Journal of Machine Learning Research 24 (2023) 1-47 Submitted 8/20; Published 2/23

Online Stochastic Gradient Descent with Arbitrary Initialization Solves Non-smooth, Non-convex Phase Retrieval

Yan Shuo Tan YANSHUO@NUS.EDU.SG
Department of Statistics and Data Science
National University of Singapore
Singapore

Roman Vershynin RVERSHYN@UCI.EDU
Department of Mathematics
University of California
Irvine, CA 92697-3875, USA

Editor: Prateek Jain

Abstract

In recent literature, a general two step procedure has been formulated for solving the problem of phase retrieval. First, a spectral technique is used to obtain a constant-error initial estimate, following which, the estimate is refined to arbitrary precision by first-order optimization of a non-convex loss function. Numerical experiments, however, seem to suggest that simply running the iterative schemes from a random initialization may also lead to convergence, albeit at the cost of slightly higher sample complexity. In this paper, we prove that, in fact, constant step size online stochastic gradient descent (SGD) converges from arbitrary initializations for the non-smooth, non-convex amplitude squared loss objective. In this setting, online SGD is also equivalent to the randomized Kaczmarz algorithm from numerical analysis. Our analysis can easily be generalized to other single index models. It also makes use of new ideas from stochastic process theory, including the notion of a summary state space, which we believe will be of use for the broader field of non-convex optimization.

Keywords: Phase retrieval, SGD, non-convex, random initialization, optimization

1. Introduction

The mathematical phase retrieval problem is that of recovering a d-dimensional signal vector $\mathbf{x}^* \in \mathbb{R}^d$ (or $\mathbb{C}^d$) from the magnitudes $b^{(k)} = |\langle \mathbf{a}^{(k)}, \mathbf{x}^* \rangle|$ $(k = 1, 2 \ldots, N)$ of its projections onto a collection of known sampling vectors $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(N)} \in \mathbb{R}^d$ (or $\mathbb{C}^d$). Clearly, one can only hope to recover $\mathbf{x}^*$ up to global phase shift (i.e. up to multiplication by $e^{i\phi}$ for $\phi \in [0, 2\pi)$), but mild regularity assumptions on the sampling vectors and N large enough ensures that this is the only ambiguity.

This problem is well motivated by practical concerns, having applications to Coherent Diffraction Imaging (CDI), Electron Microscopy, and X-ray Crystallography, and as such has been a topic of study from at least the early 1980s. We refer the reader to the survey papers Fienup (1982); Shechtman et al. (2015); Bendory et al. (2017) for a comprehensive account

of the contexts in which the problem arises, as well as the techniques that practitioners employ to solve it.

Over the last decade, the phase retrieval problem has also garnered substantial attention from the optimization and machine learning communities, the reason being that it can be formulated as a relatively benign non-convex optimization problem. In other words, it can be solved by minimizing the least squares objective:

$$f(\mathbf{x}) = \frac{1}{2N}\sum_{k=1}^{N}\Big(|\langle \mathbf{a}^{(k)}, \mathbf{x}\rangle|^2 - (b^{(k)})^2\Big)^2. \tag{1}$$

Many papers have attempted to study how to optimize this objective given distributional assumptions on the sampling vectors. One popular approach is to use a two-step procedure: First, a spectral technique is used to obtain an initial estimate $\mathbf{x}^{(0)}$ so that its distance from a global minimum, $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2$, is bounded above by a small constant. Next, the estimate is refined to arbitrary precision using an iterative method such as gradient descent or stochastic gradient descent (SGD). This procedure is well-supported by theoretical guarantees for both real and complex signals. Here we note that one can also construct loss functions for phase retrieval which are different from (2). Running variations of gradient descent or SGD on these functions also leads to provable guarantees (see for instance Candès et al. (2015); Zhang et al. (2016); Wang et al. (2017); Wei (2015); Jeong and Güntürk (2017); Tan and Vershynin (2018)).

Nonetheless, this state of affairs is not entirely satisfactory from an optimization theory perspective, since the use of a spectral initialization diminishes the novelty of being able to provably minimize a non-convex objective using first-order methods. The spectral initialization essentially allows the first-order method to begin within a "basin of convexity", thus artificially escaping the difficulties of non-convexity. Such a *deus ex machina* may not be available when trying to optimize other non-convex functions.

Consequently, it is important to investigate whether first order methods converge to $\mathbf{x}^*$ from a random or even arbitrary initialization. From here on, we work with only real signals and sampling vectors, i.e. $\mathbf{x}^*, \mathbf{a}^{(k)} \in \mathbb{R}^d$. Sun et al. (2018) performed numerical experiments to show that this was indeed the case. They also analyzed the landscape of the loss function (1) in the real setting, and showed that it enjoyed favorable qualities: given $N = \Omega(d \cdot \mathrm{polylog}(d))$ measurements, there are no spurious minima and all saddle points are proper. Therefore, saddle-point escaping methods like perturbed gradient descent enjoy polynomial time convergence.

