EXACT RISK CURVES OF SIGNSGD IN HIGH-DIMENSIONS: QUANTIFYING PRECONDITIONING AND NOISE-COMPRESSION EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, SIGNSGD has garnered interest as both a practical optimizer as well as a simple model to understand adaptive optimizers like ADAM. Though there is a general consensus that SIGNSGD acts to precondition optimization and reshapes noise, quantitatively understanding these effects in theoretically solvable settings remains difficult. We present an analysis of SIGNSGD in a high dimensional limit, and derive a limiting SDE and ODE to describe the risk. Using this framework we quantify four effects of SIGNSGD: effective learning rate, noise compression, diagonal preconditioning, and gradient noise reshaping. Our analysis is consistent with experimental observations but moves beyond that by quantifying the dependence of these effects on the data and noise distributions. We conclude with a conjecture on how these results might be extended to ADAM.

1 INTRODUCTION

The success of deep learning has been driven by the effectiveness of relatively simple stochastic optimization algorithms. Stochastic gradient descent (SGD) with momentum can be used to train models like ResNet50 with minimal hyperparameter tuning. The workhorse of modern machine learning is ADAM, which was designed to give an approximation of preconditioning with a diagonal, online approximation of the Fisher information matrix (Kingma, 2014). Additional hypotheses for the success of ADAM include its ability to maintain balanced updates to parameters across layers and its potential noise-mitigating effects (Zhang et al., 2020b; 2024). Getting a quantitative, theoretical understanding of Adam and its variants is hindered by their complexity. While the multiple exponential moving averages are easy to implement, they complicate analysis.

The practical desire for simpler, more efficient learning algorithms as well as the theoretical desire for simpler models to analyze have led to a resurgence in the study of SIGNSGD. SIGNSGD is a variant of SGD where the stochastic gradient is passed through the sign function σ, leading to an update vector of ± 1s. On average, SIGNSGD's updates at every step have positive dot product with the average SGD step, but it can have dramatically different convergence properties (Bernstein et al., 2018a; Karimireddy et al., 2019). Multiple studies point towards sign-based methods as an effective proxy given that the sign component of the gradient has been shown to play an important role in ADAM (Kunstner et al., 2023; Balles & Hennig, 2018; Bernstein et al., 2018b). SIGNSGD is also the basis for new practical methods; the LION algorithm (Chen et al., 2023) combines SIGNSGD with multiple exponential moving averages, and SIGNSGD + momentum was used to train LLMs with performance comparable to ADAM (Zhao et al., 2024).

Despite the promise of SIGNSGD, a detailed quantitative understanding of its dynamics in realistic settings remain elusive—in particular the nature of the preconditioning and the effect of the σ function on the noise are not well understood. A crucial first step is to understand these effects on quadratic optimization problems.

Motivated by these questions, we provide the first analysis of the learning dynamics of SIGNSGD in a high-dimensional stochastic setting (Section 2). We make the following contributions:

- We derive a limiting stochastic differential equation (SDE) for SIGNSGD and combine it with a concentration result to derive a deterministic ordinary differential equation (ODE) that describes the dynamics of the risk in our setting (Section 3).
- We compare the dynamics of SIGNSGD and vanilla SGD, isolating 4 effects: effective learning rate, noise-compression, diagonal preconditioning, and gradient noise reshaping (Section 4).
- We quantitatively analyze these four effects and their contributions to learning, including exact results in specific settings (remainder of Section 4).

Our work addresses significant technical challenges in analyzing both the preconditioning and noise transformation effects of SIGNSGD. Our analysis is consistent with more general experimental observations about adaptive methods, but provides a more quantitative understanding in our setting. We conclude with a discussion of the implications of our results for future study of adaptive algorithms, including a conjecture on the limiting form of ADAM in an equivalent setting.

2 PROBLEM SETUP

Our work considers linear regression using the mean-squared loss $\mathcal{L}$ in the one-pass scenario, where data is not reused. SIGNSGD, without mini-batching, is first initialized by some $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and then follows the update rule:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k' \sigma\left(\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_k, \mathbf{x}_{k+1}, y_{k+1})\right), \qquad \mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y) = \| \langle \mathbf{x}, \boldsymbol{\theta} \rangle - y \|^2/2, \tag{1}$$

where σ denotes the sign function applied element-wise and $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_k, \mathbf{x}_{k+1}, y_{k+1}) = (\langle \boldsymbol{\theta}_k, \mathbf{x}_{k+1} \rangle - y_{k+1})\mathbf{x}_{k+1}$.

We will assume that the samples $\{(\mathbf{x}_k, y_k)\}_{k \geq 0}$, consisting of data $\mathbf{x}_k$ and targets y_k, satisfy the following:

Assumption 1. *The data $\mathbf{x}$ are mean 0 and Gaussian with positive definite covariance matrix $\mathbf{K} \in \mathbb{R}^{d \times d}$. The targets y are generated by $y = \langle \mathbf{x}, \boldsymbol{\theta}_* \rangle + \epsilon$, where $\boldsymbol{\theta}_*$ is the ground-truth and ϵ the label noise.*

Definition 1. *Define the population risk and the noiseless risk:*

$$\mathcal{P}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x},y)}\left[(\langle \mathbf{x}, \boldsymbol{\theta} \rangle - y)^2\right]/2 \qquad \textit{and} \quad \mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}}\left[\langle \mathbf{x}, \boldsymbol{\theta} - \boldsymbol{\theta}_* \rangle^2\right]/2. \tag{2}$$

Although our theory is framed in the setting of Gaussian data, as we will see the results are still a good description for real-world, *a priori* non-Gaussian settings (Figure 1). This is an instance of *universality*, wherein the details of the data distribution do not affect the precise high-dimensional limit law (see discussion in Tao (2023) section 2.2). Formalizing this is left to future work.

In contrast, the distribution of the label noise has a nontrivial impact on the behavior of the process. We shall require that the noise is well-behaved in a neighborhood around 0.

Assumption 2. *There exists $a_0 > 0$ such that the law of the noise ϵ has an almost-everywhere C^2 density on $(-a_0, a_0)$.*

Assumption 2 ensures our SDE (7) is Lipschitz (c.f. Lemma 8) and applies to many distributions; it encompasses heavy-tailed distributions such as α-stable laws, and we make no assumptions on any tail properties of the noise. Due to the non-smoothness of the σ function at 0, extraordinary behavior of the noise near 0 will lead to degraded performance of SIGNSGD as the risk vanishes. At the cost of a less-informative theorem, it is possible to drop Assumption 2; see Theorem 6 in the Appendix.

As we will see, an important characterizing feature of SIGNSGD is its effect on the covariance of the signed stochastic gradients. We introduce the following transformations on $\mathbf{K}$:

$$\overline{\mathbf{K}} \equiv \mathbf{D}^{-1}\mathbf{K} \quad \text{and} \quad \mathbf{K}_\sigma \equiv \left[\frac{\pi}{2}\mathbb{E}_{\mathbf{x}}[\sigma(\mathbf{x}_i)\sigma(\mathbf{x}_j)]\right]_{i,j} = \left[\arcsin\left(\frac{\mathbf{K}_{ij}}{\sqrt{\mathbf{K}_{ii}\mathbf{K}_{jj}}}\right)\right]_{i,j}, \tag{3}$$

where $\mathbf{D} = \sqrt{\operatorname{diag}(\mathbf{K})}$. We remark that $\overline{\mathbf{K}}$ is similar in the matrix-sense to $\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}}$, thus $\overline{\mathbf{K}}$ has all real, positive eigenvalues. $\mathbf{K}_\sigma$ is proportional to the covariance of $\sigma(\mathbf{x})$. We assume some properties of the matrices $\mathbf{K}$, $\overline{\mathbf{K}}$, and $\mathbf{K}_\sigma$.

Assumption 3. *Suppose:*

i). The spectrum of $\mathbf{K}$ *is bounded from above and away from* 0 *independently of* d*.*

ii). The sign-data matrix $\mathbf{K}_\sigma$ *also has operator norm bounded independent of* d*.*

iii). The resolvent of $\overline{\mathbf{K}}$ *defined by* $\mathbf{R}(z;\overline{\mathbf{K}}) = (\overline{\mathbf{K}} - z\mathbf{I})^{-1}$ *satisfies*

$$\max_{i\le d}\max_{i\neq j}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_{ij}\right\| = O\left(\frac{d^{\delta_0}}{\sqrt{d}}\right), \tag{4}$$

for all $z \in \partial B_{2\|\overline{\mathbf{K}}\|}$ *and for some* $\delta_0 < 1/12$*. (Equivalently, one may instead assume the same bounds with* $\overline{\mathbf{K}}$ *replaced by* $\mathbf{K}$*).*

The upper bound on $\mathbf{K}$ in Assumption 3 *(i)* is standard and can always be achieved by rescaling the risk. But the lower bound is a nontrivial assumption that is necessary for analyzing how the σ function affects the stochastic gradient. Assumption 3 *(ii)* is convenient for the proof. A full understanding of when it holds is highly nontrivial; there exists some theory establishing when it holds for some random $\mathbf{K}$ Fan & Montanari (2019). Assumption 3 *(iii)* can be interpreted as a condition that the eigenvectors of $\mathbf{K}$ contain no low-dimensional structure: for example, it is satisfied with high probability if the eigenvectors of $\mathbf{K}$ are taken to be uniformly random. Additionally, it is trivially satisfied for any diagonal $\mathbf{K}$. For a further discussion, including applicability in real datasets, see (Paquette & Paquette, 2022, Figure 2).

We assume the learning rates have a high-dimensional limiting profile:

Assumption 4. *The learning rates follow*

$$\eta_t' = \eta(t/d)/d, \tag{5}$$

where $\eta : \mathbb{R}^+ \to \mathbb{R}^+$ *is a continuous bounded function. We will write* η_t *for* $\eta(t)$*.*

This scaling is critical: it ensures that as the problem size grows, both the bias and variance terms in the risk evolution are balanced (see e.g. Equation (23)).

Finally, we assume the initialization remains (stochastically) bounded across d:

Assumption 5. *The difference between* $\boldsymbol{\theta}_*$ *and initialization* $\boldsymbol{\theta}_0$ *satisfies*

$$\mathbb{P}\left(\left|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*)\right| \ge t\right) \le C\exp\left(-ct^2 d/\|\mathbf{R}(z;\overline{\mathbf{K}})_i\|^2\right), \tag{6}$$

for all $1 \le i \le d$ *with absolute, positive constants* c, C*.*

For example, this assumption holds for deterministic $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_*$ with a dimension-independent bound on $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$ (e.g., $\boldsymbol{\theta}_0 = 0$ and $\|\boldsymbol{\theta}_*\|$ bounded independently of d), or for random $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_*$ with a dimension-free subgaussian bound on $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|$.

3 SIGNHSGD

The analysis of SIGNSGD in high-dimensional settings presents a unique set of technical challenges and requires careful mathematical treatment. A core difficulty lies in the transformative effect of the sign operator on the gradient. Unlike traditional SGD, where the gradient direction remains consistent with the magnitude of the update, SIGNSGD changes the gradient's direction, via a *non-Lipschitz* compression operation. This compression not only changes the optimization landscape, but also impacts the interaction between the noise and the optimizer.

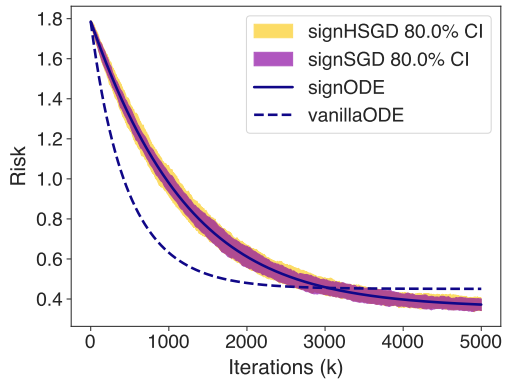
Nonetheless, we show that under the assumptions above, there is a *continuous* stochastic process Sign-Homogenized SGD (SIGNHSGD) which captures the high-dimensional behaviour of SIGNSGD.

Definition 2 (SIGNHSGD)**.** *We define* $\boldsymbol{\Theta}_t$ *as the solution of the stochastic differential equation:*

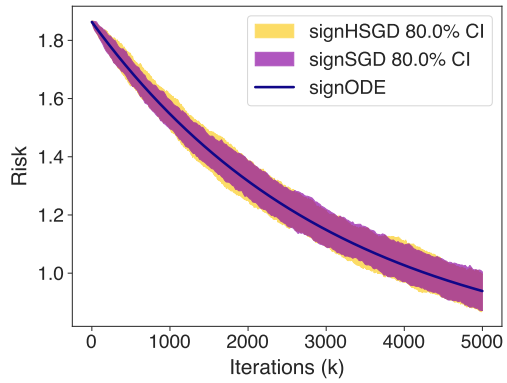
$$\mathrm{d}\boldsymbol{\Theta}_t = -\eta_t\frac{\varphi(\mathcal{R}(\boldsymbol{\Theta}_t))}{\sqrt{2\mathcal{R}(\boldsymbol{\Theta}_t)}}\overline{\mathbf{K}}(\boldsymbol{\Theta}_t - \boldsymbol{\theta}_*)\mathrm{d}t + \eta_t\sqrt{\frac{\mathbf{K}_\sigma}{\pi d}}\mathrm{d}\mathbf{B}_t, \qquad \textit{and } \boldsymbol{\Theta}_0 = \boldsymbol{\theta}_0, \tag{7}$$

where, with μ_ϵ *the law of* ϵ

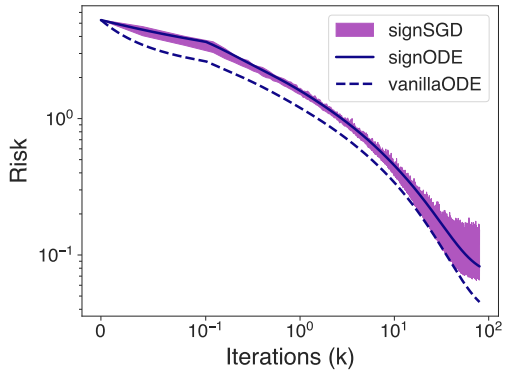
$$\varphi(\mathcal{R}(\boldsymbol{\Theta}_t)) = \frac{2}{\pi}\int_{\mathbb{R}}\exp\left(\frac{-y^2}{4\mathcal{R}(\boldsymbol{\Theta}_t)}\right)\mathrm{d}\mu_\epsilon(y) = \frac{2}{\pi}\mathbb{E}_\epsilon\left[\exp\left(\frac{-\epsilon^2}{4\mathcal{R}(\boldsymbol{\Theta}_t)}\right)\right]. \tag{8}$$



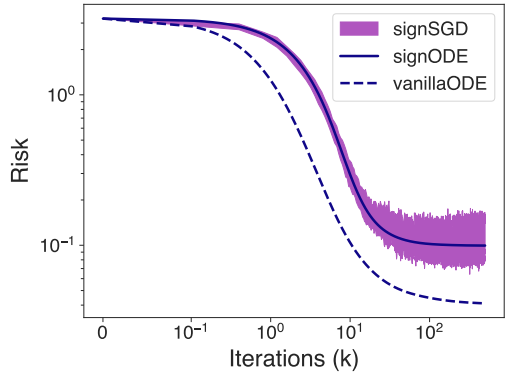
(a) Gaussian noise



(b) Non-centered Cauchy noise



(c) CIFAR10 data, assumed Gaussian noise



(d) IMDB movie reviews data, assumed GMM noise

Figure 1: Dynamics of the risk under SIGNSGD and SIGNHSGD on synthetic and real datasets. SIGNHSGD and its deterministic equivalent ODE are good models for the risk dynamics even for $d = 500$ (a, b) or on real datasets (c, d). The convergence of SIGNSGD for Cauchy noise (b) is remarkable given that SGD fails to converge there. The usefulness of the ODE on CIFAR10 and IMDB movie reviews is remarkable due to the non-Gaussian nature of the data, and the significant estimation of key quantities like $\boldsymbol{\theta}_*$ or ϵ. Details of these experiments may be found in Appendix H. See also Appendix B.2 for the definition of the VANILLAODE.

Remark 1. *In the case where $\epsilon \equiv 0$, we would take that $\varphi(x) \equiv 2/\pi$. While this ϵ does not satisfy Assumption 2, we formulate in the Appendix Theorem 6 which covers this case.*

It is worth noting that, in practice, φ is often easy and inexpensive to compute numerically; we compute it analytically for some common distributions (Figure 2). In general, it is simple to fit a Gaussian mixture model to your noise and use that to compute φ (Appendix H).

We can now state the first part of our main theorem:

Theorem 1 (Main Theorem, part 1)**.** *Given Assumptions 1–5 and choosing any fixed even moment $2p \in (0, d)$, there exists a constant $C(\overline{\mathbf{K}}, \epsilon) > 0$ such that for any $\delta \in (1/3, 1/2)$ and all $T > 0$,*

$$\sup_{0 \le t \le T} |\mathcal{R}(\boldsymbol{\theta}_{\lfloor td \rfloor}) - \mathcal{R}(\boldsymbol{\Theta}_t)| \le \frac{3d^{\delta} \|\mathbf{K}\|}{\sqrt{d}} \exp\left(C(\overline{\mathbf{K}}, \epsilon) \|\eta\|_{\infty} T\right), \tag{9}$$

with probability at least $1 - c(2p, \overline{\mathbf{K}}) d^{p(1/3-\delta)}$ for a constant $c(2p, \overline{\mathbf{K}})$ independent to d.

In other words, the risk curves of SIGNSGD are well approximated by the risk curves of SIGNHSGD and this approximation improves as dimension grows. Numerical simulations suggest that in practice this correspondence is strong even by $d = 500$ (Figure 1 (a), (b)).

One may also be interested in studying other statistics such as iterate norms or distance to optimality, for this we present a more generalized result across all quadratics in Theorem 5, which may be found in the Appendix.

Distribution	$\varphi(x)$
0	$\frac{2}{\pi}$
$N\left(0, \upsilon^2\right)$	$\frac{2\sqrt{2x}}{\pi\sqrt{2x+\upsilon^2}}$
Rademacher	$\frac{2}{\pi}\exp\left(-\frac{2}{4x}\right)$
Unif $(-1,1)$	$2\sqrt{\frac{x}{\pi}}\operatorname{erf}\left(\frac{1}{2\sqrt{x}}\right)$
$\sqrt{\text{Levy}(\lambda)}$	$\frac{2}{\pi}\exp\left(-\sqrt{\frac{\lambda}{2x}}\right)$

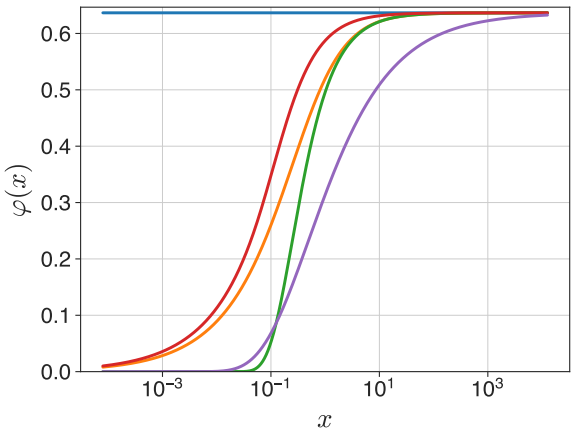


Figure 2: Examples of φ for simple noise distributions. $\sqrt{\text{Levy}}$ has Cauchy type-tails and vanishing density near 0. We note that $\varphi(x)$ is trivially bounded above by $\frac{2}{\pi}$ and converges to $\frac{2}{\pi}$ as $x \to \infty$; the rate of convergence at ∞ is related to the tail decay rate. At 0, $\varphi(x)/\sqrt{x}$ converges to the density of the noise at 0 scaled by $2/\pi$.

The risk curves of both SIGNSGD and SIGNHSGD concentrate around the same deterministic path. We will refer to this deterministic path as R_t, the *deterministic equivalent* of SIGNSGD. We call R_t SIGNODE. In order to find the deterministic equivalent we introduce a family of scalars $\{r_i\}_{i=1}^d$ which loosely correspond to the magnitudes of the residual $\boldsymbol{\Theta}_t - \boldsymbol{\theta}_*$ projected onto an eigenbasis (see Appendix B for details). The sum of these scalars then gives the deterministic equivalent for the risk:

$$R_t = \sum_{i=1}^{d} r_i(t). \tag{10}$$

The scalars follow a coupled system of ODEs:

$$\frac{\mathrm{d}r_i}{\mathrm{d}t} = -2\eta_t \frac{\varphi(R_t)}{\sqrt{2R_t}}\lambda_i(\overline{\mathbf{K}})r_i + \eta_t^2 \frac{\mathbf{w}_i^{\mathrm{T}}\mathbf{K}_\sigma\mathbf{K}\mathbf{u}_i}{\pi d}, \qquad \text{for all } 1 \le i \le d, \tag{11a}$$

$$r_i(0) = \frac{1}{2}\left\langle \boldsymbol{\theta}_0 - \boldsymbol{\theta}_*, \mathbf{K}\mathbf{u}_i\right\rangle \left\langle \mathbf{w}_i, \boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\right\rangle, \qquad \text{for all } 1 \le i \le d, \tag{11b}$$

where $\lambda_i(\overline{\mathbf{K}})$, $\mathbf{u}_i$ and $\mathbf{w}_i$ are the eigenvalues and left/right eigenvectors of $\overline{\mathbf{K}}$ respectively. We remark that by a similar argument, we may derive a coupled system of ODEs that describe the risk of vanilla SGD (Collins-Woodfin & Paquette, 2023). We call the deterministic equivalent of vanilla SGD as VANILLAODE. See Appendix B for the formulation.

We can now present a deterministic version of Theorem 1:

Theorem 2 (Main Theorem, part 2)**.** *Let R_t be given by* (10) *and* (11)*. Then given Assumptions 1–5 and choosing any fixed even moment $2p \in (0, d)$ there exists a constant $C(\overline{\mathbf{K}}, \epsilon) > 0$ such that for any $\delta \in (1/3, 1/2)$ and all $T > 0$,*

$$\sup_{0 \le t \le T} \left|\mathcal{R}(\boldsymbol{\theta}_{\lfloor td \rfloor}) - R_t\right| \le \frac{3d^\delta \left\|\mathbf{K}\right\|}{\sqrt{d}} \exp\left(C(\overline{\mathbf{K}}, \epsilon)\left\|\eta\right\|_\infty T\right), \tag{12}$$

with probability at least $1 - c(2p, \overline{\mathbf{K}})d^{p(1/3-\delta)}$ for a constant $c(2p, \overline{\mathbf{K}})$ independent to d.

This ODE captures the behavior of the risk even at finite $d = 500$ (Figure 1 (a), (b)). Moreover, it seems to capture the behavior of high dimensional linear regression on real, non-Gaussian datasets as well (Figure 1 (c), (d)).

