Electron. Commun. Probab. **29** (2024), article no. 3, 1–15.
https://doi.org/10.1214/23-ECP571
ISSN: 1083-589X

**ELECTRONIC
COMMUNICATIONS
in PROBABILITY**

High-dimensional limit of one-pass SGD on least squares

Elizabeth Collins–Woodfin* Elliot Paquette†

Abstract

We give a description of the high-dimensional limit of one-pass single-batch stochastic gradient descent (SGD) on a least squares problem. This limit is taken with non-vanishing step-size, and with proportionally related number of samples to problem-dimensionality. The limit is described in terms of a stochastic differential equation in high dimensions, which is shown to approximate the state evolution of SGD. As a corollary, the statistical risk is shown to be approximated by the solution of a convolution-type Volterra equation with vanishing errors as dimensionality tends to infinity. The sense of convergence is the weakest that shows that statistical risks of the two processes coincide. This is distinguished from existing analyses by the type of high-dimensional limit given as well as generality of the covariance structure of the samples.

Keywords: stochastic gradient descent; random matrix theory; optimization; stochastic differential equations.
MSC2020 subject classifications: 60H30.
Submitted to ECP on May 31, 2023, final version accepted on December 15, 2023.
Supersedes arXiv:2304.06847.

1 Introduction

Stochastic optimization methods are the modern day standard for many large-scale computational tasks, especially those that arise in machine learning. There is a long history of analyses of these algorithms, beginning with the seminal work of [RM51], which focused on long-time behavior in a fixed dimensional space. However, modern applications of stochastic optimization have motivated a different regime of analysis, where the problem dimensionality grows proportionally with the run-time of the algorithm.

In this article, we derive the exact scaling behavior of stochastic gradient descent (SGD) on a least squares problem, in the *one-pass* setting (see below) when dimension tends to infinity. We further draw a comparison to the recent work [Paq+22, Paq+21], in which the *multi-pass* version of this problem was considered.

Stochastic gradient descent for empirical risk minimization Most versions of (minibatch) SGD can be formulated in the context of *finite-sum problems*:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) \right\}. \tag{1.1}$$

*McGill University, Canada. E-mail: elizabeth.collins-woodfin@mail.mcgill.ca; https://sites.google.com/view/e-collins-woodfin

†McGill University, Canada. E-mail: elliot.paquette@mcgill.ca; http://elliotpaquette.github.io

Empirical risk minimization fits in this context by supposing that there are n independent samples from some data distribution, and each f_i represents the loss of how the parameters $\boldsymbol{x}$ in some model fit the i-th datapoint. In this article we will exclusively consider the case of linear regression with ℓ^2–regularizer. So we suppose that there are n iid samples $((\boldsymbol{a}_i, \boldsymbol{b}_i) : 1 \leq i \leq n)$ from some distrbution $\mathcal{D}$, with some assumptions to be specified. We arrange this data into a design matrix $\boldsymbol{A}$ and label vector $\boldsymbol{b}$, whose i-th row is given by $\boldsymbol{a}_i$. Finally, we specify the functions f_i in (1.1) by setting

$$f_i(\boldsymbol{x}) = \tfrac{1}{2}(\boldsymbol{a}_i \cdot \boldsymbol{x} - \boldsymbol{b}_i)^2 + \tfrac{\delta}{2}\|\boldsymbol{x}\|^2.$$

The parameter $\delta \geq 0$ is fixed and is the strength of the regularizer and throughout $\|\cdot\|$ will be the Euclidean norm.

Minibatch stochastic gradient descent in this context can be described as

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \gamma_k \nabla f_{i_k}(\boldsymbol{x}_k) \\ &= \boldsymbol{x}_k - \gamma_k \boldsymbol{A}^T \boldsymbol{e}_{i_k} \boldsymbol{e}_{i_k}^T (\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}) - \gamma_k \delta \boldsymbol{x}_k \end{aligned} \tag{1.2}$$

where $\{\gamma_k\}$ are stepsize parameters, $\boldsymbol{e}_i$ is the i-th standard basis vector, and $\{i_k\}$ is a sequence of choices data. In this article we consider the *one-pass* case, in which $i_k = k$ but the algorithm is terminated after n steps. In practice, the order of the data points might be shuffled once before, but in the setting we have posed, with iid data, there is no point to including this additional randomization. There are other choices for how to pick i_k, and we highlight three of them, all of which are *multi-pass* variants.

In *random (with replacement) sample* SGD, each i_k is chosen uniformly at random from $\{1, 2, \ldots, n\}$. This is the setting considered in [Paq+22], and we shall refer simply to this flavor of SGD simply as *multi-pass* SGD in the bulk of the paper. But for context, we also mention *single shuffle* SGD, in which one takes $i_k = k \bmod n$, and so only differs from one-pass SGD in that the algorithm performs the same operations every *epoch*.[1] In *random shuffle* SGD, one modifies the above strategy by randomly permuting the data between each epoch.

All of these strategies are extensively studied in the optimization literature: it is generally thought that the *single shuffle* and *random shuffle* strategies are faster than the random sample strategy [YSJ21] (see also [RR12, GOP21, SS20, AYS20]). We also note that there is a closely related story for the Kaczmarz class of iterative algorithms, which in the language above is like a single-shuffle SGD but with a adapted, non-uniform stepsize γ_k; *randomized Kaczmarz* [SV09, Nee10, NWS14], whose properties are better understood, is closely related to random sample SGD.

The one-pass case is the fundamental point of comparison for all of these methods, being both simpler phenomenologically and also representing an idealization of SGD, in which the run-time of the algorithm is the amount data. Appropriately, running for longer (meaning increasing n) can only improve the statistical performance of the SGD estimator $\boldsymbol{x}_n$, in which context this is usually referred to as *streaming SGD*.

The performance of SGD is measured through the *population risk* $\mathcal{P}$ and sometimes through an ℓ_2–regularized risk $\mathcal{R}$, which are given by

$$\mathcal{P}(\boldsymbol{x}) \coloneqq \tfrac{1}{2}\mathbb{E}_{(\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{a}\cdot\boldsymbol{x} - \boldsymbol{b})^2, \qquad (\boldsymbol{a},\boldsymbol{b}) \sim \mathcal{D}, \tag{1.3}$$

$$\mathcal{R}(\boldsymbol{x}) \coloneqq \tfrac{1}{2}\mathbb{E}_{(\boldsymbol{a},\boldsymbol{b})}(\boldsymbol{a}\cdot\boldsymbol{x} - \boldsymbol{b})^2 + \tfrac{\delta}{2}\|\boldsymbol{x}\|^2, \qquad (\boldsymbol{a},\boldsymbol{b}) \sim \mathcal{D}. \tag{1.4}$$

This regularized risk appears naturally as the mean behavior of one-pass SGD, in that

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k(\nabla\mathcal{R}(\boldsymbol{x}_k) + \xi_{k+1}),$$

[1] We take epoch to mean n steps of the algorithm in all these contexts (including the with-replacement case).

for martingale increments $(\xi_k : 1 \le k \le n)$. Another natural statistical setting to consider is out-of-distributional regression, in which case we would measure the performance of SGD trained on $\mathcal{D}$ but tested, as in (1.3) after replacing $\mathcal{D}$ with another distribution $\mathcal{D}'$. We shall not pursue this case in detail, but we note that all of the above examples are some quadratic functionals of the SGD state $\boldsymbol{x}$.

Data and stepsize assumptions The goal of this analysis is to allow the number of samples n to be large and proportional to the dimension of the problem, here d. This means that the data must be normalized to be nearly dimension independent. Further, we shall need good tail properties of some of the random variables involved, and so we recall the Orlicz norms $\|\cdot\|_{\psi_p}$ for $p \ge 1$ which are given by

$$\|X\|_{\psi_p} = \inf\{t : \mathbb{E}e^{|X|^p/t^p} \le 2\}.$$

We refer the reader to [Ver18] for further exposition, properties and equivalent formulations.

We shall suppose throughout that under $\mathcal{D}$, the labels are given by an underlying linear model with noise. Formally, we suppose that:

Assumption 1.1. *For $(\boldsymbol{a}, \boldsymbol{b})$ sampled from $\mathcal{D}$, conditionally on $\boldsymbol{a}$, the distribution of $\boldsymbol{b}$ is given by $\boldsymbol{a} \cdot \tilde{\boldsymbol{x}} + w$ where w is mean 0, variance $\eta^2 \ge 0$ and is subgaussian with $\|w\|_{\psi_2} \le d^{\varepsilon}$. The ground truth $\tilde{\boldsymbol{x}}$ is assumed to have norm at most d^{ε}.*

The constant ε will be small and fixed throughout. Anything less than $\frac{1}{36}$ will do, and we make no attempt to optimize this constant.