More recently, Chen et al. (2018) showed that vanilla gradient descent converges in $O(\log d)$ iterations, again given $N = \Omega(d \cdot \mathrm{polylog}(d))$ measurements. Up to log factors, this matches the running time guarantees for the original two-stage method. While an important step, their analysis still requires full gradient updates and does not apply to SGD. In the high-dimensional setting, SGD is particularly advantageous because it can be applied in an online, streaming fashion. This lowers the space complexity of the algorithm from $O(Nd)$ to $O(d)$, and allows progress toward the solution to be made even before the analyst gains access to the full data sample.

Moreover, with respect to the non-smooth *amplitude least squares* objective

$$f(\mathbf{x}) = \frac{1}{2N}\sum_{k=1}^{N}\Big(|\langle \mathbf{a}^{(k)}, \mathbf{x}\rangle| - b^{(k)}\Big)^2, \tag{2}$$

the question of even gradient descent convergence remains open. This objective is especially interesting because numerical simulations have shown gradient descent and SGD with respect to it to succeed with fewer measurements than are necessary for the alternative objective (1) (Wei, 2015; Wang et al., 2017).

1.1 Main results

In this paper, we prove that for a real signal vector and real sampling vectors, online stochastic gradient descent for the non-smooth objective (2) converges to a global minimum from arbitrary initializations given $\Omega(d \log d)$ Gaussian measurements. We believe our work to be among the first in establishing convergence of SGD in the non-smooth, non-convex regime. Furthermore, it will be readily apparent that our analysis framework generalizes easily to other single index models, and we conjecture that similar techniques will also work low-rank matrix sensing models in general.

We perform SGD with respect to (2), using a single data point per iteration. More formally, we form a sequence of signal estimates $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ with the update rule:

$$\mathbf{x}^{(k+1)} \coloneqq \mathbf{x}^{(k)} + \eta\Big(\operatorname{sign}(\langle \mathbf{a}^{(k)}, \mathbf{x}^{(k)}\rangle)b^{(k)} - \langle \mathbf{a}^{(k)}, \mathbf{x}^{(k)}\rangle\Big)\mathbf{a}^{(k)}. \tag{3}$$

We typically choose $\eta = \frac{1}{d}$. This is the same as in previous work that analyzed SGD as part of the two-step approach (Tan and Vershynin, 2018), and allows for a clear geometric interpretation. At each step k, we receive the datum $\big(\mathbf{a}^{(k)}, b^{(k)}\big)$; the solution set to the corresponding equation $|\langle \mathbf{a}^{(k)}, \mathbf{x}\rangle| = b^{(k)}$ is then the union of two parallel hyperplanes. Taking an SGD step projects the current iterate $\mathbf{x}^{(k-1)}$ onto the closer of these hyperplanes. This iterative projection is strongly reminiscent of the randomized Kaczmarz algorithm for solving linear systems. For a more in-depth discussion of this connection, we again refer the reader to Tan and Vershynin (2018).

We make the following assumptions for the rest of the paper.

Assumption 1 *Assume the following for each positive integer k:*

(A) (Real signal and sampling vectors) We have $\boldsymbol{x}^*, \boldsymbol{a}^{(k)} \in \mathbb{R}^d$.

(B) (Fresh independent measurements) At step k of the algorithm, we use a sampling vector $\boldsymbol{a}^{(k)}$ that is fully independent of the previous measurements $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(k-1)}$.

(C) ("Gaussian"[1] measurements) We have $\boldsymbol{a}^{(k)} \sim \sqrt{d}\mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$.

(D) (No noise) We have $b_k = |\langle \boldsymbol{a}^{(k)}, \boldsymbol{x}^*\rangle|$.

1. Because of concentration of norm, this is almost equivalent to requiring that each $\mathbf{a}_k$ has a standard Gaussian distribution in $\mathbb{R}^d$.

These assumptions are all somewhat common in the literature. We note that they are relatively strong, and each can probably be relaxed with some effort. Our paper makes a conceptual first step, and does not aim for the widest generality and weakest assumptions. We leave these extensions to future work, and discuss them in more detail in Section 8.

The following is the main result of the paper.