4 COMPARING SIGNSGD TO VANILLA SGD

To produce an apples-to-apples comparison, we compare the SIGNHSGD to the analogous SDE for vanilla SGD from Collins-Woodfin & Paquette (2023):

$$\mathrm{d}\boldsymbol{\Theta}_t^{\mathrm{SGD}} = -\eta_t^{\mathrm{SGD}} \times \mathbf{K}(\boldsymbol{\Theta}_t^{\mathrm{SGD}} - \boldsymbol{\theta}_*)\mathrm{d}t + \eta_t^{\mathrm{SGD}} \times \sqrt{\frac{2\mathbf{K}\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{SGD}})}{d}}\mathrm{d}\mathbf{B}_t \quad \text{and } \boldsymbol{\Theta}_0^{\mathrm{SGD}} = \boldsymbol{\theta}_0. \tag{13}$$

To control for the adaptive-scheduling inherent in SIGNSGD, we run vanilla SGD with a risk dependent learning rate schedule η_t^{SGD} given by

$$\eta_t^{\mathrm{SGD}} = \frac{2}{\pi}\frac{\eta_t}{\sqrt{2\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{SGD}})}} = \frac{2}{\pi}\frac{\eta_t}{\sqrt{\mathbb{E}\|\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\mathbf{x},y)\|^2}}, \tag{14}$$

which is to say that we scale the steps in SGD inversely proportional to the norm of the gradients. We note that the $\mathcal{P}$-risk requires the noise ϵ to have finite variance $\mathfrak{v}$; indeed if the variance is infinite, then SIGNSGD is overwhelmingly favored, see (Zhang et al., 2020a, Remark 1). Training SIGNSGD with learning rate η_t and SGD with learning rate η_t^{SGD}, we can use (7) to write, with ψ as in (16)),

$$\mathrm{d}\boldsymbol{\Theta}_t^{\mathrm{SGD}} = -\eta_t^{\mathrm{SGD}} \times \mathbf{K}(\boldsymbol{\Theta}_t^{\mathrm{SGD}} - \boldsymbol{\theta}_*)\mathrm{d}t + \eta_t\sqrt{\frac{4\mathbf{K}}{\pi^2 d}}\mathrm{d}\mathbf{B}_t \tag{15a}$$

$$\mathrm{d}\boldsymbol{\Theta}_t = -\eta_t^{\mathrm{SGD}} \times \underbrace{\psi(\mathcal{R}(\boldsymbol{\Theta}_t))}_{\epsilon\text{- compress.}} \times \underbrace{\mathbf{D}^{-1}}_{\text{D.Precond.}} \times \mathbf{K}(\boldsymbol{\Theta}_t - \boldsymbol{\theta}_*)\mathrm{d}t + \eta_t\underbrace{\sqrt{\frac{\mathbf{K}_\sigma}{\pi d}}}_{\text{Reshape}}\mathrm{d}\mathbf{B}_t. \tag{15b}$$

We summarize the precise effects below:

Effective learning rate: The effective learning rate of SIGNSGD can be considered as risk dependent, effectively matching the expected ℓ^2–norm of a gradient.

ϵ- compression: The distribution of the label noise (be it from model-misspecification or otherwise) rescales the bias term. Formally, letting $\mathfrak{v}^2 = \mathbb{E}[\epsilon^2]$,

$$\psi(x) = \frac{\pi\varphi(x)\sqrt{2x+\mathfrak{v}^2}}{2\sqrt{2x}}. \tag{16}$$

Diagonal preconditioner: The matrix $\mathbf{D}^{-1}$ gives the diagonal preconditioner ($\mathbf{D}_{ii} = \sqrt{\mathbf{K}_{ii}}$).

Gradient noise reshaping: Finally, passing the gradient through the σ function results in a different covariance structure to the gradients, which is accounted for in the differing diffusion term.

Although all the effects appear in concert in SIGNSGD, we will now attempt to isolate and address each one separately in the following sections.

4.1 EFFECTIVE LEARNING RATE AND CONVERGENCE

We recall that to match the learning rate of SGD to SIGNSGD, we had to use the identification (14),

$$\eta_t^{\mathrm{SGD}} = \frac{2}{\pi}\frac{\eta_t}{\sqrt{2\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{SGD}})}}.$$

In particular, the effective learning rate gets smaller when the optimizer's position is far from optimality and gets larger as it gets closer. In the convex setting this is generally undesirable at both extremes. When far from optimality, the algorithm slows far beyond what would tend to be favorable, while at small risks this behavior can impede convergence. On the other hand, it can easily be rectified by appropriately rescaling the SIGNSGD learning rate η_t by the square root of the risk.

In a nonconvex setting, identifying $2\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{SGD}})$ with the expected square-norm of the gradients (c.f. (14)) one possible benefit of this schedule is that it may be helpful in dynamically adjusting to saddle manifolds in the loss landscape.

4.1.1 STATIONARY POINT OF SIGNSGD

If the learning rate is *any* constant $\eta_t \equiv \eta$, we have a unique stationary point of the ODE system (11a) which is locally attractive. The η dependence of this stationary point demonstrates the effect of an aggressive learning rate, which is accentuated in the presence of small noise variance $\mathfrak{v}$.

Theorem 3. *With fixed learning rate $\eta_t \equiv \eta \in (0,\infty)$ and $\epsilon \sim N(0,\mathfrak{v}^2)$, the ODEs have a unique stationary point $[s_i : 1 \le i \le d]$ given by Equation (212). Then, the limiting risk, $R_\infty = \sum s_i$, is given by*

$$R_\infty = \frac{\pi\eta}{32d}\operatorname{Tr}(\mathbf{D})\left(\frac{\pi\eta\operatorname{Tr}(\mathbf{D})}{2d} + \sqrt{\frac{\pi^2\eta^2\operatorname{Tr}(\mathbf{D})^2}{4d^2} + 16\mathfrak{v}^2}\right). \tag{17}$$

Notice that the limiting risk's dependence on η changes depending on the relationship between η and $\mathfrak{v}$, for small η it will be proportional to η.

4.2 ϵ-COMPRESSION

The influence of the distribution of the noise ϵ on the optimization, in the case of finite variance, can be summarized by (c.f. (16) and (8))

$$\psi(\mathcal{R}) = \mathbb{E}\left[\exp\left(\frac{-\epsilon^2}{4\mathcal{R}}\right)\right] \times \sqrt{1 + \frac{\mathbb{E}[\epsilon^2]}{2\mathcal{R}}}. \tag{18}$$

When $\psi < 1$, the descent term of (15b) is decreased, and hence SIGNSGD is slowed with respect to SGD with learning rate η_t^{SGD}. Conversely, when $\psi > 1$ the descent term is increased, and SIGNSGD is favored. When $\mathbb{E}[\epsilon^2] = \infty$, ψ can be interpreted as ∞, corresponding to overwhelming SIGNSGD favor, although the quantitative meaning in (15b) breaks down.

In the Gaussian case $\psi = 1$; we can interpret ψ as the effect that *deviation from Gaussianity* has on the drift term of the SDE. We note that all the influence of the label noise ϵ on SIGNSGD is entirely through (18) which in turn only depends on ϵ^2. Hence SIGNSGD *symmetrizes* the noise distribution.

In general, a full comparison of SGD and SIGNSGD requires optimizing the learning rates of both algorithms independently. We will show that in the case of isotropic data, this procedure is tractable and produces a different threshold $\psi = \frac{\pi}{2}$ above which SIGNSGD is favored (see Equation (26)).

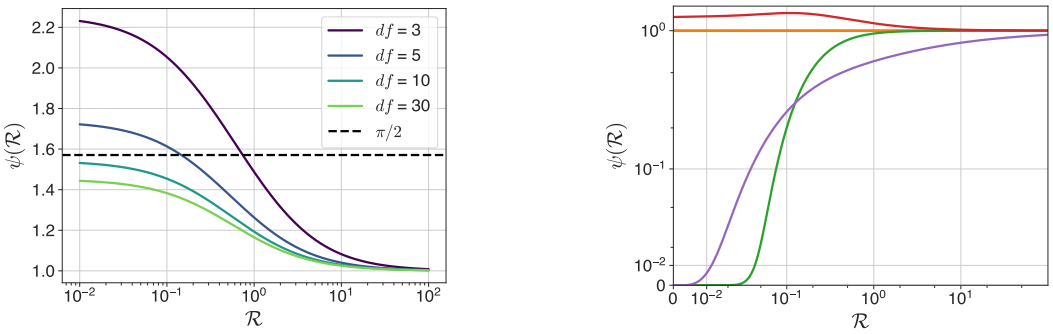


Figure 3: Left: ψ for Student-t. Here ψ is always greater than 1 and ϵ-compression accelerates SIGNSGD. For sufficiently small df, $\psi > \pi/2$ over some range of $\mathcal{R}$ and SIGNSGD also converges faster than SGD in the isotropic setting. Right: ψ for $N(0, \mathfrak{v}^2)$, Rademacher, $\mathrm{Unif}(-1, 1)$, $\sqrt{\mathrm{Levy}}$. Only $\mathrm{Unif}(-1, 1)$ admits $\psi > 1$.

Setups favoring SIGNSGD. In the presence of heavy tails, $\psi(\mathcal{R})$ can be large and hence very SIGNSGD favored. Indeed, among some parametric classes, such as the Student-t family, this is observed numerically to always be larger than 1 (Figure 3, left) and increase to ∞ as the kurtosis increases. More generally, as $\mathcal{R}$ tends to 0, letting $f_\epsilon(0)$ be the density of the noise at 0, one has

$$\psi(\mathcal{R}) \to_{\mathcal{R}\to\infty} \sqrt{2\pi} f_\epsilon(0)\mathbb{E}[\epsilon^2], \tag{19}$$

which is large for SIGNSGD when ϵ has large kurtosis.

Conversely, for *all* distributions, we also observe that when the risk is relatively large, SIGNSGD is always modestly favored over SGD under the η^{SGD} learning rate as we have

$$1 \le \mathbb{E}\left[1 - \frac{\epsilon^2}{4\mathcal{R}}\right] \times \sqrt{1 + \frac{\mathbb{E}[\epsilon^2]}{\mathcal{R}}} \le \psi(\mathcal{R}) \le \sqrt{1 + \frac{\mathbb{E}[\epsilon^2]}{\mathcal{R}}}, \quad \text{for all} \quad \frac{\mathbb{E}[\epsilon^2]}{\mathcal{R}} \le \frac{3}{2}. \tag{20}$$

Setups where SIGNSGD does not improve. For light-tailed noises, the factor ψ can only mildly favor SGD. A density f on $\mathbb{R}$ is called *log-concave* if it can be written as e^g for concave g (see Saumard & Wellner (2014) for discussion). The exponential, uniform and many other canonical noise distributions are log-concave. Note these decay no slower than exponentially at infinity. Then

as $\varphi(\mathcal{R})/\sqrt{4\pi\mathcal{R}}$ is the density at 0 of a log-concave density, we have from (Saumard & Wellner, 2014, Proposition 5.2),

$$\psi(\mathcal{R}) \leq \sqrt{2\pi}. \tag{21}$$

Hence, for these distributions, while there may be limited gains from using SIGNSGD, they are bounded by an absolute constant factor.

Setups where SIGNSGD is catastrophic. In the situation that the noise is bounded away from 0 by some δ, it follows that we have the upper bound:

$$\psi(\mathcal{R}) \leq e^{-\frac{\delta^2}{4\mathcal{R}}} \times \sqrt{1 + \frac{\mathbb{E}[\epsilon^2]}{2\mathcal{R}}}. \tag{22}$$

This tends to 0 *exponentially* in $1/\mathcal{R}$ (e.g. see the Rademacher case of Figure 3). For such noise distributions, SIGNSGD will effectively experience a floor on the risk, which is completely induced by distributional properties of the noise (and unrelated to the underlying optimization problem geometry). In this situation, SGD is heavily favored for small risks, which would be seen late in training.

Scheduling SIGNSGD. We have discussed adjusting the SGD learning rate to match the behaviour of SIGNSGD. However, when using SIGNSGD there is the reciprocal question of how to select its learning rate. We briefly discuss this in the case of isotropic data $\mathbf{K} = \mathbf{I}_d$, in which $\overline{\mathbf{K}} = \mathbf{K}$ and $\mathbf{K}_\sigma = \frac{\pi}{2}\mathbf{I}_d$ which allows us to isolate the effects of the label noise. It is easy to check that the d-system of ODEs for SIGNSGD in (10) may be reduced to the following single ODE:

$$\frac{\mathrm{d}R_t}{\mathrm{d}t} = -\frac{2\eta_t\varphi(R_t)}{\sqrt{2R_t}}R_t + \frac{\eta_t^2}{2}, \qquad R_0 = \mathcal{R}(\boldsymbol{\theta}_0). \tag{23}$$

If we greedily optimize in η_t we arrive at

$$\frac{\mathrm{d}R_t}{\mathrm{d}t} = -\varphi(R_t)^2 R_t, \qquad \text{where} \quad \eta_t^* = \varphi(R_t)\sqrt{2R_t}. \tag{24}$$

So generally for large risks, the optimal stepsize compensates for the effective gradient rescaling in (14). This compensation is seen for all risks in the Gaussian ϵ setting.

As a point of comparison, we may repeat the same procedure for the SGD risk ODE R^{S} with learning rate η^{S}, which can be derived from (13):

$$\frac{\mathrm{d}R_t^{\mathrm{S}}}{\mathrm{d}t} = -2\eta_t^{\mathrm{S}}R_t^{\mathrm{S}} + \frac{(\eta_t^{\mathrm{S}})^2}{2}(2R_t^{\mathrm{S}} + \mathfrak{v}^2) \xrightarrow{\text{optimizing in } \eta^{\mathrm{S}}} \frac{\mathrm{d}R_t^{\mathrm{S}}}{\mathrm{d}t} = -\frac{2R_t^{\mathrm{S}}}{2R_t^{\mathrm{S}} + \mathfrak{v}^2}R_t^{\mathrm{S}}. \tag{25}$$

Hence (24) can also be expressed as

$$\frac{\mathrm{d}R_t}{\mathrm{d}t} = -\left(\frac{4}{\pi^2}\psi^2(R_t)\right) \times \frac{2R_t}{2R_t + \mathfrak{v}^2}R_t. \tag{26}$$

Thus the performance benefits of SIGNSGD having selected the optimal learning rate can again be reduced to a question of the magnitude of ψ, albeit with a crossover at $\psi = \pi/2$.

In the non-isotropic setting, locally greedy stepsizes can be very far from optimal, even with two eigenvalues (Collins-Woodfin et al., 2024). But we expect the conclusion of (26) remains mostly true in well-conditioned settings.

4.3 Diagonal preconditioner

Next, and strikingly, we see that SIGNSGD performs a diagonal preconditioning step on the gradients, with the preconditioner given by $\mathbf{D}_{ii} = \sqrt{\overline{\mathbf{K}}_{ii}} = \sqrt{\mathbb{E}[\mathbf{x}_i^2]}$, where $\mathbf{x}$ is a sample. To produce this bias term in SGD, we would need to run the algorithm

$$\theta_{k+1} = \theta_k - \eta_k \mathbf{D}^{-1}\left(\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}, y)\right). \tag{27}$$

We expect to see the same preconditioner in ADAM; for details, see Appendix E.

As $\overline{\mathbf{K}}$ appears naturally in (7), its spectrum regulates the rate of convergence of the optimization to stationarity. By utilizing our d-systems of ODEs we can establish the following convergence rate:

Theorem 4. *Assume $\epsilon \sim N(0, \mathfrak{v}^2)$ and let s_i be the stationary points to (11a). Then there is an absolute constant $c > 0$ so that if*

$$\eta \frac{\mathrm{Tr}(\overline{\mathbf{K}})}{2d} \leq \min\left\{c, \frac{4\mathfrak{v}}{\pi}\right\}, \quad \textit{and} \quad R_0 \leq c\mathfrak{v}, \tag{28}$$

then we have, setting $R_\infty = \sum_{i=1}^d s_i$ to be the limit risk,

$$|R_t - R_\infty| \leq 2(R_0 + R_\infty) e^{-t\eta\lambda_{\min}(\overline{\mathbf{K}})/(\pi\mathfrak{v})}. \tag{29}$$

The proof is given in Appendix C. In contrast to vanilla SGD, where the risk converges (in a high-dimensional setting) with rate $\frac{1}{\overline{\kappa}}$, where $\overline{\kappa}(\mathbf{K}) = \frac{\mathrm{Tr}(\mathbf{K})}{d\lambda_{\min}(\mathbf{K})}$ is the average condition number (Paquette et al. (2022)). Theorem 4 states that the risk of SIGNSGD converges at a rate $\frac{\mathrm{Tr}(\overline{\mathbf{K}})}{d\lambda_{\min}(\overline{\mathbf{K}})}$, after selecting the largest allowed η.

Settings in which the preconditioned $\overline{\mathbf{K}}$ is preferable. Theorem 4 shows that the rate of convergence is governed entirely by $\overline{\mathbf{K}}$. The clearest setting when this is favourable is if $\mathbf{K}$ is diagonal, so that $\overline{\mathbf{K}} = \sqrt{\mathbf{K}}$. In this case, the convergence rate is, up to constants

$$\frac{1}{d}\frac{\mathrm{Tr}(\overline{\mathbf{K}})}{\lambda_{\min}(\overline{\mathbf{K}})} = \frac{1}{d}\frac{\mathrm{Tr}(\overline{\mathbf{K}})}{\sqrt{\lambda_{\min}(\mathbf{K})}} \leq \frac{\sqrt{\frac{1}{d}\mathrm{Tr}(\mathbf{K})}}{\sqrt{\lambda_{\min}(\mathbf{K})}}. \tag{30}$$

Hence on diagonal problems, SIGNSGD attains a speedup commensurate with optimal deterministic convex optimization algorithms such as Heavy-Ball momentum or Conjugate gradient (Nocedal & Wright, 2006). As for stochastic algorithms, these are also attained using optimally tuned minibatch SGD with momentum (Lee et al., 2022).

A strictly diagonal $\mathbf{K}$ is not necessary to attain this speedup. Diagonally dominant matrices, which are well-known to benefit from Jacobi preconditioning (in which one would rescale by $\mathbf{D}^{-2}$), should see similar benefits. This supports the prior work of Balles et al. (2020) who show that SIGNSGD is effective when the the Hessian of the risk, which in our setting is $\mathbf{K}$, is sufficiently diagonally concentrated.

A second situation in which one may have substantial speedups are for block-tridiagonal $\mathbf{K}$, where the blocks are scaled by greatly differing constants; diagonal preconditioning by $\mathbf{D}$ partially corrects for this effect. It has been argued that one of the principal advantages of ADAM is that it correctly adapts learning rates across different layers of Transformers and MLPs (Zhang et al., 2024), which can have similar structures in their Jacobians.

Settings in which $\overline{\mathbf{K}}$ does not help. Like preconditioning generally, $\overline{\mathbf{K}}$ does not always have a smaller condition number than $\mathbf{K}$. See Appendix G for a counter example.

In addition, if the eigenvectors of $\overline{\mathbf{K}}$ are randomized to make a new covariance matrix $\mathbf{A}$, say by performing a uniformly random orthogonal change of basis, the entries of the diagonal of $\mathbf{A}$ will concentrate to be

$$\max_i \left|\mathbf{A}_{ii} - \frac{\mathrm{Tr}(\mathbf{K})}{d}\right| = O((\log d)d^{-1/2}), \tag{31}$$

and so the preconditioner $\mathrm{diag}(\mathbf{A})^{-1/2}$ does not affect the condition number of $\mathbf{A}$. Hence the diagonal preconditioner is effective when $\mathbf{A}$ has special structure represented in the basis in which the optimization is performed.

4.4 GRADIENT NOISE RESHAPING

Finally, there is gradient noise reshaping, wherein the SGD gradient noise matrix $\mathbf{K}$ is replaced by the matrix $\mathbf{K}_\sigma$ up to constants. This is a complicated mapping, and there is no short answer about the impact of this replacement. In Figure 4, we show a simulation of the spectra illustrating that for CIFAR10, a practical, non-diagonal dataset, passing from $\mathbf{K} \to \mathbf{K}_\sigma$ might affect the magnitudes of the eigenvalues but not their structure.

In the case that $\mathbf{K}$ itself is a sample covariance matrix, the matrix $\mathbf{K}_\sigma$ is strongly related to a *Kernel inner product matrix*, for which there is a large literature. This includes properties of bulk spectra (Karoui, 2010; Cheng & Singer, 2013), norms (Fan & Montanari, 2019) and more.

When $\mathbf{K}$ is a diagonal matrix then $\mathbf{K}_\sigma = \frac{\pi}{2}\mathbf{I}$ and so this can be considered a type of preconditioning of the gradient noise, albeit with a *more* aggressive preconditioner than $\mathbf{D}$.

We *expect* that for power-law type covariances, in which $\mathbf{K}$ has powerlaw spectral dependence and which are often seen in practice (e.g. in Figure 4), in language embeddings, and in image and video datasets, $\mathbf{K}_\sigma$ again has powerlaw spectra of the same exponent. Beyond the spectral distribution, replacing $\mathbf{K}$ by $\mathbf{K}_\sigma$ may also serve to slightly break the alignment of large directions of gradient variance from large gradient biases (they are perfectly aligned in SGD), which should be beneficial both to stability of the algorithm and performance.

5 Discussion

Our high-dimensional limit sheds a quantitative light on the precise ways in which SignSGD can be compared to SGD, via change of effective learning rate, noise compression, preconditioning, and reshaping of the gradient noise.

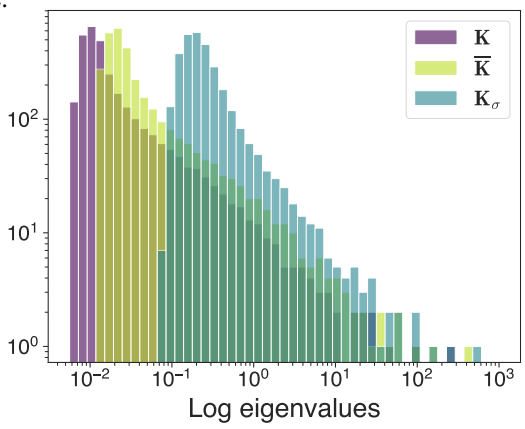


Figure 4: Log eigenvalues of $\mathbf{K}, \overline{\mathbf{K}}, \mathbf{K}_\sigma$ computed for the CIFAR10 dataset.

Theorem 2, the main technical contribution of this work, required substantial technical efforts. Although similar in formulation to existing work like Collins-Woodfin et al. (2024), there are technical complexities in working with the nonsmooth σ function: both in terms of deriving the relevant concentration of measure estimates (the textbook versions of which require smoothness) and in terms of the additional pathology of the resulting SignHSGD (especially the φ). We believe that a version of Theorem 2 is true in much greater generality than we have proven it, even for the linear setting: two desirable mathematical generalizations are quantifying dimension *intrinsically* (instead of through the ambient dimension) and generalizing the theory to settings of non-Gaussian data.