The data covariance is assumed to be normalized in such a way that it is bounded in norm, which is to say:

Assumption 1.2. *The covariance matrix $\boldsymbol{K} := \mathbb{E}[\boldsymbol{a}\boldsymbol{a}^T]$ has operator norm bounded independent of d.*

Note that while we do not explicitly assume that $\boldsymbol{a}$ is centered, the mean would have to be small in some sense to achieve the assumption above.

Finally, we suppose that $\boldsymbol{a}$ has good tail properties, namely that

Assumption 1.3. *The data vector $\boldsymbol{a}$ satisfies that, for any deterministic $\boldsymbol{x}$ of norm less than 1, $\|\boldsymbol{a} \cdot \boldsymbol{x}\|_{\psi_2} \le d^{\varepsilon}$, and the data vector $\boldsymbol{a}$ satisfies the Hanson-Wright inequality: for all $t \ge 0$ and for any deterministic matrix $\boldsymbol{B}$*

$$\mathbb{P}\left(\left|\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} - \mathbb{E}\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}\right| \ge t\right) \le 2\exp\left(-\min\left\{\frac{t^2}{d^{4\varepsilon}\|\boldsymbol{B}\|_{HS}^2}, \frac{t}{d^{2\varepsilon}\|\boldsymbol{B}\|}\right\}\right).$$

We remark that these assumptions hold for two important settings:

(a) when $\boldsymbol{a} = \sqrt{\boldsymbol{K}}\boldsymbol{u}$ where $\boldsymbol{K}$ is some deterministic matrix of bounded operator norm and $\boldsymbol{u}$ is a vector of iid subgaussian random variables or

(b) when $\boldsymbol{a}$ is a vector with the *convex concentration property*, see [Ada15] for details.

There are natural examples of the second case, such as random features models [RR08] with Lipschitz activation functions (see also [Paq+22, Proposition 6.2] for specifics in the case of random features). We note that by truncation, it is also possible to work in the setting (a) above but solely under a uniform bound on a sufficiently high but finite moment, but we do not pursue this.

Finally, the step-size parameters γ_k must be normalized approporiately:

Assumption 1.4. *The stepsize $\gamma_k = \frac{\gamma}{d}$ for all k and fixed $\gamma > 0$.*

We note that we may also pick $\gamma_k = \gamma(k/d)/d$ for a bounded continuous function $\gamma : [0,\infty) \to [0,\infty)$, and this leads to no change anywhere in the arguments.

In light of all these assumptions, we note that SGD finally reduces to the following stochastic recurrence

$$\boldsymbol{x}_k - \tilde{\boldsymbol{x}} = (\boldsymbol{I}(1 - \tfrac{\gamma\delta}{d}) - \gamma \boldsymbol{m}_k \boldsymbol{m}_k^T)(\boldsymbol{x}_{k-1} - \tilde{\boldsymbol{x}}) - \tfrac{\gamma\delta}{d}\tilde{\boldsymbol{x}} + \gamma \boldsymbol{m}_k \eta_k, \tag{1.5}$$

where $\eta_k = w_k/\sqrt{d}$ and $\boldsymbol{m}_k = \boldsymbol{a}_k/\sqrt{d}$.

Homogenized SGD Our theorem is most easily formulated as showing that the state of SGD can be compared to a certain diffusion model in high dimensions. Homogenized SGD is defined to be a continuous time process with initial condition $\boldsymbol{X}_0 = \boldsymbol{x}_0$ that solves the stochastic differential equation

$$\mathrm{d}\boldsymbol{X}_t = -\gamma \nabla \mathcal{R}(\boldsymbol{X}_t)\mathrm{d}t + \gamma \sqrt{\tfrac{2}{d}\mathcal{P}(\boldsymbol{X}_t)\boldsymbol{K}}\mathrm{d}B_t \tag{1.6}$$

where B_t is standard Brownian motion in dimension d, and $\mathcal{R}$ and $\mathcal{P}$ are the regularized and unregularized risks, respectively (recall (1.3) and (1.4)).

Our main theorem shows that for quadratic statistics (in particular the risks (1.3) and (1.4)), homogenized SGD and SGD are interchangeable to leading order. We use the probabilistic modifier *with overwhelming probability* to mean a statement holds except on an event of probability at most $e^{-\omega(\log d)}$ where $\omega(\log d)$ tends to ∞ faster than $\log d$ as $d \to \infty$. We further introduce a norm $\|\cdot\|_{C^2}$ on quadratic functions $q : \mathbb{R}^d \to \mathbb{C}$

$$\|q\|_{C^2} \coloneqq \|\nabla^2 q\| + \|\nabla q(0)\| + |q(0)|,$$

with the norms on the right hand side being given by the operator and Euclidean norm respectively.

Theorem 1.5. *For any quadratic $q : \mathbb{R}^d \to \mathbb{R}$, and for any deterministic initialization $\boldsymbol{x}_0$ with $\|\boldsymbol{x}_0\| \leq 1$, there is a constant $C(\|\boldsymbol{K}\|)$ so that the processes $\{\boldsymbol{x}_k\}_{k=0}^n$ and $\{\boldsymbol{X}_t\}_{t=0}^{n/d}$ satisfy for any n satisfying $n \leq d \log d/(8C(\|\boldsymbol{K}\|))$*

$$\sup_{0\leq k\leq n} \left| q(\boldsymbol{x}_k) - q(\boldsymbol{X}_{k/d}) \right| < \|q\|_{C^2} \cdot e^{C(\|\boldsymbol{K}\|)\frac{n}{d}} \cdot d^{-\frac{1}{2}+9\varepsilon} \tag{1.7}$$

with overwhelming probability.

The processes $\boldsymbol{x}_k$ and $\boldsymbol{X}_t$ are independent, and hence this is also a statement about concentration. In particular, the statement is also true if we replace $q(\boldsymbol{X}_{k/d})$ by $\mathbb{E}q(\boldsymbol{X}_{k/d})$.

Explicit risk curves Using existing theory (see [Paq+22, Theorem 1.1]), $\mathcal{P}(\boldsymbol{X}_{k/d})$ and $\mathcal{R}(\boldsymbol{X}_{k/d})$ can be seen to concentrate around their means, which solve a convolution Volterra equation. Specifically, $\mathbb{E}\mathcal{P}(\boldsymbol{X}_t) = \Psi(t)$ and $\mathbb{E}\mathcal{R}(\boldsymbol{X}_t) = \Omega(t)$ where

$$\begin{pmatrix}\Psi(t)\\ \Omega(t)\end{pmatrix} = \begin{pmatrix}\mathcal{P}(\mathcal{X}_{\gamma t})\\ \mathcal{R}(\mathcal{X}_{\gamma t})\end{pmatrix} + \int_0^t \begin{pmatrix}K(t-s;\nabla^2\mathcal{P})\Psi(s)\\ K(t-s;\nabla^2\mathcal{R})\Psi(s)\end{pmatrix}\mathrm{d}s, \quad \begin{cases}K(t;\boldsymbol{M}) \coloneqq \frac{\gamma^2}{d}\operatorname{tr}(\boldsymbol{K}\boldsymbol{M}e^{-2\gamma t(\boldsymbol{K}+\delta\boldsymbol{I})}),\\ \mathrm{d}\mathcal{X}_{\gamma t} \coloneqq -\gamma\nabla\mathcal{R}(\mathcal{X}_{\gamma t}), \quad \mathcal{X}_0 = \boldsymbol{X}_0.\end{cases} \tag{1.8}$$

Note the equation for Ψ is autonomous, and the solution of Ω is then solvable in terms of it. The process $\mathcal{X}_t$ is gradient flow and is explicitly solvable. For example, in the special case that $\delta = \eta = 0$, this gradient flow is exactly

$$\mathcal{X}_{\gamma t} - \tilde{\boldsymbol{x}} = \exp(-\gamma t\boldsymbol{K})(\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}).$$

Similar, more complicated formulas can also be stated for the general $\delta, \eta \geq 0$ case. Hence for example in the case $\boldsymbol{K} \succ \rho > 0$ for some dimension independent constant ρ, this converges like $\mathcal{R}(\mathcal{X}_{\gamma t}) \leq Ce^{-\rho\gamma t}$. With assumptions on the spectral distribution on $\boldsymbol{K}$ and $\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}$, more precise statements can be made, including cases of power-law rates and non-zero limiting risks.