Theorem 2 (Main result) *Suppose we run the iterative update* (3) *in* $\mathbb{R}^d$ *under the assumptions above, and with constant step size* $\eta = \frac{\eta_0}{d}$ *where* η_0 *is a small enough universal constant. Then with probability at least* $0.8 - 1/d - C/\log d$*, there is a stopping time* $T \lesssim d \cdot \left(\log d + \log\left(\frac{\|\boldsymbol{x}^{(0)}\|}{\|\boldsymbol{x}^*\|} \vee 1\right)\right)$*, such that for all* $k \geq T$*, we have*

$$\|\boldsymbol{x}^{(k)} - \sigma \boldsymbol{x}^*\|^2 \leq \left(1 - \frac{1}{2d}\right)^{k-T} \|\boldsymbol{x}^*\|^2, \tag{4}$$

where $\sigma = \operatorname{sign}(\langle \boldsymbol{x}^{(k)}, \boldsymbol{x}^* \rangle)$*. Furthermore, there is some constant* D*, such that if* $d \geq D$*, we may choose* $\eta_0 = 1$*.*

This theorem tells us that in T steps, SGD brings us to a point in the "basin of convexity" around a global minimum, within which we get linear convergence. As a consequence of this theorem, it is easy to see that we can get an ϵ-relative-error estimate using $O(d \log(d/\epsilon))$ measurements (and the same number of steps). In fact, the analysis in Tan and Vershynin (2018) tells us that within the "basin of convexity", linear convergence still holds if we resample from amongst $O(d)$ measurements, and we no longer need Assumption A to hold for this stage of this of the process. By doing this resampling, our final sample complexity is $O(d \log d)$ even for exact recovery.

The novelty of this result is not simply the convergence of the algorithm, but rather its convergence with *near optimal sample and time complexity*. Indeed, it is well known that as we let the step size decrease to zero, SGD under our assumptions approximates gradient flow for the population loss function. It is easy to show that gradient flow converges to a global minimum. With smaller steps, however, the algorithm takes a longer time to converge.

In addition, our chosen step size $\frac{1}{d}$ is the smoothness parameter for SGD if we were optimizing a least squares system objective under the same assumptions on $\mathbf{a}^{(k)}$. It is interesting that although our objective is no longer smooth, $\frac{1}{d}$ nonetheless remains the right scaling.

The proof of the result is somewhat complicated and makes use of several new ideas. The first idea is to think of the sequence of SGD iterates as a Markov chain on a two-dimensional *summary state space* $\mathcal{Y}$. The two coordinates are the squared Euclidean norm of the iterate, $r^2 = \|\mathbf{x}\|^2$, and the correlation with the signal, $s = \langle \mathbf{x}, \frac{\mathbf{x}^*}{\|\mathbf{x}^*\|} \rangle$. Just as how in thermodynamics state variables such as temperature, pressure, and volume suffice to determine the evolution of a thermodynamic system, so too in our case do the state variables r^2 and s suffice to determine the progress of the SGD algorithm.

The state space is obviously independent of the dimension d. In fact, one can show that the distribution of the update in the state space is effectively independent of the dimension up to overall scaling (see Theorem 3). As such, as d tends to infinity, the stochastic dynamics of the Markov chain when initialized at a *fixed* $\mathbf{y} \in \mathcal{Y}$ approximates the solution of an ODE system whose corresponding vector field is given by the rescaled drift of the process.

Unfortunately, we are hit with the curse of dimensionality: if we take a random initialization $\mathbf{x}^{(0)} \sim \mathcal{N}(0, \mathbf{I})$, then it is well-known that with high probability,

$$s_0 = \frac{|\langle \mathbf{x}^{(0)}, \mathbf{x}^* \rangle|}{\|\mathbf{x}^{(0)}\| \|\mathbf{x}^*\|} \lesssim \frac{1}{\sqrt{d}}. \tag{5}$$

This implies that the initial correlation with the signal decays as the dimension increases, and in fact, at the corresponding point (r_0^2, s_0) in the state space, the drift and the fluctuations have the same magnitude, and it is no longer appropriate to approximate the stochastic dynamics with a deterministic process.

Overcoming this is the most difficult part of the proof. To do so, we use a small-ball probability argument. In other words, we show that the distribution of $s_K = s(\mathbf{x}^{(K)})$ is anti-concentrated away from 0 when K is large enough. This involves comparing the process $s_0, s_1, \dots$ with a more well-understood process $\hat{s}_0, \hat{s}_1, \dots$ via stochastic dominance. The distribution of this new sequence can in turn be controlled via recursive inequalities bounding 4th moments from above and 2nd moments from below. We conclude by applying the Paley-Zygmund inequality.

1.2 Related work

1.2.1 Phase retrieval

There is already a large body of work on phase retrieval, and it is impossible to give a full account of the literature. We have already mentioned survey papers for how phase retrieval arises in various engineering problems. On the theoretical side, we have already discussed the two-step non-convex optimization approach, and will further mention here the convex relaxation approaches pioneered in the papers Candès et al. (2013); Goldstein and Studer (2018); Bahmani and Romberg (2017); Hand and Voroninski (2016).