Our high-dimensional SDE differs from the previously studied Weak-Approximation (WA) SDE framework (Li et al., 2019; Malladi et al., 2022) in some key ways: first, our approximation improves with dimension, whereas in WA one fixes a dimension and sends stepsize to 0. Secondly, (Malladi et al., 2022) does not provide an explicit optimization problem, while SignHSGD is fully determined given a learning rate schedule and covariance structure, which allows us to draw conclusions about SignHSGD applied to these optimization problems (Theorems 3, 4). Finally, previous works using WA to study adaptive algorithms like Adam fail to quantitatively or qualitatively capture the dynamics of SignSGD; see Appendix F for details.

Though our work focuses on the case of MSE loss and linear regression, there is a path towards extending results to more general settings using recent results in high-dimensional optimization (Collins-Woodfin et al., 2023). In practical settings, models undergo dramatic changes in local geometry during training; nonetheless, stability analysis of the linearized problem is still useful for understanding aspects of the non-linear dynamics of these systems (Cohen et al., 2022; Agarwala & Pennington, 2024).

Finally, our analysis of SignSGD gives hints towards understanding Adam in a similar setting. A heuristic analysis shows that Adam has a homogenized process similar to SignHSGD: it appears to share the preconditioner $\mathbf{D}$ while differing from SignHSGD by setting $\varphi \to 1$ and again modifying the shape of the gradient noise $\mathbf{K}_\sigma$ (Appendix E). Thus for well-behaved noises ϵ, SignSGD should be nearly path-identical to Adam; we note that Lion has been recently observed to do just that (Zhao et al., 2024). We leave investigation of Adam for future work.

REFERENCES

Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer New York, 2007. ISBN 9780387481166. doi: 10.1007/978-0-387-48116-6.

Atish Agarwala and Jeffrey Pennington. High dimensional analysis reveals conservative sharpening and a stochastic edge of stability. *arXiv preprint arXiv:2404.19261*, 2024.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent, 2020. URL https://openreview.net/forum?id=rJe9lpEFDH.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 2018a. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. *arXiv preprint arXiv:1810.05291*, 2018b.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ne6zeqLFCZ.

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

Elizabeth Collins-Woodfin and Elliot Paquette. High-dimensional limit of one-pass sgd on least squares, 2023. URL https://arxiv.org/abs/2304.06847.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ode for sgd learning dynamics on glms and multi-index models. *arXiv preprint arXiv:2308.08977*, 2023.

Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew W Mackenzie, Elliot Paquette, and Courtney Paquette. The high line: Exact risk and learning rate curves of stochastic adaptive learning rate algorithms. *arXiv preprint arXiv:2405.19585*, 2024.

Zhou Fan and Andrea Montanari. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173:27–85, 2019.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, 2019.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 – 50, 2010. doi: 10.1214/08-AOS648. URL https://doi.org/10.1214/08-AOS648.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a65YK0cqH8g.

Kiwon Lee, Andrew Cheng, Elliot Paquette, and Courtney Paquette. Trajectory of mini-batch momentum: batch size saturation and convergence in high dimensions. *Advances in Neural Information Processing Systems*, 35:36944–36957, 2022.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40): 1–47, 2019.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms. *Advances in Neural Information Processing Systems*, 35: 7697–7711, 2022.

Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006. ISBN 978-0387-30303-1; 0-387-30303-0.

Courtney Paquette and Elliot Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. *Advances in Neural Information Processing Systems*, 34:9229–9240, 2021.

Courtney Paquette and Elliot Paquette. High-dimensional optimization. *SIAM Views and News*, 20: 16pp, December 2022. https://siagoptimization.github.io/assets/views/ViewsAndNews-30-1.pdf.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? exact risk trajectories of sgd in high dimensions. *Advances in Neural Information Processing Systems*, 35:35984–35999, 2022.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society, 2023.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15383–15393. Curran Associates, Inc., 2020a.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *arXiv preprint https://arxiv.org/abs/1912.03194*, 2020b.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. *arXiv preprint arXiv:2402.16788*, 2024.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for language models. *arXiv preprint arXiv:2407.07972*, 2024.

OVERVIEW OF SUPPLEMENTARY MATERIAL

The supplementary material is primarily dedicated to the proofs of the main theorems, Theorem 1 and 2. Here we give the organization of the appendices.

In Appendix A, we give the proof of these main theorems, including their extensions in Theorem 5 and 6. The key approximations to the update rules of SIGNSGD are given in Appendix A.1, including the key technical Lemma 1. In Appendix A.2, we show how these tools are used to give the main proof (but we defer the estimates on the stochastic errors to Appendix A.4), culminating in Lemma 3, which in fact proves the main theorem statement (that of Theorem 5). In Appendix A.3, we discuss the extension Theorem 6 – as this is a modification of the proof of Theorem 5, we do not go into details.

In Appendix B, we give the derivation of the ODEs SIGNODE and VANILLAODE from their homogenized counterparts, and discuss the proof of Theorem 2, which follows the same strategy as Theorem 5 (for full details of this type of ODE comparison, see Collins-Woodfin et al. (2023)). Here also, we discuss the derivation of the VANILLAODE, which is a special case of Collins-Woodfin et al. (2023).

In Appendix C, we proof the analysis of the SIGNODE and VANILLAODE that gives its limit level (Theorem 3) and a local convergence rate (Theorem 4).

In Appendix D, we provide additional supporting simulations, corroborating aspects of the main theorems.

In Appendix E, we give a heuristic derivation of the high-dimensional limit of ADAM.

In Appendix F we show the "Weak Approximation" theory of ADAM produces a different SDE prediction (see the discussion there as well).

In Appendix G, we give an example of a matrix where diagonal preconditioning hurts.

Finally in Appendix H, we give some additional information on how the experiments were performed.

A Proof of Main Theorem

A.1 Approximation of the conditional updates

For simplicity of our proofs, we will assume η is constant. The proof remains unchanged if η is defined as in Assumption 4. For convenience for the reader and to avoid confusion, we provide the typical notion of convergence in high-dimensions.

Definition 3. *An event $A \subset \mathbb{R}^d$ holds with high-probability, if there exists some $\delta > 0$ independent to d such that $\mathbb{P}(A) \geq 1 - Cd^{-\delta}$ for some C independent to d.*

Definition 4. *An event $A \subset \mathbb{R}^d$ holds with overwhelming-probability, if for all $\delta > 0$ there exists C_δ such that $\mathbb{P}(A) \geq 1 - C_\delta d^{-\delta}$.*

Following the SGD update rule (1), we start by computing the conditional update of the i-th entry of the iterates. For notational convenience, we define $\boldsymbol{\nu}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_*$. Moreover, when it is clear we will denote $\mathcal{R}_k = \mathcal{R}(\boldsymbol{\theta}_k)$. Then,

$$\mathbb{E}\left[\boldsymbol{\nu}_{k+1}^i - \boldsymbol{\nu}_k^i | \mathcal{F}_k\right] = -\frac{\eta}{d}\mathbb{E}\left[\sigma(\mathbf{x}_{k+1}^i)\sigma(\langle \mathbf{x}_{k+1}, \boldsymbol{\nu}_k\rangle - \epsilon)|\mathcal{F}_k\right] \tag{32}$$

$$= -\frac{\eta}{d}\mathbb{E}\left[\sigma(\mathbf{x}_{k+1}^i)\mathbb{E}\left[\sigma\left(\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i + \sum_{j\neq i}\boldsymbol{\nu}_k^j\mathbf{x}_{k+1}^j - \epsilon\right)\middle|\mathcal{F}_k,\, \mathbf{x}_{k+1}^i\right]\middle|\mathcal{F}_k\right]. \tag{33}$$

Given that the data is Gaussian distributed, upon conditioning on $\mathcal{F}_k$ we see that

$$\sum_{j\neq i}\boldsymbol{\nu}_k^j\mathbf{x}_{k+1}^j \sim N(0, 2\mathcal{R}_k - 2\boldsymbol{\nu}_k^i\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle + \mathbf{K}_{ii}(\boldsymbol{\nu}_k^i)^2). \tag{34}$$

Additionally, if c_i is any constant, we can write

$$\sum_{j\neq i}\boldsymbol{\nu}_k^j\mathbf{x}_{k+1}^j = \left(\sum_{j\neq i}\boldsymbol{\nu}_k^j\mathbf{x}_{k+1}^j - c_i\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i\right) + c_i\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i. \tag{35}$$

Let $y^i = \sum_{j\neq i}\boldsymbol{\nu}_k^j\mathbf{x}_{k+1}^j - c_i\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i$. One may check that if $c_i = \frac{\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle - \mathbf{K}_{ii}\boldsymbol{\nu}_k^i}{\mathbf{K}_{ii}\boldsymbol{\nu}_k^i}$, then y^i is independent to $\mathbf{x}_{k+1}^i$ and $y^i \sim N(0, 2\mathcal{R}_k - \left\langle \overline{\mathbf{K}}_i, \boldsymbol{\nu}_k\right\rangle^2)$. Moreover, since y^i is independent to ϵ their difference $y_i - \epsilon$ has density given by

$$h_k^i(x) = \frac{1}{\sqrt{2\pi\left(2\mathcal{R}_k - \left\langle \overline{\mathbf{K}}_i, \boldsymbol{\nu}_k\right\rangle^2\right)}}\int_{\mathbb{R}}\exp\left(\frac{-(x+y)^2}{2\left(2\mathcal{R}_k - \left\langle \overline{\mathbf{K}}_i, \boldsymbol{\nu}_k\right\rangle^2\right)}\right)\mathrm{d}\mathcal{L}_\epsilon(y). \tag{36}$$

Using (36), we compute

$$\begin{aligned}
&\mathbb{E}\left[\sigma\left(\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i + \sum_{j\neq i}\boldsymbol{\nu}_k^j\mathbf{x}_{k+1}^j - \epsilon\right)\middle|\mathcal{F}_k, \mathbf{x}_{k+1}^i\right] \\
&= \mathbb{E}\left[\sigma\left(y^i - \epsilon + (1+c_i)\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i\right)\middle|\mathcal{F}_k, \mathbf{x}_{k+1}^i\right] \\
&= \mathbb{P}_{\mathcal{F}_k, \mathbf{x}_{k+1}^i}\left(y_i - \epsilon > -(1+c_i)\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i\right) - \mathbb{P}_{\mathcal{F}_k, \mathbf{x}_{k+1}^i}\left(y_i - \epsilon \leq -(1+c_i)\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i\right) \\
&= \int_{-(1+c_i)\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i}^{\infty} h_k^i(x)\,\mathrm{d}x - \int_{-\infty}^{-(1+c_i)\boldsymbol{\nu}_k^i\mathbf{x}_{k+1}^i} h_k^i(x)\,\mathrm{d}x.
\end{aligned} \tag{37}$$

Define

$$H(s) = \int_{-s}^{\infty} h_k^i(x)\,\mathrm{d}x - \int_{-\infty}^{-s} h_k^i(x)\,\mathrm{d}x, \tag{38}$$

where upon differentiating it is easy to see that

$$H'(s) = 2h_k^i(-s). \tag{39}$$

Taylor expanding around 0 we obtain,

$$\begin{aligned}\mathbb{E}\left[\sigma\left(\boldsymbol{\nu}_k^i \mathbf{x}_{k+1}^i + \sum_{j\neq i} \boldsymbol{\nu}_k^j \mathbf{x}_{k+1}^j - \epsilon\right)\middle|\mathcal{F}_k,\, \mathbf{x}_{k+1}^i\right] &= H((1+c_i)\boldsymbol{\nu}_k^i \mathbf{x}_{k+1}^i) \\ &= H(0) + 2h_k^i(0)\frac{\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle}{\mathbf{K}_{ii}}\mathbf{x}_{k+1}^i \\ &\quad + \frac{\mathrm{d}}{\mathrm{d}s}h_k^i(0)\left(\frac{\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle}{\mathbf{K}_{ii}}\mathbf{x}_{k+1}^i\right)^2 + R_{k+1}^i,\end{aligned} \tag{40}$$

where

$$R_{k+1}^i = O\left(\left(\frac{\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle}{\mathbf{K}_{ii}}\mathbf{x}_{k+1}^i\right)^3\right). \tag{41}$$

Plugging this back into (33) yields

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\nu}_{k+1}^i - \boldsymbol{\nu}_k^i|\mathcal{F}_k\right] &= -\frac{\eta}{d}\mathbb{E}\left[\sigma(\mathbf{x}_{k+1}^i)\left(H(0) + 2h_k^i(0)\frac{\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle}{\mathbf{K}_{ii}}\mathbf{x}_{k+1}^i\right)\middle|\mathcal{F}_k\right] \\ &\quad - \frac{\eta}{d}\mathbb{E}\left[\sigma(\mathbf{x}_{k+1}^i)\left(\frac{\mathrm{d}}{\mathrm{d}s}h_k^i(0)\left(\frac{\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle}{\mathbf{K}_{ii}}\mathbf{x}_{k+1}^i\right)^2 + R_{k+1}^i\right)\middle|\mathcal{F}_k\right] \\ &= -\frac{\eta}{d}2h_k^i(0)\frac{\langle \mathbf{K}_i, \boldsymbol{\nu}_k\rangle}{\mathbf{K}_{ii}}\mathbb{E}|\mathbf{x}_{k+1}^i| - \frac{\eta}{d}\mathbb{E}[\sigma(\mathbf{x}_{k+1}^i)R_{k+1}^i] \\ &= -\frac{\eta}{d}2h_k^i(0)\left\langle \overline{\mathbf{K}}_i, \boldsymbol{\nu}_k\right\rangle\sqrt{\frac{2}{\pi}} - \frac{\eta}{d}\mathbb{E}[\sigma(\mathbf{x}_{k+1}^i)R_{k+1}^i].\end{aligned} \tag{42}$$

If $\left\langle \overline{\mathbf{K}}_i, \boldsymbol{\nu}_k\right\rangle = O(d^{-s})$ for some $s > 0$, then we should be able to approximate $2\sqrt{\frac{2}{\pi}}h_k^i(0)$ by $\frac{1}{\sqrt{2\mathcal{R}_k}}\varphi(\mathcal{R}_k)$ where

$$\varphi(\mathcal{R}_k) = \frac{2}{\pi}\int_{\mathbb{R}} \exp\left(\frac{-y^2}{4\mathcal{R}_k}\right)\mathrm{d}\mathcal{L}_\epsilon(y), \tag{43}$$

for all $1 \leq i \leq d$. As such, assuming the error R_{k+1}^i vanishes, we obtain the approximation of Equation (33):

$$\mathbb{E}\left[\boldsymbol{\nu}_{k+1}^i - \boldsymbol{\nu}_k^i\middle|\mathcal{F}_k\right] \approx -\frac{\eta}{d}\frac{\varphi(\mathcal{R}_k)}{\sqrt{2\mathcal{R}_k}}\left\langle \overline{\mathbf{K}}_i, \boldsymbol{\nu}_k\right\rangle. \tag{44}$$

We will justify (44) in the upcoming lemma. Before we do so, it will be convenient to work under the setting that the risk is bounded. As such, let $L > 0$ and define the following stopping time,

$$\tau_0 = \min\left\{k;\, \|\mathbf{v}_k\| > L\right\}, \tag{45}$$

as well as the stopped process $\mathbf{v}_k = \boldsymbol{\nu}_{k\wedge\tau_0}$. We will show in Lemma 4 that L may be chosen so that $\mathbf{v}_k = \boldsymbol{\nu}_k$ with overwhelming probability.

Next, we consider the following implication of Cauchy's integral formula

$$\left\langle \overline{\mathbf{K}}_i, \mathbf{v}_k\right\rangle = -\frac{1}{2\pi i}\oint_\Gamma z\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k\,\mathrm{d}z. \tag{46}$$

By Assumption 3, we may bound $\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i\right\|$ for all $z \in \Gamma$ by a finite collection of $z_0 \in \Gamma$. Indeed, if Γ_0 is a $1/\sqrt{d}$-net on Γ then $|\Gamma_0| = O(d)$. It follows that for all $z \in \Gamma$, there exists some $z_0 \in \Gamma_0$ such that $|z - z_0| \leq 1/\sqrt{d}$. Then, by resolvent identities we see that for all $1 \leq i \leq d$ and $\mathbf{a} \in \mathbb{R}^d$,

$$\begin{aligned}||\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{a}|| &= ||\mathbf{R}(z_0;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{a} + (z - z_0)[\mathbf{R}(z;\overline{\mathbf{K}})\mathbf{R}(z_0;\overline{\mathbf{K}})]_i^{\mathrm{T}}\mathbf{a}|| \\ &\leq ||\mathbf{R}(z_0;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{a}|| + \frac{1}{\sqrt{d}}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i\right\|\left\|\mathbf{R}(z_0;\overline{\mathbf{K}})\mathbf{a}\right\| \\ &\leq (1 + M_R)\max_{1\leq i\leq d}||\mathbf{R}(z_0;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{a}||.\end{aligned} \tag{47}$$

In particular,

$$\max_{z\in\Gamma}\max_{1\le i\le d}||\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{a}|| \le (1+M_R)\max_{z_0\in\Gamma_0}\max_{1\le i\le d}||\mathbf{R}(z_0;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{a}||. \tag{48}$$

Plugging this into (46),

$$\begin{aligned}|\left\langle\overline{\mathbf{K}}_i,\mathbf{v}_k\right\rangle| &\le \frac{1}{2\pi}\oint_\Gamma \|\mathbf{z}\|\,(1+M_R)\max_{\mathbf{z}_0\in\Gamma_0}\max_{1\le i\le d}\left\|\mathbf{R}(z_0;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k\right\|\,\mathrm{d}z\\ &= 4(1+M_R)\left\|\overline{\mathbf{K}}\right\|^2\max_{z\in\Gamma_0}\max_{1\le i\le d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k\right\|\\ &= C_{\overline{\mathbf{K}}}\max_{z\in\Gamma_0}\max_{1\le i\le d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k\right\|.\end{aligned} \tag{49}$$

Note that terms such as η and $\left\|\overline{\mathbf{K}}\right\|$ are bounded by assumption, thus we make the convention moving forward any constants independent to d such as $C_{\overline{\mathbf{K}}}$ may change from line to line. Therefore, to show that $\left\langle\overline{\mathbf{K}}_i,\mathbf{v}_k\right\rangle$ shrinks as $d\to\infty$, it suffices to show that $\max_{z\in\Gamma_0}\max_{1\le i\le d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k\right\|$ shrinks as $d\to\infty$.

Lemma 1. *Given Assumptions 1 - 5, there exists a net $\Gamma_0\subset\Gamma$ of order $O(d)$, such that for all $t>0$ and $1/6+\delta_0<\delta<1/4$,*

$$\max_{z\in\Gamma_0}\max_{1\le i\le d}\max_{0\le k\le\lfloor td\rfloor}||\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k|| < \frac{d^\delta}{\sqrt{d}} \tag{50}$$

with high probability.

Proof. For clarity of notation, let $\widetilde{h}_k^i = 2\sqrt{\frac{2}{\pi}}h_k^i(0)$ and $\widetilde{h}_k = \frac{1}{\sqrt{2\mathcal{R}_k}}\varphi(\mathcal{R}_k)$. In addition, define A_k be a diagonal matrix with entries given by $\widetilde{h}_k^i$, as well as the vector $E_{k+1} = (\mathbb{E}[\sigma(\mathbf{x}_{k+1}^i)R_{k+1}^i])_{i=1}^d$. By (42), for a fixed $z\in\Gamma_0$ and $1\le i\le d$,

$$\begin{aligned}\mathbb{E}\left[\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}(\mathbf{v}_{k+1}-\mathbf{v}_k)|\mathcal{F}_k\right] &= \mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbb{E}\left[\mathbf{v}_{k+1}-\mathbf{v}_k|\mathcal{F}_k\right]\\ &= -\frac{\eta}{d}\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\left(A_k\overline{\mathbf{K}}\mathbf{v}_k+E_{k+1}\right)\\ &= -\frac{\eta}{d}\widetilde{h}_k\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_k\\ &\quad -\frac{\eta}{d}\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\left(A_k\overline{\mathbf{K}}\mathbf{v}_k-\widetilde{h}_k\overline{\mathbf{K}}\mathbf{v}_k+E_{k+1}\right)\\ &= -\frac{\eta}{d}\widetilde{h}_k\left(z\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k+\mathbf{v}_k^i\right)\\ &\quad -\frac{\eta}{d}\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\left(A_k\overline{\mathbf{K}}\mathbf{v}_k-\widetilde{h}_k\overline{\mathbf{K}}\mathbf{v}_k+E_{k+1}\right)\\ &= -\frac{\eta}{d}\widetilde{h}_k z\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k\\ &\quad +\underbrace{\frac{\eta}{d}\left(-\widetilde{h}_k\mathbf{v}_k^i+\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\left(\widetilde{h}_k\overline{\mathbf{K}}\mathbf{v}_k-A_k\overline{\mathbf{K}}\mathbf{v}_k-E_{k+1}\right)\right)}_{:=\mathcal{E}_k^i(z)}.\end{aligned} \tag{51}$$

By the Doob decomposition we see that,

$$\begin{aligned}\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_{k+1} &= \mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k+\mathbb{E}\left[\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}(\mathbf{v}_{k+1}-\mathbf{v}_k)|\mathcal{F}_k\right]+\Delta M_{k+1}^i(z)\\ &= \left(1-\frac{\eta}{d}\widetilde{h}_k z\right)\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k+\Delta M_{k+1}^i(z),\end{aligned} \tag{52}$$

where $\Delta M_{k+1}^i(z)$ are the martingale increments of $\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\left(\mathbf{v}_{k+1}-\mathbf{v}_k\right)$. Let

$$L_k = \prod_{j=0}^{k}\left(1-\frac{\eta}{d}\widetilde{h}_j z\right),$$

then upon iterating (52) we obtain

$$\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_{k+1} = L_k\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_0 + L_k\sum_{j=0}^{k}\frac{1}{L_j}\left(\mathcal{E}_j^i(z) + \Delta M_{j+1}^i(z)\right). \tag{53}$$

It is easy to check that $\sum_{j=0}^{k}\frac{1}{L_j}\Delta M_{j+1}^i(z)$ is a martingale so we shall denote it by $\overline{\mathcal{M}}_{k+1}^i(z)$. Let

$$\tau_1 = \min\left\{k;\quad ||\mathbf{R}(z,\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k|| \geq \frac{d^\delta}{\sqrt{d}} \text{ for some } 1\leq i\leq d \text{ and } z\in\Gamma_0\right\}. \tag{54}$$

It suffices to show (50) holds for the stopped process $\mathbf{v}_{k\wedge\tau_1}$ given that