This Volterra equation is non-negative and of convolution type, and it is therefore an instance of the *renewal equation*, which appears throughout probability. From general theory on convolution Volterra equations (see especially [Asm03, Section 7]), we can derive many simple conclusions. For example, the boundedness of the risk curve occurs exactly when the L^1-norm of the convolution kernel is less than 1, i.e.[2]

$$1 > \int_0^\infty K(t; \nabla^2 \mathcal{P}) \mathrm{d}t = \tfrac{\gamma}{2} \operatorname{tr}(\boldsymbol{K}^2(\boldsymbol{K} + \delta \boldsymbol{I})^{-1}).$$

One can also derive exact dimension-independent limiting risk curves, under assumptions on the convergence as $d \to \infty$ of some of the parameters in the setup. In particular, assuming that the empirical spectral distribution of $\boldsymbol{K}$ converges as $d \to \infty$, and the initialization $\boldsymbol{x}_0$ and target $\tilde{\boldsymbol{x}}$ are taken independent, mean 0, isotropic, and having norm converging as $d \to \infty$, there is a dimension independent Volterra equation limit. This "average-case" setting is taken in [Paq+21], where more details of the limit are discussed, including rates and limitings losses. Note that in particular, in this setting, we note that to achieve risk ϵ, there is a dimension-independent length of time $T(\epsilon)$ needed to achieve risk ϵ under the Volterra curve. From Theorem 1.5, SGD on the same problem would require $d(T(\epsilon) + o(1))$ many steps with probability tending to 1 in d.

Comparison to the multi-pass case In [Paq+22], the analog of Theorem 1.5 was proven for the (random sample) multi-pass case. In that case, the diffusion is different. Introduce the *empirical risk* $\mathcal{L}(\boldsymbol{x}) = \frac{1}{2n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$ and the regularized empirical risk $f(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x}) + \delta\|\boldsymbol{x}\|^2$. Then the homogenized SGD for the multi-pass case becomes

$$\mathrm{d}\boldsymbol{X}_t = -\gamma \nabla f(\boldsymbol{X}_t)\mathrm{d}t + \gamma\sqrt{\tfrac{2}{d}\mathcal{L}(\boldsymbol{X}_t)(\tfrac{1}{n}\boldsymbol{A}^T\boldsymbol{A})}\mathrm{d}B_t. \tag{1.9}$$

Hence the difference between the multi-pass and one-pass cases is that the population risks are traded for the empirical risk and the data covariance matrix $\boldsymbol{K}$ is traded for the empirical data covariance matrix $(\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A})$. Note that if one conditions on the data $(\boldsymbol{A}, \boldsymbol{b})$, then multi-pass SGD in fact *is* streaming SGD, but with a finitely supported data distribution (specifically the empirical distribution of data); however the empirical distribution of data is very far from satisfying Assumption 1.3.

From (1.9) it follows there is a convolution Volterra equation that describes the evolution of the empirical risk $\mathcal{L}$ up to vanishing factors in dimension (replacing $\mathcal{R}$ by $\mathcal{L}$ and $\boldsymbol{K}$ by $(\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A})$ in (1.8)). Moreover the population risk can be given a deterministic equivalent in the same fashion as was done for Ψ and Ω. By comparing these risk curves, this allows one to give a precise dimension-independent characterization of the value of reusing data in SGD (see Figure 1).

On a technical level, the multi-pass is more complicated than one-pass. The arguments share some fundamental common components, in particular they are based on analysis of the same function class Q_n introduced in (2.4), and at a high-level both proofs follow the analysis in Section 2. The central difference is that, due to limited randomness in the martingale updates, in the multipass case, one must show the state vector $\boldsymbol{x}_k$ remains in a good a subset of the state space in which the martingales are well-behaved. Indeed this

[2] The critical case of norm 1 is not clearly resolvable, but when the norm is greater than 1, the Volterra solution will diverge.

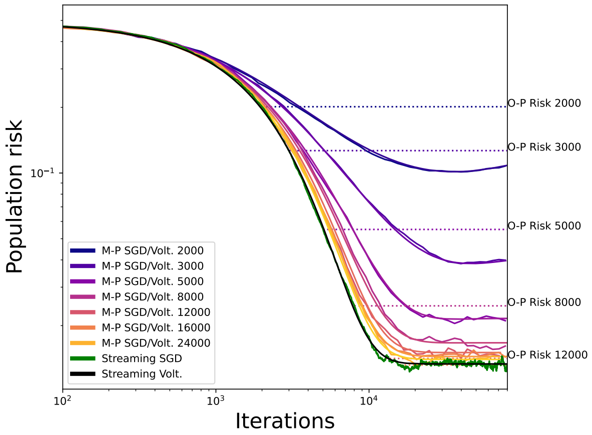

Figure 1: Risk curves for a simple linear regression problem in $d = 2000$. Multi-pass & streaming SGD, and the expected risks ("Volterra") of homogenized SGD are plotted. Risk levels for streaming SGD at various data level n given as "O-P Risk". Note at smaller dataset sizes, multi-pass SGD improves greatly over one-pass SGD. At higher dataset sizes, multi-pass SGD always underperforms.

requires an additional hypothesis that the inital conditions are good, and the argument proceeds by showing the state cannot exit the good set over $O(d)$ steps. In contrast, in the one-pass case this argument is extraneous owing to the higher degree of isotropy of the SGD updates.

Discussion We have presented an approach to taking the high-dimensional limit of one-pass SGD on a least squares problem in which the number of steps is proportional to the dimension of the problem. The limit object is described in terms of a Langevin type diffusion, which can be directly compared to the same object in the multi-pass case.

The literature on scaling limits of one-pass SGD training is large, and so we mention just some of the closest literature. [AGJ22] is perhaps the closest high-dimensional diffusion approximation, and it applies in cases where there is a hidden finite dimensional structure; it covers the case studied here when $\boldsymbol{K} = \boldsymbol{I}$ as well as cases in which $\boldsymbol{K}$ has boundedly many eigenvalues. See also [AGJ21].

There are other scaling limits that pursue a different formulation than the one here. [WL19, WML17] give a PDE limit for the state for a generalized linear model, with identity covariance. [BGH23] give a scaling limit of the SGD under smoothness assumptions on the covariance $\boldsymbol{K}$, when interpreted geometrically; they further describe fluctuations of SGD in a certain sense. Note that Theorem 1.5 essentially gives the law of large numbers for the risks and not the fluctuations.

[Ger+22] uses dynamical field theory to give closely related results for shallow neural networks with minibatch SGD of large batch-size; dynamical mean field theory provides an implicit characterization of the autocorrelation of the minibatch noise and a few other processes. See also the related work of [CCM21]. We comment that in the case of proportional batch sizes, there is also a discrete Volterra description in [Lee+22].

Organization In Section 2 we give an overview of the main proof, reducing it to its main technicalities. In Section 3, we bound the stochastic error terms, representing the main technical contribution of the paper.

2 Main argument of proof

In order to compare the SGD and homogenized SGD, we use a version of the martingale method in diffusion approximations (see [EK09]). In effect we show that $q(\boldsymbol{x}_k)$ nearly satisfies the conclusion of Itô's lemma. Further, we show the martingale terms in both of the Doob decompositions are small, and hence it suffices to show the predictable parts of $q(\boldsymbol{x}_k)$ and $q(\boldsymbol{X}_t)$ are close.