1.2.2 SGD as a Markov chain

It has long be observed that constant step-size SGD can be thought of as a Markov chain, and there seems to be a resurgence of interest in this view of SGD. For instance, Dieuleveut et al. (2017) used this approach to analyze the limiting distribution of SGD iterates for strongly convex functions. Furthermore, Mandt et al. (2017) used this interpretation to see how SGD can be used as a sampling algorithm. Both these works have drawn inspiration from the recent body of work on Langevin algorithms for sampling from log-concave distributions. The idea of analyzing SGD through diffusion approximation is also present in Li et al. (2016).

1.2.3 Non-convex optimization and first-order methods

The Kaczmarz method is a classical method in numerical analysis for solving large scale overdetermined linear systems. A randomized version of it was first analyzed by Strohmer and Vershynin (2009). In our earlier work (Tan and Vershynin, 2018), we proposed adapting the method to the setting of phase retrieval, where it coincides with SGD under the Gaussian measurement setting,

Stochastic first-order methods have emerged as the optimization method of choice for modern machine learning. In particular, deep neural networks are trained almost exclusively using SGD and variants like ADAM. The loss functions for these models, however, are non-convex functions, for which there has traditionally been little theory on how first-order methods behave.

Unsurprisingly, there has been a concerted push over the last few years to address this issue. One line of work studies how gradient descent or SGD can be made to escape saddle points quickly (Ge et al., 2015; Jin et al., 2017, 2019). Another line of work has focused on identifying regimes for shallow and deep neural networks for which gradient descent or SGD can be shown to converge (see for instance Mei et al. (2018); Allen-Zhu et al. (2018)).

1.3 Notation

Scalars are denoted in normal font, whereas vectors and matrices are denoted in bold. Subscripts (usually) denote components of a vector, while superscripts denote the index of a quantity when it is part of a sequence. Sets and events are denoted using calligraphic font. The indicator of a set $\mathcal{A}$ is denoted by $\mathbf{1}_{\mathcal{A}}$. The Euclidean norm is denoted with subscript omitted: $\|-\|$. No other norms are used in this paper, so there is no risk of confusion. When X is a subexponential random variable, $\|X\|_{\psi_1}$ denotes its subexponential norm (for a definition, see Appendix B). We use $\mathcal{B}(\mathbb{R}^d)$ to denote the Borel σ-algebra for $\mathbb{R}^d$. Sequences of the form $x_0, x_1, x_2, \ldots$ are denoted by $\{x_k\}_k$. Throughout the paper, C, c, C_1, C_2 and C_3 denote positive constants that may change from line to line.

2. Outline of proof

We start by making some simplifying assumptions. First, note that the algorithm and our guarantee (4) are both invariant with respect to scaling and rotation. As such, we may assume without loss of generality that $\mathbf{x}^* = \mathbf{e}_1$, the first coordinate basis vector. We also only analyze the case where d is large enough so that $\eta_0 = 1$ and the step size is set to be $\eta = \frac{1}{d}$. The extension to smaller d will be obvious.

2.1 State space Markov chain

The state space is the set $\mathcal{Y} \coloneqq \left\{\mathbf{y} = (r^2, s) \in \mathbb{R}^2 \,:\, s^2 \le r^2\right\}$, where r is the Euclidean norm of the SGD iterate and s is its projection onto the signal direction. In other words, the natural projection map $\pi \colon \mathbb{R}^d \to \mathcal{Y}$ is defined by

$$r^2(\mathbf{x}) \coloneqq \|\mathbf{x}\|^2 \qquad \text{and} \qquad s(\mathbf{x}) \coloneqq \langle \mathbf{x}, \mathbf{x}^* \rangle.$$

The reason we choose to use r^2 instead of r is for the convenience of obtaining formulas for the stochastic update, as will be evident later. We further define $\theta = \theta(\mathbf{x}) \coloneqq \arccos(s/r)$. This is the smaller angle between $\mathbf{x}$ and $\mathbf{x}^*$.

Note that we can track the progress of SGD purely in terms of the state variables. Indeed, we have

$$\begin{aligned}\|\mathbf{x} - \operatorname{sign}(\langle \mathbf{x}, \mathbf{x}^* \rangle)\mathbf{x}^*\|^2 &= \|\mathbf{x}\|^2 - 2|\langle \mathbf{x}, \mathbf{x}^* \rangle| + \|\mathbf{x}^*\|^2 \\ &= r^2 - 2|s| + 1 \\ &=: \Psi(r^2, s), \end{aligned} \tag{6}$$

so that the error of the k-th step estimate $\mathbf{x}^{(k)}$ is equal to $\Psi(\pi(\mathbf{x}^{(k)}))$. Note that $-\mathbf{x}^*$ and $\mathbf{x}^*$ are mapped onto $(1, -1)$ and $(1, 1)$, so that Ψ is uniquely minimized at these values. We hence wish to show that r^2 and s coordinates of our iterates converge to 1 and ± 1 respectively.