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq k\leq\lfloor td\rfloor}\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_k\right\| \geq \frac{d^\delta}{\sqrt{d}}\right)\\ &= \mathbb{P}\left(\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\mathbf{v}_{\lfloor td\rfloor\wedge\tau_1}\right\| \geq \frac{d^\delta}{\sqrt{d}}\right).\end{aligned} \tag{55}$$

For notational clarity, we will write $\widetilde{\mathbf{v}}_k = \mathbf{v}_{k\wedge\tau_1}$. Note that (53) holds all k, so it must also hold for the stopped process $\widetilde{\mathbf{v}}_k$. Given that the entries of $\widetilde{\mathbf{v}}_k$ move at increments of $\frac{\eta}{d}$, we observe the following bound on the stopped process,

$$\begin{aligned}||\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_k|| &\leq \frac{d^\delta}{\sqrt{d}} + \frac{\eta M_R}{\sqrt{d}}\\ &\leq \frac{\eta M_R' d^\delta}{\sqrt{d}}.\end{aligned} \tag{56}$$

Moreover, by Lemma 8 we know that $\widetilde{h}_k \leq M_\epsilon$. This in turn implies L_k is bounded from above and below. Indeed,

$$\begin{aligned}\|L_k\| &= \prod_{j=0}^{k}\left\|1 - \frac{\eta}{d}\widetilde{h}_k z\right\|\\ &\leq \prod_{j=0}^{k} 1 + \frac{\eta M_\epsilon\|z\|}{d}\\ &\leq \left(1 + \frac{2\eta M_\epsilon\left\|\overline{\mathbf{K}}\right\|}{d}\right)^{\lfloor td\rfloor}\\ &\leq \exp\left(\eta C_{t,\epsilon,\overline{\mathbf{K}}}\right).\end{aligned} \tag{57}$$

Similarly for the lower bound,

$$\begin{aligned}\|L_k\| &\geq \prod_{j=0}^{k} 1 - \frac{\eta}{d}\widetilde{h}_k\|z\|\\ &\geq \left(1 - \frac{2\eta M_\epsilon\left\|\overline{\mathbf{K}}\right\|}{d}\right)^{\lfloor td\rfloor}\\ &\geq \exp\left(\frac{-\frac{2\eta M_\epsilon\left\|\overline{\mathbf{K}}\right\|}{d}\lfloor td\rfloor}{1 - \frac{2\eta M_\epsilon\left\|\overline{\mathbf{K}}\right\|}{d}}\right)\\ &\geq \exp\left(-\frac{2\eta M_\epsilon\left\|\overline{\mathbf{K}}\right\|}{d}\lfloor td\rfloor\right)\\ &= \exp\left(-\eta C_{t,\epsilon,\overline{\mathbf{K}}}\right),\end{aligned} \tag{58}$$

provided that $\frac{\eta M_\epsilon \|\overline{\mathbf{K}}\|}{d} < \frac{1}{2}$. Therefore, up to a constant factor

$$\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_k\right\| \leq C_{\eta,t,\epsilon,\overline{\mathbf{K}}}\left(\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_0\right\| + \left\|\overline{\mathcal{M}}_{k+1}^i(z)\right\| + \sum_{j=0}^{k-1}\left\|\mathcal{E}_j^i(z)\right\|\right). \tag{59}$$

We will now bound the error $\mathcal{E}_j^i(z)$. By (49), we already know that

$$|\left\langle\overline{\mathbf{K}}_i,\widetilde{\mathbf{v}}_j\right\rangle| \leq C_{\overline{\mathbf{K}}}\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_j\right\|. \tag{60}$$

Similarly,

$$|\widetilde{\mathbf{v}}_j^i| \leq C_{\overline{\mathbf{K}}}\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_j\right\|. \tag{61}$$

We also observe for all $1 \leq i \leq d$,

$$\begin{aligned}
E_{j+1}^i &= \mathbb{E}\left[\sigma(\mathbf{x}_{j+1}^i)R_{j+1}^i\right] \\
&\leq O\left(\mathbb{E}\left[\left|\frac{\left\langle\mathbf{K}_i,\widetilde{\mathbf{v}}_j\right\rangle\mathbf{x}_{j+1}^i}{\mathbf{K}_{ii}}\right|^3\right]\right) \\
&= O\left(|\left\langle\overline{\mathbf{K}}_i,\widetilde{\mathbf{v}}_j\right\rangle|^3\right) \\
&= O\left(\left(\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_j\right\|\right)^3\right).
\end{aligned} \tag{62}$$

In particular, for some constant $C_{\overline{\mathbf{K}}} > 0$,

$$\begin{aligned}
\|E_{j+1}\| &= \sqrt{\sum_{i=1}^d\left(E_{j+1}^i\right)^2} \\
&\leq \sqrt{d}C_{\overline{\mathbf{K}}}\left(\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_j\right\|\right)^3.
\end{aligned} \tag{63}$$

For our last error term we apply the Lipschitz bound obtained by Lemma 8. That is the map

$$s \mapsto \psi(s) = \frac{2}{\pi\sqrt{s}}\int_{-\infty}^{\infty}\exp\left(\frac{-y^2}{2s}\right)\mathrm{d}\mu_\epsilon(y), \tag{64}$$

is Lipschitz with constant L_ϵ. Moreover, $\psi(2\mathcal{R}_j - \left\langle\overline{\mathbf{K}}_i,\widetilde{\mathbf{v}}_j\right\rangle^2) = \widetilde{h}_j^i$ and $\psi(2\mathcal{R}_j) = \widetilde{h}_j$. By (60), for all $1 \leq i \leq d$,

$$\begin{aligned}
|\widetilde{h}_j^i - \widetilde{h}_j| &\leq L_\epsilon\left\langle\overline{\mathbf{K}}_i,\widetilde{\mathbf{v}}_j\right\rangle^2 \\
&\leq L_\epsilon\left(C_{\overline{\mathbf{K}}}\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_j\right\|\right)^2.
\end{aligned} \tag{65}$$

It follows that

$$\begin{aligned}
\left\|\widetilde{h}_j\overline{\mathbf{K}}\widetilde{\mathbf{v}}_j - A_j\overline{\mathbf{K}}\widetilde{\mathbf{v}}_j\right\| &\leq \left\|\widetilde{h}_jI_d - A_j\right\|\left\|\overline{\mathbf{K}}\widetilde{\mathbf{v}}_j\right\| \\
&\leq L_\epsilon\left(C_{\overline{\mathbf{K}}}\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_j\right\|\right)^2\left\|\overline{\mathbf{K}}\widetilde{\mathbf{v}}_j\right\| \\
&\leq C_{\epsilon,\overline{\mathbf{K}}}\sqrt{d}\left(\max_{z\in\Gamma_0}\max_{1\leq i\leq d}\left\|\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_j\right\|\right)^3.
\end{aligned} \tag{66}$$

For notational clarity, let us write $\omega_k = \max_{z\in\Gamma_0}\max_{1\leq i\leq d}||\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}}\widetilde{\mathbf{v}}_k||$. Putting all this together we have up to constant factor,

$$\begin{aligned}
\left\|\mathcal{E}_j^i(z)\right\| &\leq \frac{\eta}{d}\left(|\widetilde{h}_j\widetilde{\mathbf{v}}_j^i| + \left\|\mathbf{R}(z;\overline{\mathbf{K}})\right\|\left(\left\|\widetilde{h}_jI_d - A_j\right\|\left\|\overline{\mathbf{K}}\widetilde{\mathbf{v}}_j\right\| + \|E_{j+1}\|\right)\right) \\
&\leq \frac{\eta C_{\epsilon,\overline{\mathbf{K}}}}{d}\left(\omega_j + 2\sqrt{d}\omega_j^3\right).
\end{aligned} \tag{67}$$

Returning to (59), upon taking the max across $z \in \Gamma_0$ and $1 \leq i \leq d$ and up to a constant $C_{\eta,t,\epsilon,\overline{\mathbf{K}}} > 0$, we obtain for all $k \leq \lfloor td \rfloor$

$$\omega_k \leq C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \left(\omega_0 + \max_{z\in\Gamma_0} \max_{1\leq i\leq d} \max_{1\leq k\leq \lfloor td \rfloor} \overline{\mathcal{M}}_k^i(z) + \sum_{j=0}^{k-1} \frac{\eta}{d} \left(\omega_j + 2\sqrt{d}\omega_j^3 \right) \right). \tag{68}$$

Define

$$\beta_t = C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \left(\omega_0 + \max_{z\in\Gamma_0} \max_{1\leq i\leq d} \max_{1\leq k\leq \lfloor td \rfloor} \overline{\mathcal{M}}_k^i(z) \right), \tag{69}$$

as well as the stopping time

$$\tau_2 = \min \left\{ k \,; \omega_k \geq 3\beta_t \exp\left(C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \right) \right\}. \tag{70}$$

As before, we note that ω_k can only move at increments of at-most $\frac{\eta M_r}{\sqrt{d}}$. Thus,

$$\omega_{k\wedge\tau_2} \leq 3\beta_t \exp(C_{\eta,t,\epsilon,\overline{\mathbf{K}}}) + \frac{\eta M_r}{\sqrt{d}} =: \beta_t', \tag{71}$$

for all $k \in \mathbb{N}$. Plugging this into (68),

$$\begin{aligned} \omega_{\lfloor td \rfloor\wedge\tau_2} &\leq \beta_t + C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \left(\sum_{j=0}^{(\lfloor td \rfloor - 1)\wedge\tau_2} \frac{\eta}{d}\omega_j + \sum_{j=0}^{\lfloor td \rfloor} \frac{\eta}{d} \left(2\sqrt{d}\omega_j^3 \right) \right) \\ &\leq \beta_t + C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \left[2\sqrt{d}(\beta_t')^3 \right] + \sum_{j=0}^{(\lfloor td \rfloor - 1)\wedge\tau_2} C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \frac{\eta}{d}\omega_j. \end{aligned} \tag{72}$$

By Gronwall's inequality,

$$\omega_{\lfloor td \rfloor\wedge\tau_2} \leq \left(\beta_t + C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \sqrt{d} \left[2(\beta_t')^3 \right] \right) \exp\left(C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \right). \tag{73}$$

If we can show that β_t can be made sufficiently small so that

$$C_{\eta,t,\epsilon,\overline{\mathbf{K}}} \left[\sqrt{d}(\beta_t')^3 \right] \leq \beta_t, \tag{74}$$

then $\omega_{\lfloor td \rfloor\wedge\tau_2} = \omega_{\lfloor td \rfloor}$. To see this, recall by Assumption 5 we know that for any constant $\xi > 0$, the former term of β_t has the following tail bound,

$$\mathbb{P}\left(\max_{z\in\Gamma_0} \max_{1\leq i\leq d} ||\mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}} \mathbf{v}_0|| \geq \frac{\xi d^\delta}{\sqrt{d}} \right) \leq Cd^2 \exp\left(-c'\xi^2 d^{2\delta} \right). \tag{75}$$

To bound the martingale term, we first fix $z \in \Gamma_0$ and $1 \leq i \leq d$, then let

$$\tau_3 = \min \left\{ k \,; \, |\overline{\mathcal{M}}_k^i(z)| \geq \frac{\xi d^\delta}{\sqrt{d}} \right\}. \tag{76}$$

Let $X_k = \overline{\mathcal{M}}_{k\wedge\tau_3}^i(z)$. Notice that $\mathbb{E}\left[\overline{\mathcal{M}}_k^i(z) \right] = 0$, so $\mathbb{E}\left[X_k \right] = 0$. It follows that

$$\mathbb{P}\left(\max_{1\leq k\leq \lfloor td \rfloor} |\overline{\mathcal{M}}_k^i(z)| \geq \frac{\xi d^\delta}{\sqrt{d}} \right) = \mathbb{P}\left(\left\| X_{\lfloor td \rfloor} \right\| \geq \frac{\xi d^\delta}{\sqrt{d}} \right). \tag{77}$$

Notice that

$$\begin{aligned} \left\| \overline{\mathcal{M}}_{k+1}^i(z) - \overline{\mathcal{M}}_k^i(z) \right\| &= \frac{1}{\|L_k\|} \left\| \mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}} (\mathbf{v}_{k+1} - \mathbf{v}_k) - \mathbf{R}(z;\overline{\mathbf{K}})_i^{\mathrm{T}} \mathbb{E}\left[\mathbf{v}_{k+1} - \mathbf{v}_k \middle| \mathcal{F}_k \right] \right\| \\ &\leq \frac{C_{t,\overline{\mathbf{K}}} \eta M_R}{\sqrt{d}}. \end{aligned} \tag{78}$$

Hence, $\|X_k - X_{k-1}\| \leq \frac{C_{t,\overline{\mathbf{K}}}\eta M_R}{\sqrt{d}}$ almost surely for all k. However, we may improve this increment bound by $\frac{d^s}{d}$ for $\frac{1}{6} + \delta_0 < s < \delta$. Indeed, by Corollary 2 for all even moments $2p < d$, there exists a constant $C(2p, \eta, \mathbf{K})$ such that

$$\mathbb{P}\left(\|X_{k+1} - X_k\| \geq \frac{d^s}{d}\right) \leq C(2p, \eta, \mathbf{K}) d^{p\left(\frac{1}{3} - 2s + 2\delta_0\right)}. \tag{79}$$

It follows by Lemma 11,

$$\mathbb{P}\left(\left\|X_{\lfloor td \rfloor}\right\| \geq \frac{\xi d^\delta}{\sqrt{d}}\right) \leq 2\exp\left(-\frac{\xi^2 d^{2(\delta - s)}}{C_{\eta,t,\overline{\mathbf{K}}}}\right) + \lfloor td \rfloor C(2p, \eta, \mathbf{K}) d^{p\left(\frac{1}{3} - 2s + 2\delta_0\right)} \tag{80}$$

Thus, taking union bounds across $z \in \Gamma_0$ and $1 \leq i \leq d$ gives,

$$\mathbb{P}\left(\max_{z \in \Gamma_0} \max_{1 \leq i \leq d} \max_{1 \leq k \leq \lfloor td \rfloor + 1} \left\|\overline{\mathcal{M}}_k^i(z)\right\| \geq \frac{\xi d^\delta}{\sqrt{d}}\right) \leq C d^2 \exp\left(-\frac{\xi^2 d^{2(\delta - s)}}{C_{\eta,t,\overline{\mathbf{K}}}}\right) + 2Cd^2 \lfloor td \rfloor C(2p, \eta, \mathbf{K}) d^{p\left(\frac{1}{3} - 2s + 2\delta_0\right)} \tag{81}$$

It is easy to see that for d sufficiently large, we may choose p large so that $s > \frac{1}{6} + \delta_0 + \frac{3}{2p}$, implying the latter term converges to 0 as $d \to \infty$. Therefore, $\beta_t \leq \frac{\xi d^\delta}{\sqrt{d}}$ with high probability. Returning to (74), up to a constant factor that is independent to d,

$$\begin{aligned}
\mathbb{P}\left(\sqrt{d}\beta_t^3 > \beta_t\right) &= \mathbb{P}\left(\beta_t^2 > \frac{1}{\sqrt{d}}\right) \\
&= \mathbb{P}\left(\frac{d^{2\delta}}{d} > \beta_t^2 > \frac{1}{\sqrt{d}}\right) + \mathbb{P}\left(\beta_t^2 > \frac{1}{\sqrt{d}}, \beta_t^2 \geq \frac{d^{2\delta}}{d}\right) \\
&= \mathbb{P}\left(\beta_t \geq \frac{d^\delta}{\sqrt{d}}\right),
\end{aligned} \tag{82}$$

provided that $\delta < 1/4$. Thus, (74) is satisfied and $\omega_{\lfloor td \rfloor \wedge \tau_2} = \omega_{\lfloor td \rfloor}$ with high-probability. By choosing ξ appropriately in accordance to (70), we conclude $\omega_{\lfloor td \rfloor} \leq \frac{d^\delta}{\sqrt{d}}$ with high-probability. □

A.2 Convergence of signSGD to signHSGD

In this section we will show convergence of the dynamics of signSGD to that of signHSGD. Recall signHSGD is defined as in (7). Similarly to signSGD we will impose a stopping time onto signHSGD,

$$\tau_0' = \min_{t > 0} \{t;\ \|\mathbf{\Theta}_t - \boldsymbol{\theta}_*\| > L\}. \tag{83}$$

We will also define the stopped process by $\mathbf{V}_t = \mathbf{\Theta}_{t \wedge \tau_0'} - \boldsymbol{\theta}_*$. In light of Lemma 1 we will equate

$$\mathbb{E}\left[\mathbf{v}_{k+1} - \mathbf{v}_k \middle| \mathcal{F}_k\right] = -\frac{\eta \varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}} \overline{\mathbf{K}} \mathbf{v}_k. \tag{84}$$

We shall use the following:

Definition 5 (Quadratic)**.** *A function $q : \mathbb{R}^d \to \mathbb{R}$ is quadratic if it may be written in the form*

$$q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x} + \mathbf{b}^{\mathrm{T}} \mathbf{x} + c$$

for some $\mathbf{A} \in \mathbb{R}^{d \times d}$, $\mathbf{b} \in \mathbb{R}^d$, and $c \in \mathbb{R}$.

Now if $q : \mathbb{R}^d \to \mathbb{R}$ is quadratic, it is easy to see that

$$q(\mathbf{v}_{k+1}) - q(\mathbf{v}_k) = -\frac{\eta}{d} \nabla q(\mathbf{v}_k)^{\mathrm{T}} (\sigma_{k+1}) + \frac{\eta^2}{2d^2} (\sigma_{k+1})^{\mathrm{T}} \nabla^2 q(\mathbf{v}_k) (\sigma_{k+1}). \tag{85}$$

Thus, taking its conditional expectation we obtain

$$\mathbb{E}[q(\mathbf{v}_{k+1}) - q(\mathbf{v}_k)|\mathcal{F}_k] = -\frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}\nabla q(\mathbf{v}_k)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_k + \frac{\eta^2}{d^2\pi}\left\langle \nabla^2 q(\mathbf{v}_k), \mathbf{K}_\sigma \right\rangle. \tag{86}$$

By the Doob-decomposition, we have

$$q(\mathbf{v}_{k+1}) - q(\mathbf{v}_k) = -\frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}\nabla q(\mathbf{v}_k)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_k + \frac{\eta^2}{d^2\pi}\left\langle \nabla^2 q(\mathbf{v}_k), \mathbf{K}_\sigma \right\rangle + \Delta\mathcal{M}_{k+1}^{lin} + \Delta\mathcal{M}_{k+1}^{quad}, \tag{87}$$

where

$$\Delta\mathcal{M}_{k+1}^{lin} = -\frac{\eta}{d}\nabla q(\mathbf{v}_k)^{\mathrm{T}}\left(\sigma_{k+1} - \mathbb{E}[\sigma_{k+1}|\mathcal{F}_k]\right), \tag{88}$$

and

$$\Delta\mathcal{M}_{k+1}^{quad} = \frac{\eta^2}{2d^2}\left(\sigma_{k+1}^{\mathrm{T}}\nabla^2 q(\mathbf{v}_k)\sigma_{k+1} - \mathbb{E}[\sigma_{k+1}^{\mathrm{T}}\nabla^2 q(\mathbf{v}_k)\sigma_{k+1}|\mathcal{F}_k]\right). \tag{89}$$

Similarly, by Ito's lemma on $\mathbf{V}_t$, we see that

$$\mathrm{d}q(\mathbf{V}_t) = \left(-\frac{\eta\varphi(\mathcal{R}(\mathbf{V}_t))}{\sqrt{2\mathcal{R}(\mathbf{V}_t)}}\nabla q(\mathbf{V}_t)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{V}_t + \frac{\eta^2}{\pi d}\left\langle \nabla^2 q(V_t), \mathbf{K}_\sigma \right\rangle\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_t^\sigma, \tag{90}$$

where

$$\mathrm{d}\mathcal{M}_t^\sigma = \eta\nabla q(\mathbf{V}_t)^{\mathrm{T}}\left(\sqrt{\frac{2\mathbf{K}_\sigma}{d\pi}}\,\mathrm{d}\mathbf{B}_t\right). \tag{91}$$

Comparing (87) and (90), we see that predictable part of signSGD and the total variation part of HSGD depend only on $\nabla q(x)^{\mathrm{T}}\mathbf{K}x$ and $\mathcal{R}(x)$. We capture these statistics in a "closed" manifold defined by

$$Q_q = \left\{\mathbf{x}^{\mathrm{T}}\mathbf{x},\ q(\mathbf{x}),\ \nabla q(\mathbf{x})^{\mathrm{T}}\mathbf{R}(z;\overline{\mathbf{K}})\mathbf{x},\ \mathbf{x}^{\mathrm{T}}\mathbf{R}(z;\overline{\mathbf{K}})^{\mathrm{T}}\nabla^2 q(\mathbf{x})\mathbf{R}(y;\overline{\mathbf{K}})\mathbf{x};\ z, y \in \Gamma\right\}. \tag{92}$$

To be precise in our notion of closure, given any $g \in Q_q$, the predictable part of (87) and the total variation part of (90) may be expressed via contour integral around Γ by a linear combination of functions from Q_q. Let us look at an example. Suppose $g(x) = \nabla q(x)^{\mathrm{T}}\mathbf{R}(z;\overline{\mathbf{K}})x$. It is easy to see that

$$\begin{aligned}
\mathbb{E}\left[\nabla g(\mathbf{v}_k)^{\mathrm{T}}(\mathbf{v}_{k+1} - \mathbf{v}_k)|\mathcal{F}_k\right] &= -\frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}\left(\mathbf{v}_k^{\mathrm{T}}\nabla^2 q\mathbf{R}(z;\overline{\mathbf{K}})\overline{\mathbf{K}}\mathbf{v}_k + \mathbf{v}_k^{\mathrm{T}}\mathbf{R}(z;\overline{\mathbf{K}})^{\mathrm{T}}\nabla^2 q\overline{\mathbf{K}}\mathbf{v}_k\right) \\
&= -\frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}\left(\mathbf{v}_k^{\mathrm{T}}\nabla^2 q\mathbf{v}_k\right) \\
&\quad - \frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}\underbrace{\left(z\mathbf{v}_k^{\mathrm{T}}\nabla^2 q\mathbf{R}(z;\overline{\mathbf{K}})\mathbf{v}_k + \mathbf{v}_k^{\mathrm{T}}\mathbf{R}(z;\overline{\mathbf{K}})^{\mathrm{T}}\nabla^2 q\overline{\mathbf{K}}\mathbf{v}_k\right)}_{p(\mathbf{v}_k)}.
\end{aligned} \tag{93}$$