To advance the discussion, we compute this Doob decomposition. To take advantage of the simpler structure afforded by removing $\tilde{\boldsymbol{x}}$, introduce

$$\boldsymbol{v}_k \coloneqq \boldsymbol{x}_k - \tilde{\boldsymbol{x}} \quad \text{and} \quad \boldsymbol{V}_t \coloneqq \boldsymbol{X}_t - \tilde{\boldsymbol{x}}. \tag{2.1}$$

We shall extend the first integer indexed function to real-valued indices by setting $\boldsymbol{v}_t = \boldsymbol{v}_{\lfloor t \rfloor}$. We also let $(\mathcal{F}_t : t \geq 0)$ be the filtration generated by $(\boldsymbol{v}_t : t \geq 0)$ and $(\boldsymbol{V}_{t/d} : t \geq 0)$. Hence for all $k \in \mathbb{N}$, $\boldsymbol{v}_k$ is measurable with respect to $\mathcal{F}_k$. Recalling the recurrence (1.5) for a quadratic q

$$\begin{aligned} q(\boldsymbol{v}_k) - q(\boldsymbol{v}_{k-1}) = -\gamma(\nabla q(\boldsymbol{v}_{k-1}))^T(\boldsymbol{y}_k) + \tfrac{\gamma^2}{2}\boldsymbol{y}_k^T(\nabla^2 q)\boldsymbol{y}_k, \quad \text{where} \\ \boldsymbol{y}_k \coloneqq u_{k-1} + \Delta_k, \quad \Delta_k \coloneqq \boldsymbol{m}_k(\boldsymbol{m}_k^T \boldsymbol{v}_{k-1} - \eta_k) \quad \text{and} \quad \boldsymbol{u}_{k-1} = \tfrac{\delta}{d}(\boldsymbol{v}_{k-1} + \tilde{\boldsymbol{x}}). \end{aligned} \tag{2.2}$$

The equation above can each be decomposed as a predictable part and two martingale increments

$$\begin{aligned} q(\boldsymbol{v}_k) - q(\boldsymbol{v}_{k-1}) = &- \gamma(\nabla q(\boldsymbol{v}_{k-1}))^T \big((\tfrac{\delta}{d}\boldsymbol{I} + \tfrac{1}{d}\boldsymbol{K})\boldsymbol{v}_{k-1} + \tfrac{\delta \tilde{\boldsymbol{x}}}{d} \big) + \Delta\mathcal{M}_k^{\text{lin}} \\ &+ \tfrac{\gamma^2}{2} \operatorname{tr}(\tfrac{1}{d}\boldsymbol{K}(\nabla^2 q)) \big(\tfrac{1}{d}\boldsymbol{v}_{k-1}^T \boldsymbol{K} \boldsymbol{v}_{k-1} + \mathbb{E}[\eta_k^2] \big) + \Delta\mathcal{E}_k^{\text{quad}} + \Delta\mathcal{M}_k^{\text{quad}}, \\ \text{where} \quad \Delta\mathcal{M}_k^{\text{quad}} \coloneqq &\Delta_k^T \nabla^2 q \Delta_k - \mathbb{E}[\Delta_k^T \nabla^2 q \Delta_k \mid \mathcal{F}_{k-1}]. \end{aligned} \tag{2.3}$$

The remainder of the martingale increments are given by $\Delta\mathcal{M}_k^{\text{lin}}$ and are all linear in Δ_k. The predictable parts have been further decomposed into the leading order terms and an error term $\Delta\mathcal{E}_k^{\text{quad}}$. For convenience, we have displayed these below:

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{lin}} &= \big(\boldsymbol{w}_{k-1}^T \boldsymbol{m}_k \big) \big(\boldsymbol{m}_k^T \boldsymbol{v}_{k-1} - \eta_k \big) - \tfrac{1}{d} \boldsymbol{w}_{k-1}^T \boldsymbol{K} \boldsymbol{v}_{k-1}, \\ \Delta\mathcal{E}_k^{\text{quad}} &= \tfrac{\gamma^2}{2} \Big(\mathbb{E}[\boldsymbol{y}_k^T (\nabla^2 q) \boldsymbol{y}_k \;\mid\; \mathcal{F}_{k-1}] - \operatorname{tr}(\tfrac{1}{d}\boldsymbol{K}(\nabla^2 q)) \big(\tfrac{1}{d} \boldsymbol{v}_{k-1}^T \boldsymbol{K} \boldsymbol{v}_{k-1} + \mathbb{E}[\eta_k^2] \big) \Big) \\ \text{where} \quad \boldsymbol{w}_{k-1} &\coloneqq -\gamma \nabla q(\boldsymbol{v}_{k-1}^\tau) + \tfrac{\gamma^2 \delta}{d}(\boldsymbol{v}_{k-1}^\tau + \tilde{\boldsymbol{x}}). \end{aligned}$$

These predictable parts, in turn, depend on different statistics $q_1(\boldsymbol{v}_{k-1})$. In finite dimensional settings, we would be able to relate this (or some suitably large finite set of summary statistics $q, q_1, \ldots, q_r$) to itself through a closed system of recurrences. In this setting, this is not possible. On the other hand, for the problem at hand, we show there is a manifold of functions which approximately closes. Specifically, we let

$$\begin{aligned} Q_n(q) \coloneqq Q_n(q, \boldsymbol{K}) = \Big\{ & q(\boldsymbol{x}), \quad (\nabla q(\boldsymbol{x}))^T R(z; \boldsymbol{K}) \boldsymbol{x}, \quad \boldsymbol{x}^T R(y; \boldsymbol{K}) (\nabla^2 q) R(z; \boldsymbol{K}) \boldsymbol{x}, \\ & (\nabla q(\boldsymbol{x}))^T R(z; \boldsymbol{K}) \tilde{\boldsymbol{x}}, \quad \boldsymbol{x}^T R(y; \boldsymbol{K}) (\nabla^2 q) R(z; \boldsymbol{K}) \tilde{\boldsymbol{x}}, \quad \forall\, z, y \in \mathit{\Gamma} \Big\}. \end{aligned} \tag{2.4}$$

Here $R(z; \boldsymbol{K}) = (\boldsymbol{K} - z\boldsymbol{I})^{-1}$ is the resolvent matrix, and $\mathit{\Gamma}$ is a circle of radius $\max\{1, 3\|\boldsymbol{K}\|\}$. In order to control the martingales, it is convenient to impose a stopping time

$$\tau \coloneqq \inf \{ k : \|\boldsymbol{v}_k\| > d^{\varepsilon} \} \cup \{ td : \|\boldsymbol{V}_t\| > d^{\varepsilon} \}, \tag{2.5}$$

and we introduce the corresponding stopped processes

$$\boldsymbol{v}_k^\tau = \boldsymbol{v}_{k\wedge\tau}, \quad \boldsymbol{V}_t^\tau = \boldsymbol{V}_{t\wedge(\tau/d)}. \tag{2.6}$$

We prove a version of our theorem for the stopped processes and then show that the stopping time is greater than n with overwhelming probability.

Our key tool for comparing $\boldsymbol{v}_{td}$ and $\boldsymbol{V}_t$ is the following lemma.

Lemma 2.1. *Given a quadratic q with $\|q\|_{C^2} \le 1$, with $Q = Q_n(q) \cup Q_n(\mathcal{P}) \cup Q_n(\|\cdot\|^2)$ as above, for any $g \in Q$*

$$\begin{aligned}\max_{0\le t\le \frac{n}{d}} |g(\boldsymbol{v}_{td}^\tau) - g(\boldsymbol{V}_t^\tau)| \le \max_{0\le t\le\frac{n}{d}} \Big(|\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{lin},\tau}| + |\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{quad},\tau}| + |\mathcal{E}_{\lfloor td\rfloor}^{\mathrm{quad},\tau}| + |\mathcal{M}_t^{HSGD,\tau}|\Big) \\ + C(\|\boldsymbol{K}\|)\cdot \int_0^{n/d} \sup_{h\in Q} |h(\boldsymbol{v}_{sd}^\tau) - h(\boldsymbol{V}_s^\tau)| ds.\end{aligned} \tag{2.7}$$

Here $\mathcal{M}_t^{HSGD,\tau}$ is the martingale part in the semimartingale decomposition of $q(\boldsymbol{V}_t^\tau)$.

Sketch of Proof. Owing to the similarities of this claim with the proof in [Paq+22, Proposition 4.1], we just illustrate the main idea. The idea is that if we take a $g \in Q$, and we apply (2.3), then in the predictable part of $g(\boldsymbol{v}_t)$ we have

$$I_1 := \int_0^t \nabla g(\boldsymbol{v}_{sd})^T(\delta \boldsymbol{I} + \boldsymbol{K})\boldsymbol{v}_{sd} \mathrm{d}s, \quad I_2 := \int_0^t \nabla g(\boldsymbol{v}_{sd})^T \tilde{\boldsymbol{x}} \mathrm{d}s, \quad I_3 := \int_0^t \boldsymbol{v}_{sd}^T \boldsymbol{K} \boldsymbol{v}_{sd} \mathrm{d}s.$$