It is clear that the SGD sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$ is a Markov chain on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, with the update rule (3) giving a random mapping representation for the transition kernel. Now, consider the sequence $\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots$ where we define $\mathbf{y}^{(k)} \coloneqq \pi(\mathbf{x}^{(k)})$ for each index k. We are now ready to state the first key insight:

Theorem 3 *The sequence $\boldsymbol{y}^{(0)}, \boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \dots$ is a Markov chain on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ whose transition kernel has the random mapping representation*

$$\boldsymbol{y}^{(k+1)} = \boldsymbol{y}^{(k)} + \frac{1}{d}\Big(\alpha(\boldsymbol{y}^{(k)}), \beta(\boldsymbol{y}^{(k)})\Big), \tag{7}$$

where

$$\alpha(r^2, s) \coloneqq (1 - r^2\cos^2\theta) \cdot u^2 - r^2 \sin^2\theta \cdot v^2 + r^2 \sin\theta\cos\theta \cdot uv, \tag{8}$$

$$\beta(r^2, s) \coloneqq (1 - r\cos\theta - 2\mathbf{1}_{\mathcal{A}}) \cdot u^2 - r\sin\theta \cdot uv. \tag{9}$$

Here, the randomness is supplied by (u, v), which is a 2-dimensional marginal of the uniform distribution on $\sqrt{d}\mathbb{S}^{d-1}$, while $\mathcal{A} = \mathcal{A}(\theta)$ is the event that $\operatorname{sign}(\cos\theta u + \sin\theta v) \neq \operatorname{sign}(u)$.

Proof Deferred to Appendix A. ■

This theorem tells us that the state space sequence $\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots$ suffices not just to track our progress, as discussed earlier in the section, but also to determine its own dynamics. We hence no longer need to concern ourselves with the original SGD sequence, and instead work with this object for the rest of the paper. Henceforth, we let $\{\mathcal{F}_k\}_k$ denote the filtration defined by this sequence.

2.2 Doob decomposition and continuous time limit as $d \to \infty$

Let us try to understand the random mappings (8) and (9) better. It is well-known that (u, v) converges in distribution to a standard 2-dimensional Gaussian $\mathcal{N}(0, \mathbf{I}_2)$ as the ambient dimension d tends to infinity. Therefore, the only essential dependence of the update formula (7) on d is through the overall $\frac{1}{d}$ scaling. If we think of the indices $k = 1, 2, \dots$ as a time variable, rescale time by a factor of $\frac{1}{d}$, we can think of the sequence as being generated by an Euler discretization of a continuous-time process.

While we do not actually take this approach in our rigorous analysis, is it instructive to see what intuition this gives us. To do this, we do a Doob decomposition of the process $\{\mathbf{y}^{(k)}\}_{k=0}^{\infty}$, separating it into a drift term and a fluctuation term. Denote the drift terms using

$$\bar{\alpha}(\mathbf{y}) \coloneqq \mathbb{E}\{\alpha(\mathbf{y})\} \qquad \text{and} \qquad \bar{\beta} \coloneqq \mathbb{E}\{\beta(\mathbf{y})\}.$$

Letting (α_j, β_j) denote the random mapping used in the j-th step of the Markov chain, we have

$$\mathbf{y}^{(k)} - \mathbf{y}^{(0)} = \underbrace{\frac{1}{d}\sum_{j=1}^{k}\Big(\bar{\alpha}(\mathbf{y}^{(j-1)}), \bar{\beta}(\mathbf{y}^{(j-1)})\Big)}_{\text{drift}} + \underbrace{\frac{1}{d}\sum_{j=1}^{k}\Big(\alpha_j(\mathbf{y}^{(j-1)}) - \bar{\alpha}(\mathbf{y}^{(j-1)}), \beta_j(\mathbf{y}^{(j-1)}) - \bar{\beta}(\mathbf{y}^{(j-1)})\Big)}_{\text{fluctuation}}. \tag{10}$$