However, by Cauchy's integral theorem we may express $p(\mathbf{v}_k)$ by

$$p(\mathbf{v}_k) = -\frac{1}{2\pi i}\oint_\Gamma z\mathbf{v}_k^{\mathrm{T}}\mathbf{R}(y;\overline{\mathbf{K}})^{\mathrm{T}}\nabla^2 q\mathbf{R}(z;\overline{\mathbf{K}})\mathbf{v}_k + y\mathbf{v}_k^{\mathrm{T}}\mathbf{R}(z;\overline{\mathbf{K}})^{\mathrm{T}}\nabla^2 q\mathbf{R}(y;\overline{\mathbf{K}})\mathbf{v}_k\,\mathrm{d}y, \tag{94}$$

as well as

$$\mathbf{v}_k^{\mathrm{T}}\nabla^2 q\mathbf{v}_k = \frac{1}{4\pi^2}\oint_\Gamma\oint_\Gamma \mathbf{v}_k^{\mathrm{T}}\mathbf{R}(z;\overline{\mathbf{K}})\nabla^2 q\mathbf{R}(y;\overline{\mathbf{K}})\mathbf{v}_k\,\mathrm{d}x\mathrm{d}y. \tag{95}$$

Consequently, we see that

$$\begin{aligned}
\left|\mathbb{E}\left[\nabla g(\mathbf{v}_k)^{\mathrm{T}}(\mathbf{v}_{k+1} - \mathbf{v}_k)|\mathcal{F}_k\right]\right| &\le \frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}16\left\|\overline{\mathbf{K}}\right\|^2\max_{g\in Q_q}|g(\mathbf{v}_k)| &(96)\\
&\le \frac{16\eta M_\epsilon}{d}\left\|\overline{\mathbf{K}}\right\|^2\max_{g\in Q_q}|g(\mathbf{v}_k)|, &(97)
\end{aligned}$$

where we applied Lemma 8 in the second inequality. Note the constant factor of $16\left\|\overline{\mathbf{K}}\right\|^2$ depended on g. We may work around this choice dependent constant to obtain a uniform bound on (97) for all g with the following lemma:

Lemma 2. *Let Q_q be defined as above then for all $n > 0$ there exists $\overline{Q}_q \subset Q_q$ such that $|\overline{Q}_q| \leq C(\overline{\mathbf{K}})d^{4n}$ and for all $g \in Q_q$, there exists $g_0 \in \overline{Q}_q$ satisfying $\|g - g_0\|_{C^2} \leq d^{-2n}$.*

The proof of Lemma 2 may be found in Collins-Woodfin et al. (2024). If $g \in Q$ then there exists $g_0 \in \overline{Q}_q$ such that

$$\left|\mathbb{E}\left[\nabla g(\mathbf{v}_k)^{\mathrm{T}}(\mathbf{v}_{k+1} - \mathbf{v}_k)|\mathcal{F}_k\right]\right| = \left|\frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}\nabla g(\mathbf{v}_k)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_k\right| \tag{98}$$

$$\leq \frac{\eta\varphi(\mathcal{R}_k)}{d\sqrt{2\mathcal{R}_k}}\left(\left|\nabla(g - g_0)(\mathbf{v}_k)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_k\right| + \left|\nabla g_0(\mathbf{v}_k)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_k\right|\right) \tag{99}$$

$$\leq \frac{\eta M_\epsilon}{d}\left(\|g - g_0\|_{C^2}\left\|\overline{\mathbf{K}}\right\|\|\mathbf{v}_k\|^2 + C_{g_0}(\overline{\mathbf{K}})\max_{g\in Q_q}|g(\mathbf{v}_k)|\right) \tag{100}$$

$$\leq \frac{\eta M_\epsilon}{d}\left(d^{-2n}\left\|\overline{\mathbf{K}}\right\| + C_{g_0}(\overline{\mathbf{K}})\right)\max_{g\in Q_q}|g(\mathbf{v}_k)|, \tag{101}$$

where $C_{g_0}(\overline{\mathbf{K}})$ is the choice dependent constant as in (97). By taking the max across our finite net $\overline{Q}_q$, there exists $C(\overline{\mathbf{K}}) > 0$ such that for all $g \in Q_q$,

$$\left|\mathbb{E}\left[\nabla g(\mathbf{v}_k)^{\mathrm{T}}(\mathbf{v}_{k+1} - \mathbf{v}_k)|\mathcal{F}_k\right]\right| \leq \frac{\eta M_\epsilon}{d}C(\overline{\mathbf{K}})\max_{g\in Q_q}|g(\mathbf{v}_k)|. \tag{102}$$

We are now ready to prove our main result. It would be convenient to extend the indexing of $\mathbf{v}_k$ from $\mathbb{N}$ to $\mathbb{R}$ by defining the sequence $t_k = k/d$. With some slight abuse of notation, let $\mathbf{v}_{t_k} = \mathbf{v}_k$. If $t_{k-1} \leq t < t_k$, then define $\mathbf{v}_t = \mathbf{v}_{t_{k-1}}$.

Lemma 3. *Given $0 < 2p < d$ and a quadratic q such that $||q||_{C^2} \leq 1$, define $Q = Q_q \cup Q_{\mathcal{R}}$, where $\mathcal{R}$ is the risk. For all $T > 0$ and $1/3 < \delta < 1/2$, there exists $C(\overline{\mathbf{K}}, \epsilon) > 0$ such that*

$$\sup_{0\leq t\leq T}|q(\mathbf{v}_t) - q(\mathbf{V}_t)| \leq \frac{3d^\delta}{\sqrt{d}}\exp\left(C(\overline{\mathbf{K}}, \epsilon)\|\eta\|_\infty T\right), \tag{103}$$

with probability at least $1 - c(2p, \overline{\mathbf{K}})d^{p(1/3-\delta)}$.

Proof. Let $g \in Q$, by (87), we see that

$$\begin{aligned} g(\mathbf{v}_t) &= g(\mathbf{v}_0) - \frac{\eta}{d}\sum_{i=0}^{\lfloor td\rfloor}\frac{\varphi(\mathcal{R}(\mathbf{v}_i))}{\sqrt{2\mathcal{R}(\mathbf{v}_i)}}\nabla g(\mathbf{v}_i)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_i + \frac{\eta^2}{d^2\pi}\sum_{i=0}^{\lfloor td\rfloor}\left\langle\nabla^2 g(\mathbf{v}_k), \mathbf{K}_\sigma\right\rangle + \mathcal{M}_t^{lin} + \mathcal{M}_t^{quad} \\ &= g(\mathbf{v}_0) - \eta\int_0^t\frac{\varphi(\mathcal{R}(\mathbf{v}_s))}{\sqrt{2\mathcal{R}(\mathbf{v}_s)}}\nabla g(\mathbf{v}_s)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_s\,\mathrm{d}s + \frac{\eta^2}{d\pi}\int_0^t\left\langle\nabla^2 g(\mathbf{v}_s), \mathbf{K}_\sigma\right\rangle\mathrm{d}s \\ &\quad + \mathcal{M}_t^{lin} + \mathcal{M}_t^{quad}. \end{aligned} \tag{104}$$

Taking the difference with σ-HSGD, we see that

$$\begin{aligned} |g(\mathbf{v}_t) - g(\mathbf{V}_t)| \leq \eta\int_0^t\left|\frac{\varphi(\mathcal{R}(\mathbf{v}_s))}{\sqrt{2\mathcal{R}(\mathbf{v}_s)}}\nabla g(\mathbf{v}_s)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{v}_s - \frac{\varphi(\mathcal{R}(\mathbf{V}_s))}{\sqrt{2\mathcal{R}(\mathbf{V}_s)}}\nabla g(\mathbf{V}_s)^{\mathrm{T}}\overline{\mathbf{K}}\mathbf{V}_s\right|\mathrm{d}s \\ + \sup_{0\leq s\leq t}\left(|\mathcal{M}_s^{lin}| + |\mathcal{M}_s^{quad}| + |\mathcal{M}_s^{\sigma}|\right). \end{aligned} \tag{105}$$

However, Lemma 8 tells us the map

$$(a, b) \mapsto \frac{\varphi(a)}{\sqrt{2a}}b, \tag{106}$$

is Lipschitz continuous with constant $L_\epsilon > 0$. Thus, using the same argument as in (102) we may bound the integrand by

$$\begin{aligned}
&\left| \frac{\varphi(\mathcal{R}(\mathbf{v}_s))}{\sqrt{2\mathcal{R}(\mathbf{v}_s)}} \nabla g(\mathbf{v}_s)^{\mathrm{T}} \overline{\mathbf{K}} \mathbf{v}_s - \frac{\varphi(\mathcal{R}(\mathbf{V}_s))}{\sqrt{2\mathcal{R}(\mathbf{V}_s)}} \nabla g(\mathbf{V}_s)^{\mathrm{T}} \overline{\mathbf{K}} \mathbf{V}_s \right| \\
&\leq L_\epsilon \sqrt{\left(\nabla g(\mathbf{v}_s)^{\mathrm{T}} \overline{\mathbf{K}} \mathbf{v}_s - \nabla g(\mathbf{V}_s)^{\mathrm{T}} \overline{\mathbf{K}} \mathbf{V}_s\right)^2 + \left(\mathcal{R}(\mathbf{v}_s) - \mathcal{R}(\mathbf{V}_s)\right)^2} \\
&\leq L_\epsilon C(\overline{\mathbf{K}}) \max_{g \in Q} |g(\mathbf{v}_s) - g(\mathbf{V}_s)|. \qquad (107)
\end{aligned}$$

Plugging into (105) we get

$$\begin{aligned}
\sup_{g \in Q} |g(\mathbf{v}_t) - g(V_t)| \leq \sup_{0 \leq s \leq t} &\left(|\mathcal{M}_s^{lin}| + |\mathcal{M}_s^{quad}| + |\mathcal{M}_s^{\sigma}|\right) \\
&+ \eta L_\epsilon C(\overline{\mathbf{K}}) \int_0^t \max_{g \in Q} |g(\mathbf{v}_s) - g(\mathbf{V}_s)| \, \mathrm{d}s. \qquad (108)
\end{aligned}$$

By Gronwall's inequality,

$$\sup_{g \in Q} |g(\mathbf{v}_t) - g(\mathbf{V}_t)| \leq \sup_{0 \leq s \leq t} \left(|\mathcal{M}_s^{lin}| + |\mathcal{M}_s^{quad}| + |\mathcal{M}_s^{\sigma}|\right) \exp\left(\eta L_\epsilon C(\overline{\mathbf{K}}) t\right). \qquad (109)$$

Lemmas 5, 6 and 7 bound the martingales, concluding the proof. □

We have now shown that the stopped processes satisfy the conclusion of Theorem 1. We will conclude the proof of Theorem 1 by showing that, with high-probability, the process is not stopped.

Lemma 4. *For all $T > 0$, there exists $C(\overline{\mathbf{K}}, \mathbf{K}_\sigma) > 0$ such that*

$$\max_{0 \leq t \leq T} \|\mathbf{V}_t\| \leq \exp\left(T C(\overline{\mathbf{K}}, \mathbf{K}_\sigma)\right), \qquad (110)$$

with overwhelming probability.

Proof. For $\mathbf{z} \in \mathbb{R}^d$ let $\psi(\mathbf{z}) = \log(1 + \|\mathbf{z}\|^2)$. By Itô's lemma,

$$\begin{aligned}
\mathrm{d}\psi(\mathbf{V}_t) = &\left[\frac{-2\eta\varphi(\mathcal{R}_t)}{\sqrt{2\mathcal{R}_t}(1 + \|\mathbf{V}_t\|^2)} \mathbf{V}_t^{\mathrm{T}} \overline{\mathbf{K}} \mathbf{V}_t - \frac{\eta^2}{d\pi(1 + \|\mathbf{V}_t\|^2)} \mathbf{V}_t^{\mathrm{T}} \mathbf{K}_\sigma \mathbf{V}_t \right] \mathrm{d}t \\
&+ \left[\frac{2\eta^2}{d\pi(1 + \|\mathbf{V}_t\|^2)} \operatorname{Tr}(\mathbf{K}_\sigma) \right] \mathrm{d}t + \frac{2\eta}{(1 + \|\mathbf{V}_t\|^2)} \mathbf{V}_t^{\mathrm{T}} \sqrt{\frac{\mathbf{K}_\sigma}{d\pi}} \, \mathrm{d}\mathbf{B}_t. \qquad (111)
\end{aligned}$$

It is easy to check by the Cauchy-Schwarz inequality that the deterministic terms of may be uniformly bounded by some constant $C(\overline{\mathbf{K}}, \mathbf{K}_\sigma) > 0$. Denote the martingale term by $\mathcal{M}_t^{\sigma-HSGD}$ then the quadratic variation is given by,

$$\begin{aligned}
\langle \mathcal{M}^{\sigma-HSGD} \rangle_t &= \frac{4\eta^2}{d\pi(1 + \|\mathbf{V}_t\|^2)^2} \int_0^t \mathbf{V}_s^{\mathrm{T}} \mathbf{K}_\sigma \mathbf{V}_s \, \mathrm{d}s \qquad (112) \\
&\leq \frac{\eta^2 \|\mathbf{K}_\sigma\| t}{d\pi}. \qquad (113)
\end{aligned}$$

By subgaussian concentration,

$$\begin{aligned}
\mathbb{P}\left(\max_{0 \leq t \leq T} \psi(\mathbf{V}_t) \geq 2TC(\overline{\mathbf{K}}, \mathbf{K}_\sigma) \right) &\leq \mathbb{P}\left(\max_{0 \leq t \leq T} \mathcal{M}_t^{\sigma-HSGD} \geq TC(\overline{\mathbf{K}}, \mathbf{K}_\sigma) \right) \\
&\leq 2 \exp\left(\frac{-C(\overline{\mathbf{K}}, \mathbf{K}_\sigma)^2 T d\pi}{2\eta^2 \|\mathbf{K}_\sigma\|} \right). \qquad (114)
\end{aligned}$$

That is

$$\max_{0 \leq t \leq T} \|\mathbf{V}_t\| \leq \exp\left(TC(\overline{\mathbf{K}}, \mathbf{K}_\sigma)\right), \qquad (115)$$

with overwhelming probability. □

Therefore by choosing the upper bound in our stopping τ_0 and τ_0' in accordance to Lemma 4, we obtain $\mathbf{v}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_*$ and $\mathbf{V}_t = \boldsymbol{\Theta}_t - \boldsymbol{\theta}_*$ for all $0 \le t \le T$ with overwhelming probability. Combining this with Lemma 3 proves Theorem 1 as well as the following generalization:

Theorem 5. *Given Assumptions 1–5 and a quadratic $q : \mathbb{R}^d \to \mathbb{R}$, if $g(\mathbf{x}) = q(\mathbf{x} - \boldsymbol{\theta}_*)$ then choosing any fixed even moment $2p \in (0, d)$, there exists a constant $C(\overline{\mathbf{K}}, \epsilon) > 0$ such that for any $\delta \in (1/3, 1/2)$ and all $T > 0$,*

$$\sup_{0 \le t \le T} |g(\boldsymbol{\theta}_{\lfloor td \rfloor}) - g(\boldsymbol{\Theta}_t)| \le \frac{3d^\delta \|g\|_{C^2}}{\sqrt{d}} \exp\left(C(\overline{\mathbf{K}}, \epsilon) \|\eta\|_\infty T\right), \tag{116}$$

with probability at least $1 - c(2p, \overline{\mathbf{K}}) d^{p(1/3-\delta)}$ for a constant $c(2p, \overline{\mathbf{K}})$ independent to d.

A.3 Main theorem with badly behaved noise

In this section we formulate a version of Theorem 5 without Assumption 2. The key is that we must work on subsets of the state space where the risk remains away from 0. So suppose that we let

$$\vartheta := \min_{t>0} \{t;\ \|\boldsymbol{\Theta}_t - \boldsymbol{\theta}_*\| < \varrho\},$$

for a fixed positive $\varrho > 0$.

We note that the map $x \mapsto \varphi(x)$ is Lipschitz on $[\varrho, \infty)$, even without Assumption 2, since

$$\varphi'(s) = \frac{1}{s} \int_{\mathbb{R}} \frac{y^2}{2s} \exp\left(-\frac{y^2}{2s}\right) \mu(\mathrm{d}y).$$

The function xe^{-x} is uniformly bounded on $x \ge 0$ by e^{-1}, and hence $|\varphi'(s)| \le 1/\varrho$ on the interval $[\varrho, \infty)$.

Thus, we can now proceed with the same proof as Theorem 5, although we do not remove the stopping time ϑ. The end result is the following:

Theorem 6. *Given Assumptions 1, 3, 4, 5 and a quadratic $q : \mathbb{R}^d \to \mathbb{R}$, if $g(\mathbf{x}) = q(\mathbf{x} - \boldsymbol{\theta}_*)$ then choosing any fixed even moment $2p \in (0, d)$ and choosing any $\varrho > 0$, there exists a constant $C(\overline{\mathbf{K}}, \epsilon, \varrho) > 0$ such that for any $\delta \in (1/3, 1/2)$ and all $T > 0$,*

$$\sup_{0 \le t \le T \wedge \vartheta} |g(\boldsymbol{\theta}_{\lfloor td \rfloor}) - g(\boldsymbol{\Theta}_t)| \le \frac{3d^\delta \|g\|_{C^2}}{\sqrt{d}} \exp\left(C(\overline{\mathbf{K}}, \epsilon, \varrho) \|\eta\|_\infty T\right), \tag{117}$$

with probability at least $1 - c(2p, \overline{\mathbf{K}}) d^{p(1/3-\delta)}$ for a constant $c(2p, \overline{\mathbf{K}})$ independent to d.

We remark that if the risk of SIGNHSGD remains bounded away from 0, which will be the case for constant stepsize and nonzero noise, one could additionally show that ϑ does not occur with high probability. In that case, one can derive as a corollary of Theorem 6 a statement without ϑ.

A.4 Bounding martingale terms

Lemma 5. *For all $g \in Q$ and $1/3 < \delta < 1/2$,*

$$\sup_{0 \le k \le \lfloor Td \rfloor} |\mathcal{M}_k^{lin}| < \frac{d^\delta}{\sqrt{d}}, \tag{118}$$

with high-probability.

Proof. Recall that under τ_0, $\mathbf{v}_k \le L$. Moreover, given that $\|q\|_{C^2} \le 1$ and $\left\|\mathbf{R}(z; \overline{\mathbf{K}})\right\| \le M_R$, we also have that $\|g\|_{C^2}$ is uniformly bounded for all $g \in Q$. Therefore,

$$\begin{aligned} |\Delta \mathcal{M}_k^{lin}| &\le \frac{\eta}{\sqrt{d}} \|g\|_{C^2} (1 + \|\mathbf{v}_k\|) \\ &\le \frac{\eta}{\sqrt{d}} \|g\|_{C^2} (1 + L), \end{aligned} \tag{119}$$

almost-surely for all k. Now by Corollary 3, for $1/3 < s < \delta$ and even moment $2p < d$, there exists $C(2p, \mathbf{K}) > 0$ such that

$$\mathbb{P}\left(|\Delta\mathcal{M}_k^{lin}| \geq \frac{d^s}{d}\right) \leq \frac{C(2p,\mathbf{K})\mathbb{E}\left[\|\nabla g(\mathbf{v}_{k-1})\|^{4p}\right]^{1/2} d^{2p/3}}{d^{2ps}}$$
$$\leq C(2p,\mathbf{K})\,(1+L)^{2p}\,d^{2p\left(\frac{1}{3}-s\right)}. \tag{120}$$

Choose $s = \frac{1}{6} + \frac{\delta}{2}$. By Lemma 11 the result follows.

□

Lemma 6. *For all $g \in Q$ and $0 < s < 1/2$,*

$$\sup_{0\leq k\leq \lfloor Td\rfloor} |\mathcal{M}_k^{quad}| < \frac{1}{d^s} \tag{121}$$

with overwhelming probability.

Proof. From Cauchy-Schwarz, we see that

$$\left|\Delta\mathcal{M}_k^{quad}\right| \leq \frac{\eta^2}{d}\|g\|_{C^2} \tag{122}$$

Then, Azuma's inequality shows that

$$\mathbb{P}\left(\max_{1\leq k\leq \lfloor Td\rfloor} |\mathcal{M}_k^{quad}| \geq \frac{1}{d^s}\right) \leq 2\exp\left(\frac{-d^{-2s+1}}{C_T\eta^4\|g\|_{C^2}^2}\right), \tag{123}$$

which gives the result. □

Lemma 7. *For all $g \in Q$ and $s < 1$,*

$$\sup_{0\leq t\leq T} |\mathcal{M}_t^{\sigma}| \leq \frac{1}{d^s}, \tag{124}$$

with overwhelming probability.

Proof. From Equation (91), we know that

$$\mathcal{M}_t^{\sigma} = \eta\int_0^t \nabla q(\mathbf{V}_s)^{\mathrm{T}}\sqrt{\frac{2\mathbf{K}_\sigma}{d\pi}}\mathrm{d}\mathbf{B}_t. \tag{125}$$

Using the $\|q\|_{C^2}$ norm we can bound

$$\|\nabla g(\mathbf{V}_s)\| \leq \|g\|_{C^2}\,(1+\|\mathbf{V}_s\|). \tag{126}$$

Then, with Assumption 3 and Equation (126) we can bound the quadratic variation as,

$$\langle\mathcal{M}^{\sigma}\rangle_t = \frac{2\eta^2}{d\pi}\int_0^t \nabla g(\mathbf{V}_s^{\tau})^{\mathrm{T}}\mathbf{K}_\sigma\nabla g(\mathbf{V}_s^{\tau})\,\mathrm{d}s$$
$$\leq \frac{2\eta^2}{d\pi}\int_0^t \|\mathbf{K}_\sigma\|\,\|\nabla g(\mathbf{V}_s)\|^2\,\mathrm{d}s$$
$$\leq \frac{2\eta^2}{d\pi}\|\mathbf{K}_\sigma\|\,\|g\|_{C^2}^2\,(1+M)^2 t. \tag{127}$$

Then, using the subgaussian tail bound for continuous martingales we see that the stopped martingale satisfies,

$$\mathbb{P}\left(\sup_{0\le t\le T}|\mathcal{M}_t^\sigma|\ge t\right)\le 2\exp\left(\frac{-t^2 d\pi}{4\eta^2\left\|\mathbf{K}_\sigma\right\|\|g\|_{C^2}^2(1+M)^2T}\right). \tag{128}$$

□

Lemma 8. *If μ is a probability measure on $\mathbb{R}$ with the property that there exists $a_0>0$ such that $\frac{\mathrm{d}\mu}{\mathrm{d}x}=g(x)$ on $[-a_0,a_0]$ for $g\in C^2([-a_0,a_0])$, then the map $\alpha:\mathbb{R}^+\to\mathbb{R}$ defined by*

$$s\mapsto\frac{1}{\sqrt{s}}\int_{\mathbb{R}}\exp\left(\frac{-y^2}{2s}\right)\mathrm{d}\mu(y), \tag{129}$$

is bounded as well as Lipschitz.