These also appear with coefficients that can be bounded solely using $\|g\|_{C^2}$ and $\|\boldsymbol{K}\|$. We get the same, applying Itô's lemma to $g(\boldsymbol{V}_t)$, albeit with the replacement $\boldsymbol{v}_t \to \boldsymbol{V}_t$. We wish to bound for example $I_1(\boldsymbol{v}_t) - I_1(\boldsymbol{V}_t)$. We do this by expressing its integrand as $p(\boldsymbol{v}_t) - p(\boldsymbol{V}_t)$ for polynomial p. If g is linear (the final row of (2.4)), then p is again linear. For example, if it is $g(\boldsymbol{x}) = \nabla q(\boldsymbol{x})^T R(z; \boldsymbol{K})\tilde{\boldsymbol{x}}$, then p is again linear and is given by

$$p(\boldsymbol{x}) = \boldsymbol{x}^T(\delta \boldsymbol{I} + \boldsymbol{K})R(z;\boldsymbol{K})\tilde{\boldsymbol{x}} = \delta \boldsymbol{x}^T R(z;\boldsymbol{K})\tilde{\boldsymbol{x}} + \boldsymbol{x}^T R(z;\boldsymbol{K})\tilde{\boldsymbol{x}} - z\boldsymbol{x}^T\tilde{\boldsymbol{x}},$$

where we have used the resolvent identity $(\boldsymbol{K} - z)R(z;\boldsymbol{K}) = \boldsymbol{I}$. Note the function $\boldsymbol{x}^T R(z;\boldsymbol{K})\tilde{\boldsymbol{x}}$ is contained in Q by virtue of being in $Q_n(\|\cdot\|^2)$. Moreover, by Cauchy's integral formula, we can represent $\boldsymbol{x}^T\tilde{\boldsymbol{x}}$ by averaging $\frac{-1}{2\pi i}\boldsymbol{x}^T R(y;\boldsymbol{K})\tilde{\boldsymbol{x}}$ over $y \in \mathit{\Gamma}$. Hence

$$|p(\boldsymbol{v}_{td}) - p(\boldsymbol{V}_t)| \le (\delta + 1 + 3\|\boldsymbol{K}\|)\sup_{h\in Q} |h(\boldsymbol{v}_{td}) - h(\boldsymbol{V}_t)|.$$

The same manipulations lead finally to showing every term included in Q can be controlled in a similar manner, using the other elements of the class Q. □

The second important idea is to discretize the set Q.

Lemma 2.2. *There exists $\bar{Q} \subseteq Q$ with $|\bar{Q}| \le C(\|\boldsymbol{K}\|)d^{4m}$ such that, for every $q \in Q$, there is some $\bar{q} \in \bar{Q}$ satisfying $\|q - \bar{q}\|_{C^2} \le d^{-2m}$.*

Proof. On the spectral curve $\mathit{\Gamma}$, we can bound the norm of the resolvent. Since

$$\frac{\mathrm{d}}{\mathrm{d}z} R(z;\boldsymbol{K}) = (\boldsymbol{K} - z\boldsymbol{I})^{-2},$$

we have it is norm bounded by an absolute constant. The arc length of the curve is at most $C(\|\boldsymbol{K}\|)$, and so by choosing a minimal d^{-2m}–net of the manifold $\mathit{\Gamma} \times \mathit{\Gamma}$, the lemma follows. □

Now the main technical part of the argument is to control the martingales and errors. As we work with the stopped process $\boldsymbol{v}_k^\tau$ we introduce the stopped proccesses $\mathcal{M}_k^{\mathrm{lin},\tau}, \mathcal{M}_k^{\mathrm{quad},\tau}, \mathcal{E}_k^{\mathrm{quad},\tau}$, which are defined analogously to (2.6).

Lemma 2.3. *For any quadratic q with $\|q\|_{C^2} \leq 1$, the terms $\mathcal{M}_k^{\mathrm{lin},\tau}, \mathcal{M}_k^{\mathrm{quad},\tau}, \mathcal{E}_k^{\mathrm{quad},\tau}$ satisfy the following bounds with overwhelming probability (with a bound which is uniform in q) for $n \leq d \log d$*

(i) $\sup_{1\leq k\leq n} |\mathcal{M}_k^{\mathrm{lin},\tau}| \leq d^{-\frac{1}{2}+5\varepsilon}$,

(ii) $\sup_{1\leq k\leq n} |\mathcal{M}_k^{\mathrm{quad},\tau}| \leq d^{-\frac{1}{2}+9\varepsilon}$,

(iii) $\sup_{1\leq k\leq n} |\mathcal{E}_k^{\mathrm{quad},\tau}| \leq d^{-1+9\varepsilon}$.

Combining Lemmas 2.1 and 2.2, along with the above (see also Lemma 3.2 in which the homogenized SGD martingales are bounded), we conclude that, for any $\bar{q} \in \bar{Q}$ with $\|q\|_{C^2} = 1$,

$$|\bar{q}(\boldsymbol{v}_{td}^\tau) - \bar{q}(\boldsymbol{V}_t^\tau)| \leq 4d^{-\frac{1}{2}+9\varepsilon} + C(\|\boldsymbol{K}\|)\int_0^t \sup_{g\in Q} |g(\boldsymbol{v}_{sd}^\tau) - g(\boldsymbol{V}_s^\tau)| ds. \tag{2.8}$$

Hence by Lemma 2.2 and by bounding $\|g\|_{C^2}$ over all Q,

$$\sup_{g\in Q} |g(\boldsymbol{v}_{td}^\tau) - g(\boldsymbol{V}_t^\tau)| \leq C(\|\boldsymbol{K}\|)\bigg(d^{-2} + d^{-\frac{1}{2}+9\varepsilon} + \int_0^t \sup_{g\in Q}|g(\boldsymbol{v}_{sd}^\tau) - g(\boldsymbol{V}_s^\tau)|ds\bigg). \tag{2.9}$$

By Gronwall's inequality, this gives us that with overwhelming probability

$$\sup_{g\in Q} \max_{0\leq t\leq n/d} |g(\boldsymbol{v}_{td}^\tau) - g(\boldsymbol{V}_t^\tau)| \leq C(\|\boldsymbol{K}\|)(d^{-2} + 4d^{-\frac{1}{2}+9\varepsilon})e^{C(\|\boldsymbol{K}\|)n/d}. \tag{2.10}$$

Hence this is small provided n/d is controlled by a sufficiently small multiple of $\log d$. Now we note that the norm function $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|^2$ is one of the quadratics included in Q. Hence if we let $\mathcal{G}$ be the event in the above display, and we let $\mathcal{E} = \{\max_{0\leq s\leq n/d} \|\boldsymbol{V}_s\| \leq d^{\varepsilon/2}\}$, then we have

$$\mathcal{G} \cap \mathcal{E} \cap \{\tau \leq n/d\} \subseteq \{\|\boldsymbol{v}_\tau\| - \|\boldsymbol{v}_{\tau-1}\| \geq d^{\varepsilon/2}\} \cap \{\tau \leq n/d\}.$$

This is because on the event $\{\tau \leq n/d\} \cap \mathcal{E}$ we must have had $\|\boldsymbol{v}_\tau\| > d^\varepsilon$, but in the step before τ, we had $\boldsymbol{v}_{\tau-1}$ could be compared to $\boldsymbol{V}_{\tau-1}$ (due to $\mathcal{G}$, and we had the norm of $\boldsymbol{V}_{\tau-1}$ was small). Now with overwhelming probability, no increment of SGD between time 0 and n/d can increase the norm by a power of d. So to complete the proof it suffices to show $\mathcal{E}$ holds with overwhelming probability.

Thus the proof is completed by the following:

Lemma 2.4. *For any $\delta > 0$ and any $t > 0$ with overwhelming probability*

$$\max_{0\leq s\leq t} \|\boldsymbol{X}_s\|^2 \leq e^{C(\|\boldsymbol{K}\|)t} d^\delta.$$

Proof. We apply Itô's formula to $\phi(\boldsymbol{X}_t) := \log(1 + \|\boldsymbol{X}_t\|^2)$, from which we have

$$\mathrm{d}\phi(\boldsymbol{X}_t) = -2\gamma \tfrac{\boldsymbol{X}_t\cdot\nabla\mathcal{R}(\boldsymbol{X}_t)}{1+\|\boldsymbol{X}_t\|^2}\mathrm{d}t + \frac{\boldsymbol{X}_t\cdot\gamma\sqrt{\frac{2}{d}\mathcal{P}(\boldsymbol{X}_t)\boldsymbol{K}}\mathrm{d}B_t}{1+\|\boldsymbol{X}_t\|^2} + \big(\tfrac{\mathcal{P}(\boldsymbol{X}_t)}{1+\|\boldsymbol{X}_t\|^2}\tfrac{2\gamma^2}{d}\operatorname{tr}(\boldsymbol{K}) - \tfrac{2\gamma^2\mathcal{P}(\boldsymbol{X}_t)\boldsymbol{X}_t^T\boldsymbol{K}\boldsymbol{X}_t}{d}\big)\mathrm{d}t$$