We now try to do a heuristic comparison of the relative magnitudes of the two terms. Suppose k is small enough so that we have $\mathbf{y}^{(j)} \approx \mathbf{y}^{(0)}$ for $j = 1, \ldots, k$. Then the drift can be approximated by

$$\frac{1}{d}\sum_{j=1}^{k}\Big(\bar{\alpha}(\mathbf{y}^{(j-1)}), \bar{\beta}(\mathbf{y}^{(j-1)})\Big) \approx \frac{k}{d}\Big(\bar{\alpha}(\mathbf{y}^{(0)}), \bar{\beta}(\mathbf{y}^{(0)})\Big).$$

Meanwhile, denoting $\tilde{\alpha}_j \coloneqq \alpha_j - \bar{\alpha}_j$, and $\tilde{\beta}_j \coloneqq \beta_j - \bar{\beta}_j$ for $j = 1, \ldots, k$, we also have

$$\frac{1}{d}\sum_{j=1}^{k}\Big(\tilde{\alpha}_j(\mathbf{y}^{(j-1)}), \tilde{\beta}_j(\mathbf{y}^{(j-1)})\Big) \approx \frac{1}{d}\sum_{j=1}^{k}\Big(\tilde{\alpha}_j(\mathbf{y}^{(0)}), \tilde{\beta}_j(\mathbf{y}^{(0)})\Big),$$

so that the fluctuation term has components whose standard deviations are approximately equal to

$$\frac{\sqrt{k}}{d}\mathrm{Var}\{\alpha(\mathbf{y}^{(0)})\}^{1/2}, \quad \text{and} \quad \frac{\sqrt{k}}{d}\mathrm{Var}\{\beta(\mathbf{y}^{(0)})\}^{1/2}.$$

Therefore, for any *fixed* $\mathbf{y}^{(0)}$, we see that the drift dominates the fluctuations as k and d tend to infinity. This means that the continuous time limit of the process trajectory should be an *integral curve* associated to the vector field on the state space $\mathcal{Y}$ defined by $\big(\bar{\alpha}, \bar{\beta}\big)$. While this picture is incomplete, it offers a good first approximation, and the next step we take is to analyze the solutions to this first order ODE system.

2.3 Drift in continuous time limit

Miraculously, it is actually possible to derive a closed form formula for the vector field. We state it in the following lemma.

Lemma 4 (Formula for drift) *With the notation* $\bar{\alpha}(\boldsymbol{y}) \coloneqq \mathbb{E}\{\alpha(\boldsymbol{y})\}$ *and* $\bar{\beta} \coloneqq \mathbb{E}\{\beta(\boldsymbol{y})\}$, *we have*

$$\bar{\alpha}(r^2, s) = 1 - r^2, \tag{11}$$

$$\bar{\beta}(r^2, s) = 1 - s - \frac{1}{\pi}(2\theta - \sin(2\theta)). \tag{12}$$

Proof Deferred to Appendix A. ■

Studying the vector field plot in Figure 1, it is obvious that $\mathbf{y}^* := (1,1)$ and $-\mathbf{y}^* = (-1,1)$ are the only attracting fixed points, with basins of attraction the sets $\mathcal{Y}_+ := \mathcal{Y} \cap \{s > 0\}$ and $\mathcal{Y}_- := \mathcal{Y} \cap \{s < 0\}$ respectively. While this assures us that the system has the right qualitative long-term behavior, the visualization alone is not sufficient to give quantitative bounds on convergence rates. This analysis turns out to be somewhat tricky. Given an integral curve $\bar{\mathbf{y}}^{(t)} = (\bar{r}_t^2, \bar{s}_t)$ starting from an arbitrary initialization $\bar{\mathbf{y}}^{(0)}$, we will analyze its convergence rate by breaking it into three separate phases, as depicted in the figure.