Proof. Notice that it suffices to show (129) is bounded and Lipschitz for $0<s<1$. Let $f_s(y)=\frac{2}{\pi\sqrt{s}}\exp\left(\frac{-y^2}{2s}\right)$, as well as $G(y)=\mu((-\infty,y])$. Decomposing the integral into

$$\int_{\mathbb{R}}f_s(y)\,\mathrm{d}\mu(y)=\int_{[-a_0,a_0]}f_s(y)\,\mathrm{d}\mu(y)+\int_{\mathbb{R}\setminus[-a_0,a_0]}f_s(y)\,\mathrm{d}\mu(y), \tag{130}$$

we see that the latter term may be easily bounded by

$$\int_{\mathbb{R}\setminus[-a_0,a_0]}f_s(y)\,\mathrm{d}\mu(y)\le\frac{1}{\sqrt{s}}\exp\left(\frac{-a_0^2}{2s}\right), \tag{131}$$

which decays to 0 as $s\to 0$. The former term we apply the integration by parts formula to get

$$\int_{[-a_0,a_0]}f_s(y)\,\mathrm{d}\mu(y)=f_s(y)G(y)\Big|_{y=-a_0}^{y=a_0}+\int_{[-a_0,a_0]}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)G(y)\,\mathrm{d}y. \tag{132}$$

Further decomposing the latter integral into positive and negative regions we get

$$\int_0^{a_0}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)G(y)\,\mathrm{d}y=\int_0^{a_0}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)\left[G(-a_0)+\mu((-a_0,y])\right]\,\mathrm{d}y, \tag{133}$$

and

$$\int_{-a_0}^{0}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)G(y)\,\mathrm{d}y=\int_{-a_0}^{0}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)\left[G(-a_0)+\mu((-a_0,y])\right]\,\mathrm{d}y \tag{134}$$

$$=-\int_0^{a_0}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)\left[G(-a_0)+\mu((-a_0,-y])\right]\,\mathrm{d}y. \tag{135}$$

Thus,

$$\int_{[-a_0,a_0]}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)G(y)\,\mathrm{d}y=\int_0^{a_0}\frac{y}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)\mu((-y,y])\,\mathrm{d}y \tag{136}$$

$$\le C\int_0^{a_0}\frac{y^2}{s^{3/2}}\exp\left(\frac{-y^2}{2s}\right)\,\mathrm{d}y \tag{137}$$

$$=C\int_0^{a_0/\sqrt{s}}y^2\exp\left(\frac{-y^2}{2}\right)\,\mathrm{d}y. \tag{138}$$

Putting this all together, we conclude that $\varphi(s)$ is uniformly bounded for all $s>0$. To see lipschitz, we apply a similar argument. We first differentiate $f_s(y)$ with respect to s to we get

$$\frac{\mathrm{d}}{\mathrm{d}s}f_s(y)=\frac{1}{2s^{5/2}}\exp\left(\frac{-y^2}{2s}\right)\left(y^2-s\right). \tag{139}$$

Therefore,

$$\alpha'(s) = \int_{[-a_0,a_0]} \frac{\mathrm{d}}{\mathrm{d}s} f_s(y)\,\mathrm{d}\mu(y) + \int_{\mathbb{R}\setminus[-a_0,a_0]} \frac{\mathrm{d}}{\mathrm{d}s} f_s(y)\,\mathrm{d}\mu(y). \tag{140}$$

There exists $s_0 > 0$ such that if $s < s_0$, then $\sqrt{3s} < a_0$. It is easy to check that if $y > \sqrt{3s}$ then $\frac{\mathrm{d}}{\mathrm{d}s} f_s(y)$ is decreasing in y. Likewise, if $y < -\sqrt{3s}$ then $\frac{\mathrm{d}}{\mathrm{d}s} f_s(y)$ is increasing in y. It follows that

$$\int_{\mathbb{R}\setminus[-a_0,a_0]} \frac{\mathrm{d}}{\mathrm{d}s} f_s(y)\,\mathrm{d}\mu(y) \leq \frac{1}{s^{5/2}} \exp\left(\frac{-a_0^2}{2s}\right)(a_0^2 - s), \tag{141}$$

which decays to 0 as $s \to 0$. Finally, we apply integration by parts once more to get

$$\int_{[-a_0,a_0]} \frac{\mathrm{d}}{\mathrm{d}s} f_s(y) = \frac{\mathrm{d}}{\mathrm{d}s} f_s(y) G(y)\Big|_{y=-a_0}^{y=a_0} \tag{142}$$

$$+ \int_0^{a_0} \frac{1}{2s^{7/2}} \exp\left(\frac{-y^2}{2s}\right)(y^3 - 3sy)\mu((-y,y])\,\mathrm{d}y. \tag{143}$$

Since $g \in C^2([-a_0,a_0])$ we may express $\mu((-y,y])$ as

$$\mu((-y,y]) = \int_{-y}^{y} g(0) + g'(0)x + O(x^2)\,\mathrm{d}x = 2g(0)y + O(y^3). \tag{144}$$

Plugging this into (143) it is easy to check that

$$\left|\int_0^{a_0} \frac{g(0)}{s^{7/2}} \exp\left(\frac{-y^2}{2s}\right)(y^3 - 3sy)y\,\mathrm{d}y\right| = \frac{g(0)a_0^3 \exp\left(\frac{-a_0^2}{2s}\right)}{s^{5/2}}, \tag{145}$$

and

$$\left|\int_0^{a_0} \frac{1}{2s^{7/2}} \exp\left(\frac{-y^2}{2s}\right)(y^3 - 3sy)O(y^3)\,\mathrm{d}y\right| \leq C\int_0^{a_0} \frac{1}{2s^{7/2}} \exp\left(\frac{-y^2}{2s}\right)(y^3 + 3sy)y^3\,\mathrm{d}y \tag{146}$$

$$= C\int_0^{a_0/\sqrt{s}} \exp\left(-\frac{y^2}{2}\right)(y^3 + 3y)y^3\,\mathrm{d}y. \tag{147}$$

Combining this with (141), we conclude that $|\alpha'(s)|$ is uniformly bounded for all $s > 0$. □

Lemma 9. *Let $x \sim N(0, \mathbf{K})$ such that $\mathbf{K}$ is positive-definite. If $a \in \mathbb{R}^d$, then for all even moments $2k \leq d$,*

$$\mathbb{E}\left[\langle a, \sigma(\mathbf{x})\rangle^{2k}\right] \leq C(2k, \mathbf{K})\|a\|_\infty^{2k} d^{4k/3}, \tag{148}$$

where $C(2k, \mathbf{K}) > 0$ depends only on $2k$, $\lambda_{\min}(\mathbf{K})$ and $\lambda_{\max}(\mathbf{K})$.

Proof. We start by fixing a $\delta > 0$ and defining the smooth approximation of $\sigma(\mathbf{x})$ to be $\sigma_\delta(\mathbf{x}) = \rho_\delta * \sigma(\mathbf{x})$, where $\rho_\delta : \mathbb{R} \to \mathbb{R}$ is the standard compactly-supported mollifier convolved entry-wise to $\sigma(\mathbf{x})$, i.e. $(\sigma_\delta(\mathbf{x}))_i = \rho_\delta * \sigma(\mathbf{x}_i)$. It follows that

$$\|\langle a, \sigma(\mathbf{x})\rangle\|_{L^{2k}} \leq \|\langle a, \sigma(\mathbf{x}) - \sigma_\delta(\mathbf{x})\rangle\|_{L^{2k}} + \|\langle a, \sigma_\delta(\mathbf{x})\rangle\|_{L^{2k}}. \tag{149}$$

Note that ρ_δ has support contained in $[-\delta, \delta]$, thus the entry-wise difference of $\sigma(\mathbf{x}) - \sigma_\delta(\mathbf{x})$ may be bounded by

$$|\sigma(\mathbf{x}_i) - \sigma_\delta(\mathbf{x}_i)| \leq \begin{cases} 0 & |\mathbf{x}_i| > 2\delta \\ 2 & |\mathbf{x}_i| \leq 2\delta. \end{cases}$$

Define $N_\delta(\mathbf{x}) = \sum_{i=1}^d \mathbb{1}_{\{|\mathbf{x}_i| \leq 2\delta\}}$ to be the number of coordinates of $\mathbf{x}$ within the interval $(-2\delta, 2\delta)$, we see that

$$\mathbb{E}\left[\langle a, \sigma(\mathbf{x}) - \sigma_\delta(\mathbf{x})\rangle^{2k}\right] \leq \|a\|_\infty^{2k} 2^{2k}\mathbb{E}\left[N_\delta(\mathbf{x})^{2k}\right] \tag{150}$$

$$= \|a\|_\infty^{2k} 2^{2k} \sum_{s \in \mathcal{I}} \mathbb{P}\left((\mathbf{x}_i)_{i \in s'} \in [-2\delta, 2\delta]^{|s'|}\right), \tag{151}$$

where $\mathcal{I} = \{1, \ldots, d\}^{2k}$and s' is set of distinct elements of s. Let $\mathbf{K}^{(s')} = \mathbb{E}\left[(\mathbf{x}_i)_{i\in s'}^{\otimes 2}\right]$, then $(\mathbf{x}_i)_{i\in s'} \sim N(0, \mathbf{K}^{(s')})$. Recall that there exists a permutation matrix $\mathbf{P}$, such $\mathbf{K}^{(s')}$ forms the top $|s'| \times |s'|$ sub-matrix of $\mathbf{PKP}^{-1}$. Given that $\mathbf{PKP}^{-1}$ and $\mathbf{K}$ are similar, they share the same eigenvalues. Therefore, by the Cauchy interlacing-law,

$$\lambda_{\min}(\mathbf{K}) \leq \lambda_{\min}(\mathbf{K}^{(s')}). \tag{152}$$

In particular this implies

$$\det \mathbf{K}^{(s')} = \prod_{i=1}^{|s'|} \lambda_i(\mathbf{K}^{(s')}) \geq \lambda_{\min}(\mathbf{K})^{|s'|}. \tag{153}$$

Plugging this back into (151) and choosing $\delta = d^{-r}$, we get

$$\begin{aligned}
\mathbb{E}\left[\langle a, \sigma(\mathbf{x}) - \sigma_\delta(\mathbf{x})\rangle^{2k}\right] &\leq \|a\|_\infty^{2k}\, 2^{2k} \sum_{s\in\mathcal{I}} \frac{(4\delta)^{|s'|}}{(2\pi\lambda_{\min}(\mathbf{K}))^{|s'|/2}} && (154)\\
&= \|a\|_\infty^{2k}\, 2^{2k} \sum_{l=1}^{2k} \binom{d}{l} l! \left\{ {2k \atop l} \right\} \frac{(4\delta)^l}{(2\pi\lambda_{\min}(\mathbf{K}))^{l/2}} && (155)\\
&\leq \|a\|_\infty^{2k}\, 2^{2k} \max_{1\leq l\leq 2k} \left(l! \left\{ {2k \atop l} \right\} \right) \sum_{l=1}^{2k} \left(\frac{ed}{l}\right)^l \frac{(4\delta)^l}{(2\pi\lambda_{\min}(\mathbf{K}))^{l/2}} && (156)\\
&\leq \|a\|_\infty^{2k}\, 2^{2k} C(2k) \left(\frac{4ed^{1-r}}{\sqrt{2\pi \min\{\lambda_{\min}(\mathbf{K}), 1\}}} \right)^{2k} && (157)\\
&= \|a\|_\infty^{2k}\, C(2k, \mathbf{K}) d^{(1-r)2k}, && (158)
\end{aligned}$$

where $\left\{ {2k \atop l} \right\}$ is Stirling's number of a second-kind. For clarity of notation moving forward, we note that $C(2k, \mathbf{K})$ may change up to factors of constants or powers of k from line to line, while always being independent to d.

To control the second term of Equation (149), we modify the proof of concentration of Lipschitz functions of Gaussian random variables. See Lemma 2.1.5 for more details Adler & Taylor (2007). Let $G(\mathbf{x}) = \langle a, \sigma_\delta(\mathbf{x})\rangle$. and $\mathbf{z} \sim N(0, \mathbf{K})$ be independent to $\mathbf{x}$. Define the Gaussian interpolation $\mathbf{z}^\alpha$ to be

$$\mathbf{z}^{(\alpha)} = \alpha\mathbf{x} + \sqrt{1-\alpha^2}\mathbf{z},$$

and note that $\mathbf{x} \overset{\text{law}}{=} \mathbf{z}^{(\alpha)}$ for all $\alpha \in [0, 1]$. Then by Lemma 2.1.4 in Adler & Taylor (2007),

$$\mathbb{E}[G(\mathbf{x})^{2k}] = (2k-1)\int_0^1 \mathbb{E}\left[\left\langle \mathbf{K}(a \odot \sigma_\delta'(\mathbf{x})), a \odot \sigma_\delta'(\mathbf{z}^{(\alpha)})\right\rangle \cdot G^{2k-2}(\mathbf{x})\right] d\alpha, \tag{159}$$

where $\odot$ represents the Hadamard product. Going forward, we will use Hölder's inequality to break up the expectation and form a recursive equation. As such, consider

$$\mathbb{E}[\langle \mathbf{K}(a \odot \sigma_\delta'(\mathbf{x}), a \odot \sigma_\delta'(\mathbf{y}^\alpha)\rangle^{2p}], \tag{160}$$

for some p. Standard linear algebra gives us

$$\begin{aligned}
\mathbb{E}[\left\langle \mathbf{K}(a \odot \sigma_\delta'(\mathbf{x}), a \odot \sigma_\delta'(\mathbf{z}^{(\alpha)}\right\rangle^{2p}] &\leq (\|\mathbf{K}\|\, \|a\|_\infty^2)^{2p} \mathbb{E}[(\|\sigma_\delta'(\mathbf{x})\| \|\sigma_\delta'(\mathbf{z}^{(\alpha)})\|)^{2p}]\\
&\leq (\|\mathbf{K}\|\, \|a\|_\infty^2)^{2p} \mathbb{E}[\|\sigma_\delta'(\mathbf{x})\|\|^{4p}],
\end{aligned} \tag{161}$$

with the last line following from Cauchy-Schwartz and equality in law of $\mathbf{x}$ and $\mathbf{z}^{(\alpha)}$. Given that $\sigma_\delta'(\mathbf{x}_i) = 0$ for $\mathbf{x}_i \notin [-2\delta, 2\delta]$, as well as $|\sigma_\delta'(\mathbf{x}_i)| \leq \frac{L_\rho}{\delta}$ for $\mathbf{x}_i \in [-2\delta, 2\delta]$ and L_ρ a universal constant depending on our mollifier,

$$\mathbb{E}[\|\sigma'_\delta(\mathbf{x})\|\|^{4p}] \leq \frac{L_\rho^{4p}}{\delta^{4p}}\mathbb{E}[N_\delta^{2p}]. \tag{162}$$

As we have seen in Equation (158),

$$\mathbb{E}[N_\delta(\mathbf{x})^{2p}] \leq C(p,\mathbf{K})d^{(1-r)2p}. \tag{163}$$

Therefore, up to absolute constants

$$\begin{aligned}\mathbb{E}[\langle \mathbf{K}(a\odot\sigma'(\mathbf{x}), a\odot\sigma'(\mathbf{y}^\alpha)\rangle^{2p}] &\leq C(p,\mathbf{K})(\|a\|_\infty^2)^{2p}\frac{L_\rho^{4p}}{\delta^{4p}}d^{(1-r)2p} \\ &\leq C(p,\mathbf{K})(\|a\|_\infty^2)^{2p}d^{(1+r)2p}.\end{aligned} \tag{164}$$

Returning to (159) and choosing $2p = 2k-1$, we see by Hölder's inequality

$$\mathbb{E}[G(\mathbf{x})^{2k}] \leq (2k-1)\left\|G(\mathbf{x})^{2k-2}\right\|_{L^{\frac{2k-1}{2k-2}}}\|[\langle \mathbf{K}(a\odot\sigma'_\delta(\mathbf{x}), a\odot\sigma'_\delta(\mathbf{y}^\alpha)\rangle\|_{L^{2k-1}} \tag{165}$$

$$\leq (2k-1)\mathbb{E}[G(\mathbf{x})^{2k-1}]^{\frac{2k-2}{2k-1}}C(k,\mathbf{K})\|a\|_\infty^2 d^{1+r}. \tag{166}$$

Iterating the same inequalities as above for $\mathbb{E}[G^{2k-1}]$, we obtain

$$\mathbb{E}[G(\mathbf{x})^{2k}] \leq \prod_{i=1}^{2k-1}\left((2k-i)C(2k-i,\mathbf{K})\|a\|_\infty^2 d^{1+r}\right)^{\frac{2k-i}{2k-1}} \tag{167}$$

$$\leq C(2k,\mathbf{K})\|a\|_\infty^{2k} d^{(1+r)k}. \tag{168}$$

Equations (158) and (167)combined give control over Equation (149),

$$\|\langle a,\sigma(\mathbf{x})\rangle\|_{L^{2k}} \leq C(2k,\mathbf{K})\|a\|_\infty\left(d^{\frac{1+r}{2}}+d^{1-r}\right). \tag{169}$$

Optimizing over r yields $r = 1/3$ leading to

$$\mathbb{E}\left[\langle a,\sigma(\mathbf{x})\rangle\|^{2k}\right] \leq C(2k,\mathbf{K})\|a\|_\infty^{2k}d^{4k/3}, \tag{170}$$

as desired. □

Lemma 10. *Let $\mathbf{x}\sim N(0,\mathbf{K})$ such $\mathbf{K}$ is positive-definite. If $\mathbf{y}\in\mathbb{R}^d$ a random vector independent to $\mathbf{x}$, then for all even moments $2k\leq d$,*

$$\mathbb{E}\left[\langle \mathbf{y},\sigma(\mathbf{x})\rangle^{2k}\right] \leq C(2k,\mathbf{K})\mathbb{E}\left[\|\mathbf{y}\|^{4k}\right]^{1/2}d^{2k/3}, \tag{171}$$

where $C(2k,\mathbf{K})>0$ depends only on $2k$, $\lambda_{\min}(\mathbf{K})$ and $\lambda_{\max}(\mathbf{K})$.

Proof. The proof is almost identical to that of Lemma 9, but instead of taking the sup-norm of a in (151), we take the l_2 norm via the Cauchy Schwarz inequality. Now proceeding proceeding in a similar fashion we get

$$\mathbb{E}\left[\langle \mathbf{y},\sigma(\mathbf{x})-\sigma_\delta(\mathbf{x})\rangle^{2k}\right] \leq \mathbb{E}\left[\|\mathbf{y}\|^{2k}\|\sigma(\mathbf{x})-\sigma_\delta(\mathbf{x})\|^{2k}\right] \tag{172}$$

$$\leq \mathbb{E}\left[\|\mathbf{y}\|^{4k}\right]^{1/2}\mathbb{E}\left[\|\sigma(\mathbf{x})-\sigma_\delta(\mathbf{x})\|^{4k}\right]^{1/2} \tag{173}$$

$$\leq \mathbb{E}\left[\|\mathbf{y}\|^{4k}\right]^{1/2}2^{2k}\left(\mathbb{E}\left[N_\delta(\mathbf{x})^{2k}\right]\right)^{1/2} \tag{174}$$

$$\leq \mathbb{E}\left[\|\mathbf{y}\|^{4k}\right]^{1/2}C(k,\mathbf{K})d^{(1-r)k}. \tag{175}$$

Lastly, by the independence of $\mathbf{y}$ and $\mathbf{x}$, upon conditioning on $\mathbf{y}$ we see by Gaussian-concentration on $\langle \mathbf{y}, \sigma_\delta(\mathbf{x})\rangle$ that

$$\mathbb{E}\left[\langle \mathbf{y}, \sigma_\delta(\mathbf{x})\rangle^{2k}\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \mathbf{y}, \sigma_\delta(\mathbf{x})\rangle^{2k} \middle| \mathbf{y}\right]\right] \tag{176}$$

$$\leq \mathbb{E}\left[\left(\frac{C\sqrt{2k}\,\|\mathbf{y}\|\,\sqrt{\lambda_{\max}(\mathbf{K})}}{\delta}\right)^{2k}\right] \tag{177}$$

$$= C(2k, \mathbf{K})\mathbb{E}\left[\|\mathbf{y}\|^{2k}\right] d^{2kr}, \tag{178}$$

where $C > 0$ is an absolute constant. Combining (175) and (178) then optimizing in $r > 0$ yields the result. □

Corollary 1. *Let $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{x}_{k+1} \sim N(0, \mathbf{K})$, then for all even moments $2p \leq d$, there exists $C(2p, \mathbf{K}) > 0$ such that*

$$\mathbb{P}\left(\left|\langle \mathbf{a}, \sigma(\ell_{k+1}\mathbf{x}_{k+1}) - \mathbb{E}\left[\sigma(\ell_{k+1}\mathbf{x}_{k+1})|\mathcal{F}_k\right]\rangle\right| \geq t\right) \leq \frac{C(2p, \mathbf{K})\,\|\mathbf{a}\|_\infty^{2p}\, d^{4p/3}}{t^{2p}}. \tag{179}$$

Proof. For notational clarity, let us denote $Y = \langle \mathbf{a}, \sigma(\ell_{k+1}\mathbf{x}_{k+1})\rangle$. By Jensen's inequality and convexity of $x \mapsto x^{2p}$,

$$\mathbb{E}\left[|Y - \mathbb{E}[Y|\mathcal{F}_k]|^{2p}\right] \leq 2^{2p}\mathbb{E}\left[\frac{1}{2}Y^{2p} + \frac{1}{2}\mathbb{E}[Y|\mathcal{F}_k]^{2p}\right] \tag{180}$$

$$\leq 2^{2p}\mathbb{E}\left[Y^{2p}\right]. \tag{181}$$

However, notice that $\mathbb{E}\left[Y^{2p}\right] = \mathbb{E}\left[\langle \mathbf{a}, \sigma(\mathbf{x}_{k+1})\rangle^{2p}\right]$. By Markov's inequality and Lemma 9,

$$\mathbb{P}\left(|Y - \mathbb{E}[Y|\mathcal{F}_k]| \geq t\right) = \mathbb{P}\left(|Y - \mathbb{E}[Y|\mathcal{F}_k]|^{2p} \geq t^{2p}\right) \tag{182}$$

$$\leq \frac{2^{2p}\mathbb{E}\left[\langle \mathbf{a}, \sigma(\mathbf{x}_{k+1})\rangle^{2p}\right]}{t^{2p}} \tag{183}$$

$$\leq \frac{C(2p, \mathbf{K})\,\|\mathbf{a}\|_\infty^{2p}\, d^{4p/3}}{t^{2p}}. \tag{184}$$

□

Corollary 2. *If $\mathbf{a} \in \mathbb{R}^d$ such that $\max_{2\leq i\leq d}|a^i| = O\left(\frac{d^\delta}{\sqrt{d}}\right)$, then for all even moments $2p \leq d$ and $s > 0$, there exists $C(2p, \mathbf{K}) > 0$ such that,*

$$\mathbb{P}\left(\left|\langle \mathbf{a}, \sigma(\ell_{k+1}\mathbf{x}_{k+1}) - \mathbb{E}\left[\sigma(\ell_{k+1}\mathbf{x}_{k+1})|\mathcal{F}_k\right]\rangle\right| \geq d^s\right) \leq C(2p, \mathbf{K})d^{p\left(\frac{1}{3}-2s+2\delta\right)}, \tag{185}$$

provided that $d^s > 4|a^1|$.