The drift terms and the quadratic variation terms can be bounded by some $C(\|\boldsymbol{K}\|)$. Hence with this constant, for all $r \geq 0$,

$$\mathbb{P}(\max_{0\leq s\leq t} \phi(\boldsymbol{X}_s) \geq C(\|\boldsymbol{K}\|)(t + r\sqrt{t})) \leq 2\exp(-r^2/2).$$

Taking $r = \sqrt{\log d \log\log d}$, we conclude that with overwhelming probability

$$\max_{0\leq s\leq t} \phi(\boldsymbol{X}_s) \leq C(\|\boldsymbol{K}\|)(t + \sqrt{t\log d\log\log d}). \qquad \square$$

3 Controlling the errors

The main goal of this section is to control the martingale terms and error terms; in particular we prove Lemma 2.3. In order to obtain these bounds, we will need the following concentration lemma, which is standard (c.f. [Ver18, Theorem 2.8.1], where the nonmartingale bound is proven. The adaptation to the martingale case is a small extension):

Lemma 3.1 (Martingale Bernstein inequality)**.** *If $(M_n)_1^N$ is a martingale on the filtered probability space $(\Omega, (\mathcal{F}_n)_1^N, \mathbb{P}))$ and we define*

$$\sigma_{n,p} := \left\| \inf\{t \geq 0 : \mathbb{E}\left(e^{|M_n - M_{n-1}|^p/t^p} | \mathcal{F}_{n-1}\right) \leq 2\} \right\|_{L^\infty(\mathbb{P})}, \tag{3.1}$$

then there is an absolute constant $C > 0$ so that, for all $t > 0$,

$$\mathbb{P}\left(\sup_{1 \leq n \leq N} |M_n - \mathbb{E}M_0| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{t}{C \max \sigma_{n,1}}, \frac{t^2}{C\sum_1^N \sigma_{n,1}^2}\right\}\right). \tag{3.2}$$

We will also record for future use an estimate on ∇q that follows from $\|\cdot\|_{C^2}$ control.

$$\|\nabla q(\boldsymbol{x})\| \leq \|\nabla^2 q\| \cdot \|\boldsymbol{x}\| + \|\nabla q(0)\| \leq \|q\|_{C^2} \cdot (\|\boldsymbol{x}\| + 1). \tag{3.3}$$

3.1 Martingale for gradient part of recurrence

Proof of Lemma 2.3 part (i). Comparing (2.2) and (2.3), we see that for $k \leq \tau$

$$\begin{aligned} \Delta\mathcal{M}_k^{\text{lin},\tau} &= \left[\left(\boldsymbol{w}_{k-1}^T \boldsymbol{m}_k\right)\left(\boldsymbol{m}_k^T \boldsymbol{v}_{k-1}^\tau - \eta_k\right) - \tfrac{1}{d}\boldsymbol{w}_{k-1}^T \boldsymbol{K} \boldsymbol{v}_{k-1}^\tau\right] =: [\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}], \\ &\text{where} \quad \boldsymbol{w}_{k-1} := -\gamma\nabla q(\boldsymbol{v}_{k-1}^\tau) + \tfrac{\gamma^2\delta}{d}(\boldsymbol{v}_{k-1}^\tau + \tilde{\boldsymbol{x}}). \end{aligned} \tag{3.4}$$

Note for $k > \tau$, the stopped martingale increment is 0. Using (3.3), $\|\boldsymbol{w}_{k-1}\| \leq C(\gamma, \delta)d^{\varepsilon}$. We will separately bound the contributions from $\Delta\mathcal{M}_k^{\text{lin}\,1,\tau}$ and $\Delta\mathcal{M}_k^{\text{lin}\,2,\tau}$ in terms of their Orlicz norms. For the first part, for any fixed k, conditionally on $\mathcal{F}_{k-1}$ and using Assumption 1.3, we conclude

$$\|\Delta\mathcal{M}_k^{\text{lin}\,1,\tau}\|_{\psi_1} \leq \left\|\boldsymbol{w}_{k-1}^T \boldsymbol{m}_k\right\|_{\psi_2} \left\|\boldsymbol{m}_k^T \boldsymbol{v}_{k-1}^\tau - \eta_k\right\|_{\psi_2} \leq C d^{-\frac{1}{2}+2\varepsilon} \cdot d^{-\frac{1}{2}+2\varepsilon} \tag{3.5}$$

where C is some absolute constant. For the second part, we have

$$|\Delta\mathcal{M}_k^{\text{lin}\,2,\tau}| = |\tfrac{1}{d}\boldsymbol{w}_{k-1}^T \boldsymbol{K} \boldsymbol{v}_{k-1}^\tau| \leq C d^{-1+2\varepsilon}. \tag{3.6}$$

Combining these, we see that, for every k,

$$\sigma_{k,1} := \inf\{t > 0 : \mathbb{E}[\exp(|\Delta\mathcal{M}_k^{\text{lin}\,1,\tau} - \Delta\mathcal{M}_k^{\text{lin}\,2,\tau}|/t)|\mathcal{F}_{k-1}] \leq 2\} \leq C d^{-1+4\varepsilon} \tag{3.7}$$

and, by the martingale Bernstein inequality, for some other constants $C, c > 0$

$$\begin{aligned} \mathbb{P}\left(\sup_{1 \leq k \leq n} |\mathcal{M}_k^{\text{lin},\tau} - \mathbb{E}\mathcal{M}_0^{\text{lin}}| \geq t\right) &\leq 2\exp\left(-\min\left\{\frac{t}{C\max\sigma_{k,1}}, \frac{t^2}{C\sum_{k=1}^n \sigma_{k,1}^2}\right\}\right) \\ &\leq 2\exp\left(-\min\left\{ctd^{1-4\varepsilon}, ct^2 d^{2-8\varepsilon} n^{-1}\right\}\right). \end{aligned} \tag{3.8}$$

As we assume that $n \leq d\log d$ then this gives us

$$\sup_{1 \leq k \leq n} |\mathcal{M}_k^{\text{lin},\tau}| \leq d^{-\frac{1}{2}+5\varepsilon} \tag{3.9}$$

with overwhelming probability. □

3.2 Martingale for Hessian part of recurrence

Proof of Lemma 2.3 parts (ii) and (iii). Next we consider the contribution from the Hessian part of the recurrence. We write

$$\begin{aligned}&\tfrac{\gamma^2}{2}(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)^T(\nabla^2 q)(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)\\&=\mathbb{E}\left[\tfrac{\gamma^2}{2}(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)^T(\nabla^2 q)(\boldsymbol{m}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\boldsymbol{m}_k\eta_k)|\mathcal{F}_{k-1}\right]+\Delta\mathcal{M}_k^{\text{quad}}.\end{aligned} \tag{3.10}$$

Since we will be conditioning on $\mathcal{F}_{k-1}$ extensively throughout this section, we use the shorthand $\mathbb{E}_k[\,\cdot\,]:=\mathbb{E}[\,\cdot\,|\mathcal{F}_{k-1}]$. Rearranging the terms of (3.10), we get

$$\Delta\mathcal{M}_k^{\text{quad}}=A_kB_k-\mathbb{E}_k[A_kB_k] \tag{3.11}$$

where

$$A_k:=\boldsymbol{m}_k^T(\nabla^2 q)\boldsymbol{m}_k,\quad B_k:=(\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau-\eta_k)^2. \tag{3.12}$$

This can be expanded as

$$\begin{aligned}\Delta\mathcal{M}_k^{\text{quad}}&=(A_k-\mathbb{E}_k[A_k])(B_k-\mathbb{E}_k[B_k])+\mathbb{E}_k[A_k]\mathbb{E}_k[B_k]-\mathbb{E}_k[A_kB_k]\\&\quad+(A_k-\mathbb{E}_k[A_k])\mathbb{E}_k[B_k]+(B_k-\mathbb{E}_k[B_k])\mathbb{E}_k[A_k],\end{aligned} \tag{3.13}$$

so we focus first on obtaining subexponential bounds for the quantities $A_k-\mathbb{E}_k[A_k]$ and $B_k-\mathbb{E}_k[B_k]$ using the Hanson-Wright inequality. For A_k, we have

$$\begin{aligned}\mathbb{P}(|A_k-\mathbb{E}_kA_k|\geq t)&\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}\|\nabla^2 q\|_{HS}^2},\frac{t}{d^{-1+2\varepsilon}\|\nabla^2 q\|}\right)\right]\\&\leq 2\exp[-c'\min(t^2d^{1-4\varepsilon},td^{1-2\varepsilon})]\leq 2\exp[-c''td^{\frac{1}{2}-2\varepsilon}]\end{aligned} \tag{3.14}$$

and thus we have the subexponential bound

$$\|A_k-\mathbb{E}_k[A_k]\|_{\psi_1}<Cd^{-\frac{1}{2}+2\varepsilon}. \tag{3.15}$$