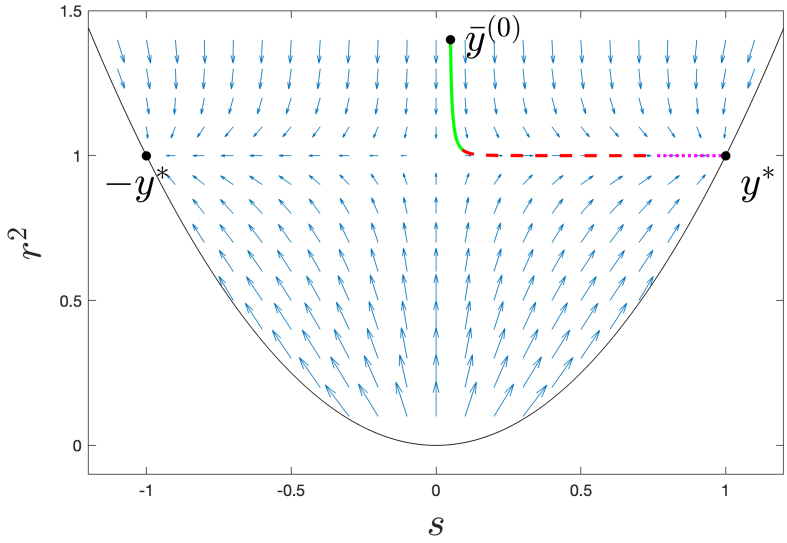

Figure 1: Vector field defined by $\left(\bar{\alpha}, \bar{\beta}\right)$. For the purposes of symmetry, we have chosen to plot r^2 along the vertical axis and s along the horizontal axis. An integral curve from an initialization $\mathbf{y}^{(0)}$ is plotted. Our analysis for the convergence rate will be broken into three phases. Phase 1 concerns the portion of the curve colored in green, Phase 2 concerns that colored in red, and Phase 3 concerns the last portion colored in magenta.

To demarcate the phases, we define two stopping times as follows. We let $\bar{\tau}_1$ be the earliest time t for which $\left|\bar{r}_t^2 - 1\right| \leq 0.1$, and we let $\bar{\tau}_2$ be the earliest time t for which the Lyapunov function Ψ defined in (6) satisfies $\Psi(\bar{\mathbf{y}}^{(t)}) \leq 0.2$. Phase 1 is then the portion of the curve traversed between time 0 and time $\bar{\tau}_1$, Phase 2 the portion traversed between time $\bar{\tau}_1$ and time $\bar{\tau}_2$, with Phase 3 the remainder of the curve traversed after $\bar{\tau}_2$.

Let us compute the duration of Phase 1, which is the same as bounding $\bar{\tau}_1$. To do this, we solve (11) to get $\bar{r}_t^2 - 1 = e^{-t}(\bar{r}_0^2 - 1)$, so that $\bar{\tau}_1 \lesssim \log(\bar{r}_0^2) \vee 1$. The second phase is

trickier due to the unwieldiness of (12). As such, we compute more amenable lower and upper bounds for the expression as follows.

Lemma 5 (Bounds for horizontal drift) *There is a constant $\bar{b}_{max}$ such that we have*

$$\sup_{r \geq 1/2} \frac{\bar{\beta}(r^2, s)}{s} \leq \bar{b}_{max}.$$

Furthermore, for any $\epsilon > 0$ small enough, there is some $\eta = \eta(\epsilon) > 0$ and some constant $\bar{b}_{min} = \bar{b}_{min}(\epsilon) > 0$ such that

$$\inf_{(r^2,s) \in \mathcal{D}} \frac{\bar{\beta}(r^2, s)}{s} \geq \bar{b}_{min},$$

where $\mathcal{D} \coloneqq \{(r^2, s) \in \mathcal{Y} : \ |s| \leq 1 - \epsilon, |r^2 - 1| \leq \eta\}$.

Proof Deferred to Appendix C. ■

One can show that $\eta(0.1) \geq 0.1$, and therefore, we have the bound $\frac{d\bar{s}_t}{dt} \geq \bar{b}_{min}\bar{s}$ for $\bar{\tau}_1 \leq t \leq \bar{\tau}_2$. Solving this gives $\bar{s}_t \geq \bar{s}_{\bar{\tau}_1} e^{\bar{b}_{min}(t - \bar{\tau}_1)}$, and we have the estimate $\bar{\tau}_2 - \bar{\tau}_1 \lesssim \log(1/|\bar{s}_{\bar{\tau}_1}|) \vee 1$.

Finally, Phase 3 corresponds to portion of the integral curve that lies within the "basin of convexity" around $\mathbf{y}^*$. Indeed we compute:

$$\begin{aligned}
\frac{d}{dt}\Psi(\bar{\mathbf{y}}^{(t)}) &= \partial_{r^2}\Psi \cdot \bar{\alpha} + \partial_s \Psi \cdot \bar{\beta} \\
&= 1 - \bar{r}_t^2 - 2\left(1 - \bar{s}_t - \frac{1}{\pi}\big(2\bar{\theta}_t - \sin(2\bar{\theta}_t)\big)\right) \\
&\lesssim -\big(\bar{r}_t^2 - 2\bar{s}_t + 1\big) \\
&= -\Psi(\bar{\mathbf{y}}^{(t)}).
\end{aligned}$$

Here, the inequality in the third line comes from a relative bound on the error term $\frac{1}{\pi}\big(2\bar{\theta}_t - \sin(2\bar{\theta}_t)\big)$ provided by the geometry of the basin region. As such, we also get linear convergence $\Psi(\bar{\mathbf{y}}^{(t)}) \leq \Psi(\bar{\mathbf{y}}^{(\bar{\tau}_2)}) e^{-c(t - \bar{\tau}_2)}$.