Proof. For ease of notation, let $\sigma_{k+1} = \sigma(\ell_{k+1}\mathbf{x}_{k+1})$. Given that $|a^1\left(\sigma_{k+1} - \mathbb{E}\left[\sigma_{k+1}|\mathcal{F}_k\right]\right)| \leq 2|a^1|$, it follows by Corollary 1,

$$\mathbb{P}\left(\left|\langle \mathbf{a}, \sigma_{k+1} - \mathbb{E}\left[\sigma_{k+1}|\mathcal{F}_k\right]\rangle\right| \geq d^s\right) \leq \mathbb{P}\left(\left|\sum_{i=2}^d a^i\left(\sigma_{k+1}^i - \mathbb{E}\left[\sigma_{k+1}^i \middle| \mathcal{F}_k\right]\right)\right| \geq \frac{d^s}{2}\right) \tag{186}$$

$$\leq \frac{C(2p, \mathbf{K})\left(\max_{2\leq i\leq d}|a^i|\right)^{2p} d^{4p/3}}{d^{2ps}} \tag{187}$$

$$\leq C(2p, \mathbf{K})d^{p\left(\frac{1}{3}-2s+2\delta\right)}. \tag{188}$$

□

Corollary 3. *If $\mathbf{x}_{k+1} \sim N(0, \mathbf{K})$ and $\mathbf{y} \in \mathbb{R}^d$ a random vector independent to $\mathbf{x}$, then for all even moments $2p \leq d$, there exists $C(2p, \mathbf{K}) > 0$ and independent to d such that*

$$\mathbb{P}\left(\left|\left\langle \mathbf{y}, \sigma(\ell_{k+1}\mathbf{x}_{k+1}) - \mathbb{E}\left[\sigma(\ell_{k+1}\mathbf{x}_{k+1})|\mathcal{F}_k\right]\right\rangle\right| \geq t\right) \leq \frac{C(2p, \mathbf{K})\mathbb{E}\left[\|\mathbf{y}\|^{4p}\right]^{1/2} d^{2p/3}}{t^{2p}}. \tag{189}$$

The proof is identical to that of Corollary 1 but using Lemma 10 instead.

Lemma 11. *Let M_k be a martingale such that $|M_k - M_{k-1}| \leq C$ almost-surely and let $S_k \leq C$. Then for all $t > 0$ and $N > 0$,*

$$\mathbb{P}\left(|M_N| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left(C^2 + \sum_{k=1}^{N-1} S_k^2\right)}\right) + \mathbb{P}\left(\exists\, k \leq N-1\,, |M_k - M_{k-1}| > S_k\right). \tag{190}$$

Proof. Let $\tau = \min\{k\,;\, |M_k - M_{k-1}| > S_k\}$ and $Y_k = M_{k\wedge\tau}$, then on the event that $\{k < \tau\}$, $|Y_k - Y_{k-1}| \leq S_k$. On the other hand, if $\{\tau \leq k\}$ then $|Y_k - Y_{k-1}| \leq C$. Breaking the probability space into the event $\{\tau \leq N-1\}$ and its complement gives,

$$\mathbb{P}\left(|M_N| \geq t\right) \leq \mathbb{P}\left(|Y_N| \geq t\right) + \mathbb{P}\left(\tau \leq N-1\right). \tag{191}$$

Azuma's inequality completes the proof as

$$\mathbb{P}\left(|Y_N| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left(C^2 + \sum_{k=1}^{N-1} S_k^2\right)}\right). \tag{192}$$

□

B THE RISK UNDER ANISOTROPIC DATA

B.1 RISK CURVE DYNAMICS

If $\mathbf{V}_t = \mathbf{\Theta}_t - \boldsymbol{\theta}_*$ where $\mathbf{\Theta}_t$ solves the SDE given by (7) then Itô's lemma applied onto

$$q(x) = \frac{1}{2} x^T \mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}}) x, \tag{193}$$

yields

$$dq(\mathbf{V}_t) = -\frac{\eta\varphi(\mathcal{R}_t)}{\sqrt{2\mathcal{R}_t}} \mathbf{V}_t^T \left(\frac{\mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}}) + \mathbf{R}(z; \overline{\mathbf{K}})^T \mathbf{K}}{2} \right) \overline{\mathbf{K}} \mathbf{V}_t \tag{194}$$

$$+ \frac{2\eta^2}{\pi d} \left\langle \mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}}), \mathbf{K}_\sigma \right\rangle + \mathcal{M}_t^\sigma. \tag{195}$$

By resolvent identities, we know that

$$\mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}})\overline{\mathbf{K}} = z\mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}}) + \mathbf{K}. \tag{196}$$

Moreover,

$$\mathbf{R}(z; \overline{\mathbf{K}})^T \mathbf{K}\overline{\mathbf{K}} = (\mathbf{K}\overline{\mathbf{K}}\mathbf{R}(z; \overline{\mathbf{K}}))^T = (z\mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}}) + \mathbf{K})^T, \tag{197}$$

so

$$2\left(zq(\mathbf{V}_t) + \mathcal{R}_t\right) = \mathbf{V}_t^T \left(\frac{\mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}}) + \mathbf{R}(z; \overline{\mathbf{K}})^T \mathbf{K}}{2} \right) \overline{\mathbf{K}} \mathbf{V}_t. \tag{198}$$

Returning to Itô's we see that

$$dq(\mathbf{V}_t) = -\frac{2\eta\varphi(\mathcal{R}_t)}{\sqrt{2\mathcal{R}_t}} \left(zq(\mathbf{V}_t) + \mathcal{R}_t\right) + \frac{\eta^2}{\pi d} \operatorname{Tr}\left(\mathbf{K}\mathbf{R}(z; \overline{\mathbf{K}})\mathbf{K}_\sigma\right) + \mathcal{M}_t^\sigma. \tag{199}$$

To recover the risk $\mathcal{R}_t$, we once again turn towards the Cauchy-integral law as well as the Spectral Theorem. Indeed,

$$\overline{\mathbf{K}} = \sum_{i=1}^d \lambda_i(\overline{\mathbf{K}}) \mathbf{u}_i \otimes \mathbf{w}_i \qquad \mathbf{R}(z; \overline{\mathbf{K}}) = \sum_{i=1}^d \frac{1}{\lambda_i(\overline{\mathbf{K}}) - z} \mathbf{u}_i \otimes \mathbf{w}_i, \tag{200}$$

where $\mathbf{u}_i$ and $\mathbf{w}_i$ are left and right eigenvectors respectively of $\overline{\mathbf{K}}$. We may then write

$$q(\mathbf{V}_t) = \frac{1}{2} \sum_{i=1}^d \frac{1}{\lambda_i(\overline{\mathbf{K}}) - z} \mathbf{V}_t^T (\mathbf{K}\mathbf{u}_i \otimes \mathbf{w}_i) \mathbf{V}_t. \tag{201}$$

Denoting $r_i(t) = \frac{1}{2}\mathbf{V}_t^T(\mathbf{K}\mathbf{u}_i \otimes \mathbf{w}_i)\mathbf{V}_t$, then upon integrating over Γ_i, a closed curve enclosing only $\lambda_i(\overline{\mathbf{K}})$, we see that

$$dr_i = \oint_{\Gamma_i} \frac{dq(\mathbf{V}_t)}{-2\pi i} dz \tag{202}$$

$$= -\frac{2\eta\varphi(\mathcal{R}_t)}{\sqrt{2\mathcal{R}_t}} \lambda_i(\overline{\mathbf{K}}) r_i + \frac{\eta^2}{\pi d} \operatorname{Tr}\left(\mathbf{K}(\mathbf{u}_i \otimes \mathbf{w}_i)\mathbf{K}_\sigma\right) + \mathcal{M}_t^{i,\sigma}, \tag{203}$$

where

$$\mathcal{M}_t^{i,\sigma} = \oint_{\Gamma_i} \frac{\mathcal{M}_t^\sigma}{-2\pi i} dz. \tag{204}$$

By Lemma 7, $\mathcal{M}_t^{i,\sigma}$ vanishes as $d \to \infty$, thus excluding the martingale part of (203), we obtain a system of d-ODEs, SIGNODE:

$$\frac{dr_i}{dt} = -\frac{2\eta\varphi(\mathcal{R}_t)}{\sqrt{2\mathcal{R}_t}} \lambda_i(\overline{\mathbf{K}}) r_i + \frac{\eta^2}{\pi d} \operatorname{Tr}\left(\mathbf{K}(\mathbf{u}_i \otimes \mathbf{w}_i)\mathbf{K}_\sigma\right), \quad \text{for all } 1 \le i \le d, \tag{205}$$

such that

$$\sum_{i=1}^d r_i = \mathcal{R}_t. \tag{206}$$

B.2 RISK CURVES FOR VANILLA SGD

Following a similar approach, but taking

$$q(x) = \frac{1}{2} x^T \mathbf{R}(z; \mathbf{K}) x, \tag{207}$$

we may derive a system of d-ODEs for SGD. We note this is not novel, and a full derivation in much greater generality is in (Collins-Woodfin et al., 2023); see also (Collins-Woodfin et al., 2024) for a shorter discussion. Using the HSGD formulation of vanilla streaming SGD (Collins-Woodfin et al., 2024), we arrive at the VANILLAODE:

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = -2\eta\lambda_i(\mathbf{K})v_i + \frac{\eta(t)^2}{d}\lambda_i(\mathbf{K})(R_t^{SGD} + \mathfrak{v}^2/2), \quad \forall 1 \leq i \leq d \tag{208}$$

$$R_t^{SGD} = \sum_{i=1}^{d} \lambda_i(\mathbf{K}) v_i. \tag{209}$$

C CONVERGENCE AND PHASE-PROPERTIES OF THE ODES

Lemma 12. *If $\epsilon \sim N(0, \mathfrak{v}^2)$, then $\mathcal{R}(\mathbf{\Theta}_t)$ is bounded from above and below for all $t > 0$.*

Proof. Take $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{D}\mathbf{x}$ then plugging this into (90) we obtain

$$\mathrm{d}q(\mathbf{V}_t) = -\frac{4\eta}{\pi\sqrt{2\mathcal{R}(\mathbf{V}_t) + \mathfrak{v}^2}}\mathcal{R}(\mathbf{V}_t) + \frac{\eta^2}{\pi d}\operatorname{Tr}(\mathbf{K}_\sigma\mathbf{D})\,\mathrm{d}t + \mathcal{M}_t^\sigma. \tag{210}$$

By concentration inequalities we know that $\mathcal{M}_t^\sigma$ vanishes as $d \to \infty$, thus we will omit the martingale term. Solving for the stationary point yields the following roots,

$$\mathcal{R}_\pm = \frac{C_\eta^2 \pm C_\eta\sqrt{C_\eta^2 + 64\mathfrak{v}^2}}{64}, \tag{211}$$

where $C_\eta = \frac{\eta}{4d}\operatorname{Tr}(\mathbf{K}_\sigma\mathbf{D}) = \frac{\pi\eta}{8d}\operatorname{Tr}(\mathbf{D})$. Phase diagram analysis shows that if $\mathcal{R}(\mathbf{V}_t) < \mathcal{R}_+$, then $q(\mathbf{V}_t)$ is increasing. Conversely, if $\mathcal{R}(\mathbf{V}_t) > \mathcal{R}_+$ then $q(\mathbf{V}_t)$ is decreasing. Since $\mathbf{D}$ is positive-definite, $q(\mathbf{V}_t) > 0$ provided that $\mathbf{V}_t \neq 0$. The growth and decay conditions of $q(\mathbf{V}_t)$ implies that $q(\mathbf{V}_t)$ cannot converge to 0, nor diverge to ∞. Therefore, $q(\mathbf{V}_t)$ is bounded from above and below. Consequently, $\|\mathbf{V}_t\|$ is bounded from above and below and so $\mathcal{R}(\mathbf{V}_t)$ is as well. □

Theorem 7. *If $\epsilon \sim N(0, \mathfrak{v}^2)$ and $\eta \in (0, \infty)$ is a fixed learning rate then there exists a unique stationary point*

$$s_i = \frac{\eta\operatorname{Tr}(\mathbf{K}(\mathbf{u}_i \otimes \mathbf{w}_i)\mathbf{K}_\sigma)}{16\lambda_i(\overline{\mathbf{K}})d}\left(\frac{\eta\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma)}{d} + \sqrt{\frac{\eta^2\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma)^2}{d^2} + 16\mathfrak{v}^2}\right), \tag{212}$$

and the limit risk is given by

$$R_\infty = \frac{\eta}{16d}\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma)\left(\frac{\eta\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma)}{d} + \sqrt{\frac{\eta^2\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma)^2}{d^2} + 16\mathfrak{v}^2}\right). \tag{213}$$

We note that in these formulas, $\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma) = \frac{\pi}{2}\operatorname{Tr}(\mathbf{D}) = \frac{\pi}{2}\operatorname{Tr}(\overline{\mathbf{K}})$ on account of $\mathbf{K}_\sigma$ having a constant diagonal.

Proof. Let $Y_t = \frac{\pi\sqrt{2R_t + \mathfrak{v}^2}}{4\eta}$ and $m_i = \frac{\operatorname{Tr}(\mathbf{K}(\mathbf{u}_i \otimes \mathbf{w}_i)\mathbf{K}_\sigma)}{\pi d}$, then our d coupled ODEs are given by

$$\frac{\mathrm{d}r_i}{\mathrm{d}t} = -\frac{\lambda_i(\overline{\mathbf{K}})r_i}{Y_t} + \eta^2 m_i, \quad r_i(0) = \frac{1}{2}\mathbf{V}_0^T(\mathbf{K}\mathbf{u}_i \otimes \mathbf{y}_i)\mathbf{V}_0. \tag{214}$$

Solving for the stationary point, we see that for all $1 \leq i \leq d$,

$$r_i = \frac{\eta^2 m_i Y_t}{\lambda_i(\overline{\mathbf{K}})}. \tag{215}$$

Thus, at equilibrium

$$R_t = \sum_{i=1}^d r_i = \eta^2 Y_t \sum_{i=1}^d \frac{m_i}{\lambda_i(\overline{\mathbf{K}})} = \frac{\eta^2 Y_t}{\pi d}\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma). \tag{216}$$

However, R_t can be expressed in terms of Y by

$$R_t = \frac{1}{2}\left(\left(\frac{4\eta Y_t}{\pi}\right)^2 - \mathfrak{v}^2\right). \tag{217}$$

Plugging into (216) we see that

$$\frac{1}{2}\left(\left(\frac{4\eta Y_t}{\pi}\right)^2 - \mathfrak{v}^2\right) = \frac{\eta^2 Y_t}{\pi d}\operatorname{Tr}(\mathbf{D}\mathbf{K}_\sigma). \tag{218}$$

Solving for Y_t yields the following positive root

$$Y_\infty = \frac{\pi}{16\eta}\left(\frac{\eta\,\mathrm{Tr}(\mathbf{D}\mathbf{K}_\sigma)}{d} + \sqrt{\frac{\eta^2\,\mathrm{Tr}(\mathbf{D}\mathbf{K}_\sigma)^2}{d^2} + 16\mathfrak{v}^2}\right) \tag{219}$$

Therefore, by (215) and (217), $\frac{\mathrm{d}r_i}{\mathrm{d}t} = 0$ if and only if

$$r_i = s_i := \frac{\eta^2 m_i Y_\infty}{\lambda_i(\overline{\mathbf{K}})}, \quad \forall 1 \le i \le d. \tag{220}$$

This concludes uniqueness. The limiting risk is then given by

$$R_\infty = \frac{\eta^2 Y_\infty}{\pi d}\,\mathrm{Tr}(\mathbf{D}\mathbf{K}_\sigma). \tag{221}$$

□

Theorem 8. *Assume $\epsilon \sim N(0, \mathfrak{v}^2)$ and let s_i be the stationary points to (11a). Then there is an absolute constant $c > 0$ so that if*

$$\eta\left(\frac{\mathrm{Tr}(\mathbf{D}\mathbf{K}_\sigma)}{\pi d}\right) \le \min\left\{c, \frac{4\mathfrak{v}}{\pi}\right\}, \quad \textit{and} \quad R_0 \le c\mathfrak{v},$$

then we have, setting $R_\infty = \sum_{i=1}^d s_i$ to be the limit risk,

$$|R_t - R_\infty| \le 2(R_0 + R_\infty)e^{-t\eta\lambda_{\min}(\overline{\mathbf{K}})/(\pi\mathfrak{v})}.$$

We note again that in these formulas, $\mathrm{Tr}(\mathbf{D}\mathbf{K}_\sigma) = \frac{\pi}{2}\,\mathrm{Tr}(\mathbf{D}) = \frac{\pi}{2}\,\mathrm{Tr}(\overline{\mathbf{K}})$ on account of $\mathbf{K}_\sigma$ having a constant diagonal.

Proof. We recall (220), in terms of which we have

$$\frac{\mathrm{d}r_i}{\mathrm{d}t} = -\frac{\lambda_i(\overline{\mathbf{K}})}{Y_t}r_i + \frac{\lambda_i(\overline{\mathbf{K}})}{Y_\infty}s_i,$$

and where we recall

$$Y_t = \frac{\pi\sqrt{2R_t + \mathfrak{v}^2}}{4\eta}.$$

Then we rewrite the evolution of r_i as

$$\frac{\mathrm{d}}{\mathrm{d}t}(r_i - s_i) = -\frac{\lambda_i(\overline{\mathbf{K}})}{Y_\infty}(r_i - s_i) + \left(\frac{\lambda_i(\overline{\mathbf{K}})}{Y_\infty} - \frac{\lambda_i(\overline{\mathbf{K}})}{Y_t}\right)r_i,$$

and we set R_∞ as $\sum s_i$. Now we observe that

$$\frac{Y_t^2 - Y_\infty^2}{Y_\infty^2} = \frac{\pi^2}{8\eta^2 Y_\infty^2}\left(R_t - R_\infty\right) =: \alpha\left(R_t - R_\infty\right), \tag{222}$$

from which it follows

$$\frac{1}{Y_\infty} - \frac{1}{Y_t} = \frac{Y_t^2 - Y_\infty^2}{Y_t Y_\infty(Y_t + Y_\infty)} = \frac{Y_t^2 - Y_\infty^2}{2Y_\infty^3} + \mathrm{Err}_t,$$

where Err_t is bounded by

$$\mathrm{Err}_t \le C\frac{1}{\mathcal{Y}}\left(\frac{Y_t^2 - Y_\infty^2}{Y_\infty^2}\right)^2 \le C\frac{\alpha^2}{\mathcal{Y}}\left(R_t - R_\infty\right)^2, \tag{223}$$

where $\mathcal{Y}$ is the minimum value of Y_t over all time and C is an absolute constant. Hence we can further develop

$$\frac{\mathrm{d}}{\mathrm{d}t}(r_i - s_i) = -\left(\frac{1}{Y_\infty} - \frac{Y_t^2 - Y_\infty^2}{2Y_\infty^3} - \mathrm{Err}_t\right)\lambda_i(\overline{\mathbf{K}})(r_i - s_i) + \left(\frac{Y_t^2 - Y_\infty^2}{2Y_\infty^3} + \mathrm{Err}_t\right)\lambda_i(\overline{\mathbf{K}})s_i.$$

Define

$$\varrho(s) = \int_0^s \left(\frac{1}{Y_\infty} - \frac{Y_t^2 - Y_\infty^2}{2Y_\infty^3} - \mathrm{Err}_t \right) \mathrm{d}t.$$

Then by variation of parameters, we have

$$(r_i - s_i)(t) = (r_i - s_i)(0)e^{-\lambda_i \varrho(t)} + \int_0^t e^{-\lambda_i(\varrho(t)-\varrho(s))} \left(\frac{Y_s^2 - Y_\infty^2}{2Y_\infty^3} + \mathrm{Err}_s \right) \lambda_i(\overline{\mathbf{K}}) s_i \mathrm{d}s.$$

Now if we sum over all i, we have

$$R_t - R_\infty = \mathcal{F}(t) + \int_0^t \mathcal{K}(t,s) \left(\frac{Y_s^2 - Y_\infty^2}{2Y_\infty^3} + \mathrm{Err}_s \right) \mathrm{d}s,$$

where

$$\mathcal{F}(t) = \sum_i (r_i - s_i)(0)e^{-\lambda_i \varrho(t)} \quad \text{where} \quad \mathcal{K}(t,s) = \sum_i e^{-\lambda_i(\varrho(t)-\varrho(s))} \lambda_i(\overline{\mathbf{K}}) s_i.$$

Now suppose that on some interval of time $[0, T]$

$$\mathrm{Err}_t \leq \frac{Y_t^2 - Y_\infty^2}{2Y_\infty^3} \quad \text{and} \quad 2\frac{Y_t^2 - Y_\infty^2}{2Y_\infty^3} \leq \frac{1}{2Y_\infty}. \tag{224}$$

Then for $s < t < T$, we have

$$\varrho(t) - \varrho(s) \geq \frac{1}{2Y_\infty}(t - s),$$

and so we have the convolution Volterra upper bound for $t \leq T$

$$|R_t - R_\infty| \leq |\mathcal{F}(t)| + \frac{\alpha}{Y_\infty} \int_0^t \left(\sum_i e^{-\lambda_i(t-s)/(2Y_\infty)} \lambda_i(\overline{\mathbf{K}}) s_i \right) |R_s - R_\infty| \, \mathrm{d}s.$$

Now we note that we have the upper bound (for $t \leq T$)

$$|\mathcal{F}(t)| \leq (R_0 + R_\infty) e^{-(\lambda_{\min}/(2Y_\infty))t}.$$

Now suppose that $0 < T' \leq T$ is such that for $s \leq T'$

$$|R_s - R_\infty| \leq M e^{-(\lambda_{\min}/(4Y_\infty))t},$$

we have for $t \leq T'$,

$$\begin{aligned}
&\int_0^t \left(\sum_i e^{-\lambda_i(t-s)/(2Y_\infty)} \lambda_i(\overline{\mathbf{K}}) s_i \right) M e^{-(\lambda_{\min}/(4Y_\infty))s} \, \mathrm{d}s \\
&= \left(\sum_i e^{-\lambda_i t/(2Y_\infty)} \left(e^{\lambda_i t/(2Y_\infty) - \lambda_{\min} t/(4Y_\infty)} - 1 \right) \frac{\lambda_i(\overline{\mathbf{K}}) s_i}{\lambda_i/(2Y_\infty) - \lambda_{\min}/(4Y_\infty)} \right) M \\
&\leq e^{-\lambda_{\min} t/(4Y_\infty)} \left(\sum_i \frac{\lambda_i(\overline{\mathbf{K}}) s_i}{\lambda_i/(2Y_\infty) - \lambda_{\min}/(4Y_\infty)} \right) M \\
&\leq e^{-\lambda_{\min} t/(4Y_\infty)} \left(\sum_i \frac{\lambda_i(\overline{\mathbf{K}}) s_i}{\lambda_i/(2Y_\infty) - \lambda_i/(4Y_\infty)} \right) M \\
&\leq e^{-\lambda_{\min} t/(4Y_\infty)} (4Y_\infty R_\infty) M.
\end{aligned}$$