Next we obtain a subexponential bound for B_k. For the part of B_k not involving η_k, we use Hanson-Wright to get

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T\boldsymbol{m}_k-\mathbb{E}_k\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T\boldsymbol{m}_k\right|\geq t\right)\\&\leq 2\exp\left[-c\min\left(\frac{t^2}{d^{-2+4\varepsilon}\|\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T\|_{HS}^2},\frac{t}{d^{-1+2\varepsilon}\|\boldsymbol{v}_{k-1}^\tau(\boldsymbol{v}_{k-1}^\tau)^T\|}\right)\right]\\&\leq 2\exp[-c\min(t^2d^{2-8\varepsilon},td^{1-4\varepsilon})].\end{aligned} \tag{3.16}$$

For the terms involving η_k, we use the Orlicz bounds from the assumptions in the set-up to obtain

$$\begin{aligned}\|\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau\eta_k\|_{\psi_1}&\leq\|\boldsymbol{m}_k^T\boldsymbol{v}_{k-1}^\tau\|_{\psi_2}\cdot\|\eta_k\|_{\psi_2}=d^{-\frac{1}{2}+2\varepsilon}d^{-\frac{1}{2}+\varepsilon}\\&=d^{-1+3\varepsilon}.\end{aligned} \tag{3.17}$$

Since also $\|\eta_k^2\|_{\psi_1}=d^{-1+2\varepsilon}$ combining the bounds (3.16) and (3.17), we have

$$\|B_k-\mathbb{E}_k[B_k]\|_{\psi_1}<Cd^{-1+4\varepsilon}. \tag{3.18}$$

Furthermore, we have

$$\mathbb{E}_k[A_k]=O(1),\qquad\mathbb{E}_k[B_k]=O(d^{-1}), \tag{3.19}$$

uniformly for all k based on the assumptions on η_k and $\boldsymbol{m}_k$. We now use (3.15), (3.18), (3.19) to bound each term of (3.13) in turn.

To bound the contribution from $(A_k - \mathbb{E}_k[A_k])(B_k - \mathbb{E}_k[B_k])$, we observe that, for each k, with overwhelming probability, $|A_k - \mathbb{E}_k[A_k]| < d^{-\frac{1}{2}+3\varepsilon}$ and $|B_k - \mathbb{E}_k[B_k]| < d^{-1+5\varepsilon}$, so we can conclude that, with overwhelming probability,

$$\sum_{k=1}^{n} \Big|(A_k - \mathbb{E}_k[A_k])(B_k - \mathbb{E}_k[B_k])\Big| < nd^{-\frac{3}{2}+8\varepsilon} < d^{-\frac{1}{2}+9\varepsilon}. \tag{3.20}$$

For the second term of (3.13) we have

$$\begin{aligned}\Big|\mathbb{E}_k[A_k]\mathbb{E}_k[B_k] - \mathbb{E}_k[A_k B_k]\Big| &= \Big|\mathbb{E}_k\big[(A_k - \mathbb{E}_k A_k)(B_k - \mathbb{E}_k B_k)\big]\Big| \\ &\leq \mathbb{E}_k\Big|(A_k - \mathbb{E}_k A_k)(B_k - \mathbb{E}_k B_k)\Big| \\ &\leq \mathbb{E}_k\Big|(A_k - \mathbb{E}_k A_k)(B_k - \mathbb{E}_k B_k)\mathbf{1}_{\mathcal{E}}\Big| + \mathbb{E}_k\Big|(A_k - \mathbb{E}_k A_k)(B_k - \mathbb{E}_k B_k)\mathbf{1}_{\mathcal{E}^c}\Big|\end{aligned} \tag{3.21}$$

where $\mathcal{E}$ is the event

$$\mathcal{E} := \Big\{|A_k - \mathbb{E}_k A_k| \leq d^{-\frac{1}{2}+3\varepsilon}\Big\} \cap \big\{|B_k - \mathbb{E}_k B_k| \leq d^{-1+5\varepsilon}\big\},$$

which holds with overwhelming probability. Thus, using this event and Cauchy-Schwartz, the right hand side of (3.21) can be bounded by

$$d^{-\frac{3}{2}+8\varepsilon} + \sqrt{\mathbb{E}_k(A_k - \mathbb{E}_k A_k)^2(B_k - \mathbb{E}_k B_k)^2}\ \mathbb{P}(\mathcal{E}^C)$$

where $\mathbb{P}(\mathcal{E}^C) < d^{-D}$ for arbitrarily large D. Using this and the moment bounds on subexponential random variables, we can conclude

$$\Big|\mathbb{E}_k[A_k]\mathbb{E}_k[B_k] - \mathbb{E}_k[A_k B_k]\Big| = O(d^{-\frac{3}{2}+8\varepsilon})$$

uniformly in k and thus

$$\sum_{k=1}^{n} \Big|\mathbb{E}_k[A_k]\mathbb{E}_k[B_k] - \mathbb{E}_k[A_k B_k]\Big| = O(nd^{-\frac{3}{2}+8\varepsilon}). \tag{3.22}$$

Finally, we note that the remaining terms of (3.13), namely $(A_k - \mathbb{E}_k[A_k])\mathbb{E}_k[B_k]$ and $(B_k - \mathbb{E}_k[B_k])\mathbb{E}_k[A_k]$, are martingale increments with

$$\|(A_k - \mathbb{E}_k[A_k])\mathbb{E}_k[B_k]\|_{\psi_1} \leq Cd^{-\frac{3}{2}+2\varepsilon}, \qquad \|(B_k - \mathbb{E}_k[B_k])\mathbb{E}_k[A_k]\|_{\psi_1} \leq Cd^{-1+4\varepsilon}. \tag{3.23}$$

Applying the Martingale Bernstein inequality, we conclude for some other constants $C, c > 0$

$$\begin{aligned}&\mathbb{P}\left(\sup_{1\leq k\leq n}\left|\sum_{j=1}^{k}(A_j - \mathbb{E}_j[A_j])\mathbb{E}_j[B_j] + (B_j - \mathbb{E}_j[B_j])\mathbb{E}_j[A_j]\right| \geq t\right) \\ &\quad\leq 2\exp\left(-\min\left\{\frac{t}{C\max\sigma_{k,1}}, \frac{t^2}{C\sum_{k=1}^{n}\sigma_{k,1}^2}\right\}\right) \\ &\quad\leq 2\exp\left(-\min\left\{ctd^{1-4\varepsilon}, ct^2 d^{2-8\varepsilon}n^{-1}\right\}\right).\end{aligned} \tag{3.24}$$

Thus, for $n \leq d\log d$, we get

$$\sup_{1\leq k\leq n}\left|\sum_{j=1}^{k}(A_j - \mathbb{E}_j[A_j])\mathbb{E}_k[B_j] + (B_j - \mathbb{E}_j[B_j])\mathbb{E}_j[A_j]\right| \leq d^{-\frac{1}{2}+5\varepsilon} \tag{3.25}$$

with overwhelming probability. Finally, combining the bounds from (3.20), (3.22), (3.25), we conclude that, for $n \leq d \log d$,

$$\sup_{1 \leq k \leq n} |\mathcal{M}_k^{\text{quad},\tau}| \leq d^{-\frac{1}{2}+8\varepsilon} \tag{3.26}$$

with overwhelming probability. This completes the proof of part (ii) of the lemma.

