN 0003-4851. doi: 10.1214/aoms/1177729586.

V. Ros, G. Ben Arous, G. Biroli, and C. Cammarota. Complex energy landscapes in spiked-tensor and simple glassy models: Ruggedness, arrangements of local minima, and phase transitions. *Phys. Rev. X*, 9:011003, Jan 2019. doi: 10.1103/PhysRevX.9.011003. URL `https://link.aps.org/doi/10.1103/PhysRevX.9.011003`.

J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. doi: 10.1007/s10208-017-9365-9.

P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1395–1403. Curran Associates, Inc., 2014.

Y. S. Tan and R. Vershynin. Phase retrieval via randomized Kaczmarz: theoretical guarantees. *Information and Inference: A Journal of the IMA*, 8(1):97–123, 04 2018. ISSN 2049-8772. doi: 10.1093/imaiai/iay005.

Y. S. Tan and R. Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *arXiv preprint arXiv:1910.12837*, 2019.

R. Vershynin. *High–Dimensional Probability*. Cambridge University Press, 2019.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009. ISSN 0018-9448. doi: 10.1109/TIT.2009.2016018.

C. Wang and Y. M. Lu. Online learning for sparse pca in high dimensions: Exact dynamics and phase transitions. *2016 IEEE Information Theory Workshop (ITW)*, pages 186–190, 2016.

C. Wang, J. Mattingly, and Y. Lu. Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and pca. *arXiv preprint arXiv:1712.04332*, 2017.

A. S. Wein, A. E. Alaoui, and C. Moore. The kikuchi hierarchy and tensor pca. *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1446–1468, 2019.

H. Zhang and Y. Liang. Reshaped wirtinger flow for solving quadratic system of equations. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2622–2630. Curran Associates, Inc., 2016.

Y. Zhang, P. Liang, and M. 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.