Hence $T' = T$, provided

$$4\alpha R_\infty < 1 \quad \text{and} \quad M = \frac{(R_0 + R_\infty)}{1 - 4\alpha R_\infty}.$$

Now we return to showing T does not occur. Recall (224), which up to T are satisfied. Then it suffices to have (compare (223)),

$$\alpha M \leq \frac{1}{2} \quad \text{and} \quad 2C\frac{Y_\infty}{\mathcal{Y}} \alpha M \leq 1, \quad \text{where} \quad \alpha = \frac{\pi^2}{8\eta^2 Y_\infty^2}.$$

in which case (224) is satisfied for all time. Note that we may always bound $\mathcal{Y}$ below by

$$\mathcal{Y} \geq (\pi \mathfrak{v})/(4\eta).$$

Define $H_{\mathbf{K}} = \frac{\mathrm{Tr}(\mathbf{D}\mathbf{K}_\sigma)}{\pi d}$. We now recall (219) and (217), from which

$$Y_\infty = \pi^2 \left(\frac{\eta H_{\mathbf{K}} + \sqrt{\eta^2 H_{\mathbf{K}}^2 + 16\mathfrak{v}^2/\pi^2}}{16\eta} \right) \quad \text{and} \quad R_\infty = Y_\infty \eta^2 H_{\mathbf{K}}.$$

Hence for $\eta H_{\mathbf{K}} \leq 4\mathfrak{v}/\pi$, we have

$$\frac{\pi \mathfrak{v}}{4\eta} \leq Y_\infty \leq \frac{\pi \mathfrak{v}}{\eta} \leq 4\mathcal{Y},$$

and hence we have

$$\alpha \leq \frac{2}{\mathfrak{v}} \quad \text{and} \quad R_\infty \leq \eta \pi \mathfrak{v} H_{\mathbf{K}}.$$

Thus we conclude there is an absolute constant $c > 0$ so that if

$$\eta H_{\mathbf{K}} \leq \min\left\{ c, \frac{4\mathfrak{v}}{\pi} \right\}, \quad \text{and} \quad R_0 \leq c\mathfrak{v},$$

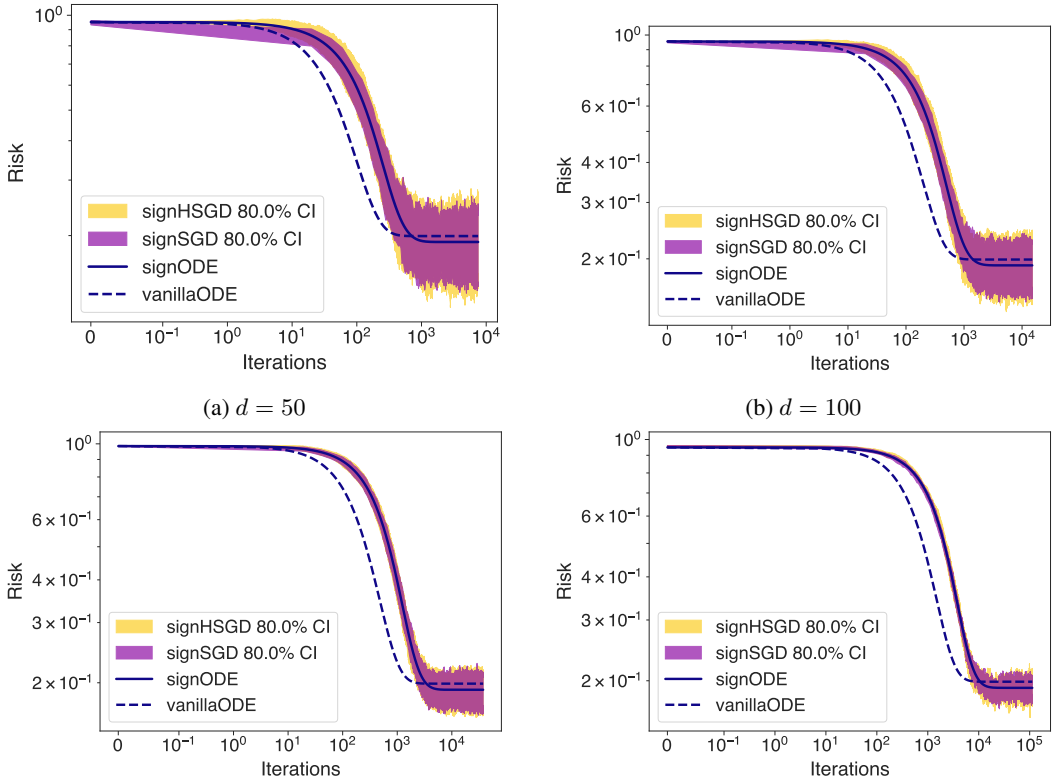
then we have

$$|R_t - R_\infty| \leq 2(R_0 + R_\infty) e^{-t\eta\lambda_{\min}(\mathbf{K})/(\pi \mathfrak{v})}.$$

□

D ADDITIONAL EXPERIMENTS

We begin by illustrating (Figure 5) the concentration effect: as d increases, the loss curves more closely match the ODEs. We also note the spread of SIGNHSGD and SIGNSGD are close across dimension.



(a) $d = 50$ (b) $d = 100$

(c) $d = 250$ (d) $d = 750$

Figure 5: A demonstration that SIGNSGD, SIGNHSGD, and their deterministic equivalent concentrate in high-dimensions over long time scales. In the limit as $d \to \infty$ our main theorem shows that all these objects become the same.

E HEURISTIC FOR ADAM

In this section we derive a heuristic for ADAM Kingma (2014). This is given, in our context, by:

Given:

η : learning rate

$\beta_1, \beta_2 \in [0, 1)$: exponential decay rates for moment estimates

ϵ_0 : small constant for numerical stability

Initialize:

$\boldsymbol{\theta}_0$: initial parameter vector

$\mathbf{m}_0 \leftarrow \mathbf{0}$: 1st moment vector

$\mathbf{v}_0 \leftarrow \mathbf{0}$: 2nd moment vector

$k \leftarrow 0$: timestep

Repeat until convergence:

$k \leftarrow k + 1$

$\mathbf{g}_k \leftarrow \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_{k-1}, \mathbf{x}_k, y_k)$ (Get gradients w.r.t. stochastic objective at timestep k)

$\mathbf{m}_k \leftarrow \beta_1 \cdot \mathbf{m}_{k-1} + (1 - \beta_1) \cdot \mathbf{g}_k$ (Update biased first moment estimate)

$\mathbf{v}_k \leftarrow \beta_2 \cdot \mathbf{v}_{k-1} + (1 - \beta_2) \cdot \mathbf{g}_k^2$ (Update biased second raw moment estimate)

$\hat{\mathbf{m}}_k \leftarrow \mathbf{m}_k / (1 - \beta_1^k)$ (Compute bias-corrected first moment estimate)

$\hat{\mathbf{v}}_k \leftarrow \mathbf{v}_k / (1 - \beta_2^k)$ (Compute bias-corrected second raw moment estimate)

$\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_{k-1} - \eta \hat{\mathbf{m}}_k / (\sqrt{\hat{\mathbf{v}}_k} + \epsilon_0)$ (Update parameters)

In a high-dimensional context, the first moment momentum β_1 has been observed to be equivalent to an effective change of learning rate, without inducing other benefits on the dynamics (see Paquette & Paquette (2021)), and so we ignore it.

The role of the second moment, in contrast, should induce a preconditioner. If we assume that exponential decay rate of β_2 is chosen sufficiently close to 1, we would have

$$\hat{\mathbf{v}}_k \approx_{\beta_2} \mathbb{E}(\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_{k-1}, \mathbf{x}_k, y_k))^2 | \mathscr{F}_{k-1}),$$

with the square applied entrywise. Using the definition of the stochastic gradient, we have

$$\hat{\mathbf{v}}_k = \mathbb{E}\left((\mathbf{x}_k)^2 \left(\langle \mathbf{x}_k, \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_* \rangle + \epsilon_k \right)^2 | \mathscr{F}_{k-1} \right).$$

This can be computed explicitly by Gaussian conditioning. Note that conditionally on the Gaussian $\mathbf{w} = \langle \mathbf{x_k}, \boldsymbol{\theta}_{\mathbf{k-1}} - \boldsymbol{\theta}_* \rangle$, $\mathbf{x}_k$ develops a mean $\mathbf{K}(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_*)$, which has norm $O(1)$. Hence provided it also has small L^∞ norm, so too will all the variances of the entries of $\mathbf{x}_k$ be nearly unaffected. Hence we essentially have independence, in that

$$\hat{\mathbf{v}}_k \approx \mathbb{E}\left((\mathbf{x}_k)^2 \right) \mathbb{E}\left(\left(\langle \mathbf{x}_k, \boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}_* \rangle + \epsilon_k \right)^2 | \mathscr{F}_{k-1} \right) = \mathrm{diag}(\mathbf{K})(2\mathcal{P}).$$

Hence, we arrive at the approximate update rule for ADAM

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \frac{\eta_k}{\sqrt{2\mathcal{P}(\boldsymbol{\theta}_k)}} \mathbf{D}^{-1} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_k, \mathbf{x}_{k+1}, y_{k+1}). \tag{225}$$

The corresponding homogenized ADAM equation is given by

$$\mathrm{d}\boldsymbol{\Theta}_t = -\frac{\eta_t}{\sqrt{2\mathcal{P}(\boldsymbol{\theta}_t)}} \overline{\mathbf{K}}(\boldsymbol{\Theta}_t - \boldsymbol{\theta}_*) + \eta_t \mathbf{D}^{-1/2} \sqrt{\overline{\mathbf{K}}} \mathrm{d}\mathbf{B}_t. \tag{226}$$

F COMPARISON WITH WEAK APPROXIMATION FRAMEWORK

Adaptive method approximations via SDEs have in fact been developed for ADAM in prior works. In particular, Malladi et al. (2022) adapts the weak approximation approach of Li et al. (2019) to produce an SDE for ADAM. Their method utilizes a noisy gradient model $\mathbf{g}_k = \nabla f(\boldsymbol{\theta}) + \sigma \mathbf{z}$, where

∇f is the expected gradient and $\mathbf{z}$ has mean 0 and covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. To more closely match our set-up, we will take f to be the quadratic-loss and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = 2\mathcal{P}(\boldsymbol{\theta})\mathbf{K}$. We note that the covariance of the quadratic-loss gradient typically involves higher moments, however in the high-dimensional setting it can be well-approximated by just $2\mathcal{P}(\boldsymbol{\theta})\mathbf{K}$. See Collins-Woodfin & Paquette (2023) for details. The ADAM weak approximation SDE is described by the following:
Let $c_1 = \frac{1-\beta_1}{\eta^2}$ and $c_2 = \frac{1-\beta_2}{\eta^2}$, then the ADAM SDE is given by the system

$$\mathrm{d}\boldsymbol{\Theta}_t^{\mathrm{WA}} = -\frac{1-\exp(-c_2 t)}{\sqrt{1-\exp(-c_1 t)}}\boldsymbol{Q}_t^{-1}\boldsymbol{m}_t \mathrm{d}t, \tag{227}$$

$$\mathrm{d}\boldsymbol{m}_t = c_1(\mathbf{K}(\boldsymbol{\Theta}_t^{\mathrm{WA}} - \boldsymbol{\theta}_*) - \boldsymbol{m}_t)\,\mathrm{d}t + \eta c_1 \sqrt{\boldsymbol{\Sigma}(\boldsymbol{\Theta}_t^{\mathrm{WA}})}\,\mathrm{d}\boldsymbol{B}_t, \tag{228}$$

$$\mathrm{d}\boldsymbol{u}_t = c_2\left(\mathrm{diag}\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Theta}_t^{\mathrm{WA}}\right)\right) - \boldsymbol{u}_t\right)\mathrm{d}t, \tag{229}$$

where

$$\boldsymbol{Q}_t = \eta\,\mathrm{diag}(\boldsymbol{u}_t)^{1/2} + \epsilon_0\sqrt{1-\exp(-c_2 t)}.$$

Heuristically, c_1 and c_2 relate to the normalizing factor of $\hat{\mathbf{m}}_k$ and $\hat{\mathbf{v}}_k$ respectively by $1-\beta_1^k \approx 1-\exp\left(-c_1 k\eta^2\right)$ and $1-\beta_2^k \approx 1-\exp\left(-c_2 k\eta^2\right)$. Malladi et al. (2022) show that the expectation of the SDE and optimizer across suitable test functions g is $O(\eta^2)$, i.e.

$$\max_{k=0,\ldots,\lfloor T/\eta^2\rfloor} \left|\mathbb{E}[g(\boldsymbol{\theta}_k)] - \mathbb{E}[g\left(\boldsymbol{\Theta}_{k\eta^2}^{\mathrm{WA}}\right)]\right| \leq C(g)\eta^2. \tag{230}$$

We can recover SIGNSGD from the ADAM algorithm by setting $\beta_1 = \beta_2 = \epsilon_0 = 0$. Formally following the recipe from Malladi et al. (2022), this means we should take $c_1 = c_2 = 1/\eta^2$ (note that this makes the heuristic fit $1-\beta_1^k \approx 1-\exp\left(-c_1 k\eta^2\right)$ and $1-\beta_2^k \approx 1-\exp\left(-c_2 k\eta^2\right)$ incorrect).

The SDE system for $\boldsymbol{\Theta}_t^{\mathrm{WA}}$ depends on η, and so in (230), as we send $\eta \to 0$, we are sending $c_1 = c_2 = 1/\eta^2$ to infinity. Therefore, to give a single SDE which approximates SIGNSGD, we can use the ideas of a slow-fast system to give a heuristic approximation for the limit:

$$\boldsymbol{m}_t \mathrm{d}t \approx \mathbf{K}(\boldsymbol{\Theta}_t^{\mathrm{WA}} - \boldsymbol{\theta}_*)\,\mathrm{d}t + \eta\sqrt{2\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{WA}})\mathbf{K}}\,\mathrm{d}\boldsymbol{B}_t, \tag{231}$$

$$\boldsymbol{u}_t \approx 2\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{WA}})\,\mathrm{diag}(\mathbf{K}). \tag{232}$$

Thus, $\boldsymbol{Q}_t \approx \eta\sqrt{2\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{WA}})\,\mathrm{diag}(\mathbf{K})}$. Plugging this into (227) gives (heuristically) the weak approximation SDE of SIGNSGD,

$$\mathrm{d}\boldsymbol{\Theta}_t^{\mathrm{sWA}} = -\frac{1}{\eta\sqrt{2\mathcal{P}(\boldsymbol{\Theta}_t^{\mathrm{sWA}})}}\overline{\mathbf{K}}(\boldsymbol{\Theta}_t^{\mathrm{sWA}} - \boldsymbol{\theta}_*)\,\mathrm{d}t + \sqrt{\mathrm{diag}(\mathbf{K})^{-1}\mathbf{K}}\,\mathrm{d}\boldsymbol{B}_t. \tag{233}$$

Recall, SIGNHSGD is given by

$$\mathrm{d}\boldsymbol{\Theta}_t = -\eta_t\frac{\varphi(\mathcal{R}(\boldsymbol{\Theta}_t))}{\sqrt{2\mathcal{R}(\boldsymbol{\Theta}_t)}}\overline{\mathbf{K}}(\boldsymbol{\Theta}_t - \boldsymbol{\theta}_*)\mathrm{d}t + \eta_t\sqrt{\frac{\mathbf{K}_\sigma}{\pi d}}\mathrm{d}\mathbf{B}_t. \tag{234}$$

Interestingly, we observe the same preconditioned effect in the form of $\overline{\mathbf{K}}$ as SIGNHSGD. However, the effects from the noise is notably different. Particularly, in the high-dimensional setting with non-Gaussian noise, higher moments of the label noise are an important feature of SIGNHSGD as seen in φ. See Section 4 on ϵ-compression. In contrast, (233) only requires up to second moments as seen in $\mathcal{P}$. This remains the case even if $\boldsymbol{\Sigma}$ is the true conditional covariance of the gradient. Moreover, the diffusion matrix between the two SDEs are also quite different. In SIGNHSGD $\frac{2}{\pi}\mathbf{K}_\sigma$ corresponds precisely to the conditional covariance of the sample sign-gradient. This may suggest that a more delicate limit approach is required or that SIGNSGD falls outside the scope of the weak approximation setting. We believe that to extend the ADAM weak approximation SDE to non-continuous gradient transformations like SIGNSGD, the constants c_1 and c_2 must be uncoupled from η in order for β_1 and β_2 to have unrestricted limits.

We develop a different method of error bounds between the statistics of the SDE approximate and the optimizer (SIGNSGD), i.e. high-dimensional bounds versus weak approximation. For ease

of comparison between (230) and (9), let us take g to be the risk $\mathcal{R}$. The weak approximation theorem states that the expected risk between the SDE and optimizer is $O(\eta^2)$. This is more akin to convergence of distributional properties between the SDE and optimizer. In contrast, in Theorem 1 we show that the exact risk dynamics of SIGNHSGD and SIGNSGD closely track each other in the high-dimensional limit. This is what allows us to directly study the behavior of the SIGNSGD by studying SIGNHSGD. Our goal is not only to derive SDEs for signSGD but to also gain insight into how adaptive methods like signSGD and eventually Adam, behave in the limit of large problem sizes. The aspect of dimensionality is not addressed in Li et al. (2019) or Malladi et al. (2022) thus requires a different set of tools.

G $\overline{\mathbf{K}}$ DOES NOT ALWAYS REDUCE THE CONDITION NUMBER

As a counter example consider the covariance matrix,

$$\mathbf{K} = \begin{bmatrix} 0.17 & -0.49 & -0.19 & -0.36 \\ -0.49 & 2.34 & 0.71 & 1.79 \\ -0.19 & 0.71 & 0.32 & 0.53 \\ -0.36 & 1.79 & 0.53 & 1.44 \end{bmatrix} \tag{235}$$

with up to two decimals the condition number is $\kappa(\mathbf{K}) = 115.88$. However, the condition number of $\overline{\mathbf{K}}$ is $\kappa(\overline{\mathbf{K}}) = 129.78$.

H EXPERIMENTAL DETAILS

H.1 SIGNSGD AND HOMOGENIZED SIGNSGD

The code to reproduce these results is available at https://anonymous.4open.science/r/signSGD-6216.

H.1.1 FIGURE 1

The experiments creating Figure 1 were carried out on an M1 Macbook Air. Homogenized SIGNHSGD is solved via a standard Euler-Maruyama algorithm. The procedure for solving for the risk is described in Appendix B.

The synthetic data was generated in dimension $d = 500$. The covariance matrix was generated by multiplying a random unitary matrix by a diagonal matrix of log-spaced eigenvalues.

φ was explicitly computed in the Gaussian data case and was solved via numerical integration in the case of Cauchy (Student-t family) noise. Note that vanilla SGD does not converge under Cauchy noise and thus we cannot provide a comparison. For both experiments, $\eta = 0.5$. We plot the 80% confidence interval across 100 runs.

The CIFAR10 (Krizhevsky, 2009) data was used to perform binary classification by regressing to ± 1 labels being animals or vehicles. The data matrix D is first passed through a random features model so that

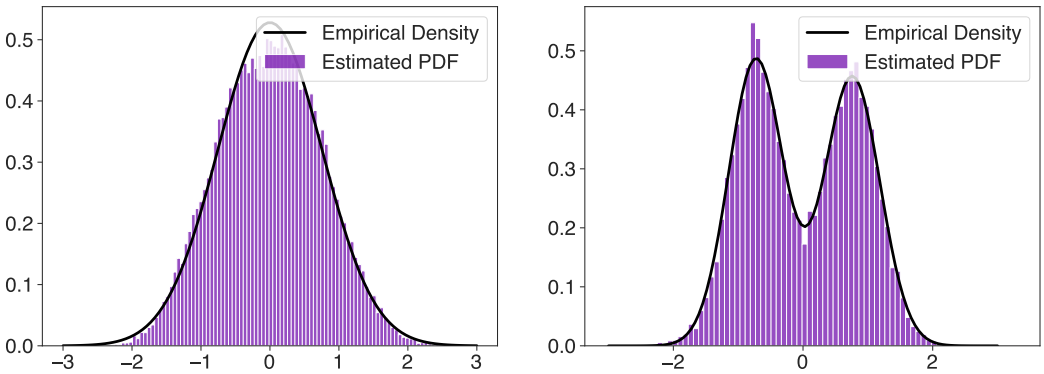
$$D_{rf} = \tanh DA \tag{236}$$

where A is a random features matrix of independent standard Gaussians. In order to estimate $\boldsymbol{\theta}_*$ the regression problem was first solved using Sci-kit learn (Pedregosa et al., 2011) and the resulting solution was taken to be $\boldsymbol{\theta}_*$. The differences $\{y_i - \langle \boldsymbol{\theta}_*, \mathbf{x}_i \rangle\}$ for all $\mathbf{x}_i \in D_{rf}$ was then assumed to be the noise. A histogram of this noise is available in Figure 6a. The noise was then fitted to a Gaussian. Finally, $\eta = 0.5$ and the SIGNSGD plot represents the 80% confidence interval over 50 runs.

The IMDB dataset (Maas et al., 2011) was first embedded using GLOVE (Pennington et al., 2014) into dimension 50. Then, a 2-layer random features model was applied as well as some noise added to regularize the problem. We add sG where G is a matrix of independent standard Gaussians. We take $s = 0.03$. This additional regularization was required in the case of text data as the covariance of the original GLOVE embedded data (ironically?) has extremely high condition number making numerically solving our ODEs impossible. Ultimately the data used is,

$$D_{rf} = \tanh(A' \tanh D(A + sG)). \tag{237}$$

Note that this regularization did not destroy the information contained in the original problem. Sci-kit learn achieves an accuracy of $\approx 75\%$ on the unregularized problem and the finally accuracy on the regularized problem was $\approx 74\%$.



(a) CIFAR10 noise

(b) IMDB noise

Figure 6: Histograms of the estimated noise distributions for the CIFAR10 and IMDB datasets. Also shown is the estimated PDFs used to compute φ for each case.

We perform the same method as in the CIFAR10 case to first estimate $\boldsymbol{\theta}_*$ and then to estimate the distribution of the noise. We fit a mixture of two Gaussians model (GMM) to this noise. φ is trivial to compute exactly when noise is assumed to come from a GMM. The noise is again available in Figure 6.