For part (iii), we observe that $\Delta\mathcal{E}_k^{\text{quad},\tau} = \mathbb{E}_k[A_k B_k] - \mathbb{E}_k[A_k]\mathbb{E}_k[B_k] + O(d^{-2+4\epsilon})$, the error terms arising from $\boldsymbol{u}_k$ cross terms, so that the bound of $\mathcal{E}_k^{\text{quad},\tau}$ follows immediately from (3.22). □

3.3 Martingale for HSGD

Recall that the HSGD process $\{\boldsymbol{V}_t\}$ satisfies the differential equation

$$d\boldsymbol{V}_t = -\gamma \nabla \mathcal{R}(\boldsymbol{V}_t + \tilde{\boldsymbol{x}})dt + \gamma \sqrt{\tfrac{2}{d}\mathcal{P}(\boldsymbol{V}_t + \tilde{\boldsymbol{x}})\boldsymbol{K}}dB_t \tag{3.27}$$

where

$$\mathcal{P}(\boldsymbol{x}) = (\boldsymbol{x} - \tilde{\boldsymbol{x}})^T \boldsymbol{K}^2 (\boldsymbol{x} - \tilde{\boldsymbol{x}}) + \eta^2. \tag{3.28}$$

Using Itô's Lemma, this gives us

$$\begin{aligned} q(\boldsymbol{V}_t^\tau) = q(\boldsymbol{V}_0^\tau) &- \gamma \int_0^t (\nabla q(\boldsymbol{V}_s^\tau))^T \nabla \mathcal{R}(\boldsymbol{V}_s^\tau + \tilde{\boldsymbol{x}}) ds \\ &+ \tfrac{\gamma^2}{2} \int_0^t \mathcal{P}(\boldsymbol{V}_s^\tau + \tilde{\boldsymbol{x}}) \operatorname{tr}(\tfrac{1}{d}\boldsymbol{K}(\nabla^2 q)) ds + \mathcal{M}_t^{HSGD,\tau}, \end{aligned} \tag{3.29}$$

where

$$\mathcal{M}_t^{HSGD,\tau} = \gamma \int_0^t (\nabla q(\boldsymbol{V}_s^\tau))^T \sqrt{\tfrac{1}{d}\mathcal{P}(\boldsymbol{V}_s^\tau + \tilde{\boldsymbol{x}})\boldsymbol{K}} dB_s. \tag{3.30}$$

Lemma 3.2. *For any quadratic q with $\|q\|_{C^2} \leq 1$, after imposing the stopping time τ, the resulting martingale $\mathcal{M}_t^{HSGD,\tau}$ satisfies*

$$\sup_{0 \leq t \leq n/d} \left[\mathcal{M}_t^{HSGD,\tau}\right] \leq C d^{-\frac{1}{2}+3\varepsilon}, \tag{3.31}$$

provided $n \leq d \log d$.

Proof. This martingale has quadratic variation

$$\left[\mathcal{M}_t^{HSGD,\tau}\right] = \frac{\gamma^2}{d} \int_0^t \mathcal{P}(\boldsymbol{V}_s^\tau + \tilde{\boldsymbol{x}})(\nabla q(\boldsymbol{V}_s^\tau))^T \nabla^2 \mathcal{P}(\boldsymbol{V}_s^\tau + \tilde{\boldsymbol{x}})(\nabla q(\boldsymbol{V}_s^\tau)) ds \tag{3.32}$$

and, for all s, we have the bound

$$|\mathcal{P}(\boldsymbol{V}_s^\tau + \tilde{\boldsymbol{x}})| \leq \|\boldsymbol{V}_s^\tau\|^2 \|\boldsymbol{K}^2\| + \eta^2 \leq C_1 d^{2\varepsilon}, \tag{3.33}$$

Thus, we have

$$\sup_{0 \leq t \leq n/d} \left[\mathcal{M}_t^{HSGD,\tau}\right] \leq C(\log d) d^{-1+4\varepsilon} \tag{3.34}$$

almost surely. By the Gaussian tail bound for continuous martingales of bounded quadratic varation, $\mathbb{P}(\sup_t [\mathcal{M}_t^{HSGD,\tau}] > a) < \exp(-a^2/(2C(\log d)d^{-1+4\varepsilon}))$. Taking $a = d^{-\frac{1}{2}+3\varepsilon}$, we conclude that

$$\sup_{0 \leq t \leq n/d} \left[\mathcal{M}_t^{HSGD,\tau}\right] \leq C d^{-\frac{1}{2}+3\varepsilon} \tag{3.35}$$

with overwhelming probability. □

References

[Ada15] R. Adamczak. “A note on the Hanson-Wright inequality for random vectors with dependencies”. *Electron. Commun. Probab.* 20 (2015), no. 72, 13. DOI: 10.1214/ECP.v20-3829. MR3407216

[AYS20] K. Ahn, C. Yun, and S. Sra. “SGD with shuffling: optimal rates without component convexity and large epoch requirements”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 17526–17535.

[AGJ21] G. B. Arous, R. Gheissari, and A. Jagannath. “Online stochastic gradient descent on non-convex losses from high-dimensional inference”. *The Journal of Machine Learning Research* 22.1 (2021), pp. 4788–4838. MR4279757

[AGJ22] G. B. Arous, R. Gheissari, and A. Jagannath. “High-dimensional limit theorems for SGD: Effective dynamics and critical scaling”. In: *Advances in Neural Information Processing Systems*. Ed. by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho. 2022.

[Asm03] S. Asmussen. *Applied probability and queues*. Second. Vol. 51. Applications of Mathematics (New York). Stochastic Modelling and Applied Probability. Springer-Verlag, New York, 2003, pp. xii+438. MR1978607

[BGH23] K. Balasubramanian, P. Ghosal, and Y. He. “High-dimensional scaling limits and fluctuations of online least-squares SGD with smooth covariance”. *arXiv e-prints*, arXiv:2304.00707 [math.PR] (Apr. 2023).

[CCM21] M. Celentano, C. Cheng, and A. Montanari. “The high-dimensional asymptotics of first order methods with random data”. *arXiv e-prints*, arXiv:2112.07572 [math.PR] (Dec. 2021).

[EK09] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. John Wiley & Sons, 2009. MR0838085

[Ger+22] C. Gerbelot, E. Troiani, F. Mignacco, F. Krzakala, and L. Zdeborova. “Rigorous dynamical mean field theory for stochastic gradient descent methods”. *arXiv e-prints*, arXiv:2210.06591 [math-ph] (Oct. 2022).

[GOP21] M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. “Why random reshuffling beats stochastic gradient descent”. *Mathematical Programming* 186 (2021), pp. 49–84. MR4214476

[Lee+22] K. Lee, A. Cheng, C. Paquette, and E. Paquette. “Trajectory of mini-batch momentum: batch size saturation and convergence in high dimensions”. In: Ed. by S. Koyejo et al. Vol. 35. *Advances in Neural Information Processing Systems*, pp. 36944–36957, 2022. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/efcb76ac1df9231a24893a957fcb9001-Paper-Conference.pdf

[Nee10] D. Needell. “Randomized Kaczmarz solver for noisy linear systems”. *BIT Numerical Mathematics* 50 (2010), pp. 395–403. MR2640019

[NWS14] D. Needell, R. Ward, and N. Srebro. “Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm”. *Advances in Neural Information Processing Systems* 27 (2014). MR3439812

[Paq+21] C. Paquette, K. Lee, F. Pedregosa, and E. Paquette. “SGD in the large: average-case analysis, asymptotics, and stepsize criticality”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by M. Belkin and S. Kpotufe. Vol. 134. Proceedings of Machine Learning Research, pp. 3548–3626. PMLR, Aug. 2021, arXiv:2102.04396 [math.OC].

[Paq+22] C. Paquette, E. Paquette, B. Adlam, and J. Pennington. “Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties”. *arXiv e-prints*, arXiv:2205.07069 [math.ST] (May 2022), 64 pp.

[RR08] A. Rahimi and B. Recht. “Random features for large-scale kernel machines”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 20. 2008, pp. 1177–1184.

[RR12] B. Recht and C. Re. “Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences”. In: *Proceedings of the 25th Annual Conference on Learning Theory*. Ed. by S. Mannor, N. Srebro, and R. C. Williamson. Vol. 23. Proceedings of Machine Learning Research. Edinburgh, Scotland: PMLR, 25–27 Jun 2012, pp. 11.1–11.24.

[RM51] H. Robbins and S. Monro. “A stochastic approximation method”. *Ann. Math. Statist.* (1951). DOI: 10.1214/aoms/1177729586.

[SS20] I. Safran and O. Shamir. “How good is SGD with random shuffling?” In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 3250–3284.

[SV09] T. Strohmer and R. Vershynin. “A randomized Kaczmarz algorithm with exponential convergence”. *Journal of Fourier Analysis and Applications* 15.2 (2009), p. 262. MR2500924

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018. MR3837109

[WL19] C. Wang and Y. M. Lu. “The scaling limit of high-dimensional online independent component analysis*”. *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (Dec. 2019), 124011. DOI: 10.1088/1742-5468/ab39d6. MR4063609

[WML17] C. Wang, J. Mattingly, and Y. M. Lu. “Scaling limit: Exact and tractable analysis of online learning algorithms with applications to regularized regression and PCA”. *arXiv e-prints*, arXiv:1712.04332 [cs.LG] (Dec. 2017).

[YSJ21] C. Yun, S. Sra, and A. Jadbabaie. “Open problem: Can single-shuffle SGD be better than reshuffling SGD and gd?” In: *Conference on Learning Theory*. PMLR. 2021, pp. 4653–4658.