Putting everything together, we see that for any $\epsilon > 0$, if we would like $\Psi(\bar{\mathbf{y}}_t) \leq \epsilon$, it suffices for

$$t \gtrsim \log(\bar{r}_0^2) \vee 1 + \log(1/|\bar{s}_{\bar{\tau}_1}|) \vee 1 + \log(1/\epsilon). \tag{13}$$

2.4 Discretizing the drift

We now examine what this means for the Markov chain $(r_k^2, s_k) = \mathbf{y}^{(k)} = \mathbf{y}^{(k,d)}$, where for clarity, we have made the dependence on d in (7) explicit as a component of the indexing. We have argued that the integral curve $\bar{\mathbf{y}}^{(t)}$ is the limit of the trajectories $\{t \mapsto \mathbf{y}^{(\lfloor t/d \rfloor, d)}\}$ as d tends to infinity. The number of iterations k needed for convergence of $\mathbf{y}^{(k)}$ to $\pm\mathbf{y}^*$ is thus the quantity (13) scaled by a factor of d.

There may also be some additional dependence on d implicit in the definitions of $\bar{r}_0^2$ and $\bar{s}_{\bar{\tau}_1}$. The first has to do with the conditioning of the problem, and it is reasonable to assume that $\bar{r}_0^2 \lesssim 1$. On the other hand, if we take a random initialization, measure concentration inflicts us with the curse of dimensionality shown in (5), so that $\log(1/|\bar{s}_{\bar{\tau}_1}|) \gtrsim \log d$. The final time complexity estimate is therefore $O(d \log(d/\epsilon))$.

2.5 Accounting for fluctuations

Comparing the time complexity estimate for the drift process given in the last section with the guarantee given by Theorem 2 suggests to us that the fluctuations do not ultimately affect the rate of convergence for the random process $\mathbf{y}^{(k)}$. This is true, and indeed, we may make our heuristic arguments rigorous and show that the fluctuation "error term" is dominated by the drift over the discretized versions of Phase 1 and Phase 3.

The situation is more complicated for Phase 2. While the length of the phase for the random process remains the same, the underlying *dynamics* of the random process is very different from that of the deterministic drift process. This is because when $|s_k| \lesssim 1/\sqrt{d}$, as would be the case under random initialization, the magnitude of the fluctuations are of the same order as the drift, thereby invalidating the approximation argument.

There is no hope of upper bounding the fluctuations of $\mathbf{y}^{(k)}$ around iteration $k \approx d\bar{\tau}_1$, so we change course and instead aim at bounding the cumulative variance of the fluctuations *from below*. We seek a small ball probability bound for the horizontal marginal $s_{d\bar{\tau}_1+k'}$ for $k' \asymp d \log d$. More precisely, we would like

$$\lim_{d\to\infty} \mathbb{P}\{|s_{d\bar{\tau}_1+cd\log d}| \geq \eta\} \geq \delta \tag{14}$$

for some universal constants c, η and δ. Once s_k is of a constant distance away from zero, we may return to using the approximation argument and treat the fluctuations as an error term.

2.6 Small ball probability bound through diffusion approximation

In order to obtain a small ball probability bound, we will need a better understanding of the distribution of the fluctuation term in (10), rather than merely control over its tail probabilities. Fortunately, this term constitutes a martingale difference process, and the martingale Central Limit Theorem tells us that it converges to Brownian motion. In other words, we should expect the random process $\mathbf{y}^{(\lfloor t/d \rfloor, d)}$ to be well-approximated by a *diffusion* process $\tilde{\mathbf{y}}^{(t)}$ satisfying the stochastic differential equation

$$d\tilde{\mathbf{y}}^{(t)} = \Big(\bar{\alpha}(\tilde{\mathbf{y}}^{(t)}), \bar{\beta}(\tilde{\mathbf{y}}^{(t)})\Big)dt + \frac{\mathbf{\Sigma}(\tilde{\mathbf{y}}^{(t)})}{\sqrt{d}} d\mathbf{B}^{(t)}.$$

Here, $d\mathbf{B}^{(t)}$ is standard, two-dimensional Brownian motion, while for each $\mathbf{y}$, $\mathbf{\Sigma}(\mathbf{y})$ is a positive semidefinite matrix reflecting the fluctuation covariance at state $\mathbf{y}$.

It is not easy to compute a closed form solution to this SDE. To perform an heuristic analysis, we instead solve a simplified form of the equation for the s-marginal:

$$d\tilde{s}_t = b\tilde{s}_t dt + \frac{\sigma}{\sqrt{d}} dB_t.$$

This is a good approximation if we assume $|\tilde{r}_t^2 - 1|, |\tilde{s}_t| \ll 1$, so that $\mathbf{\Sigma}(\mathbf{y})$ is approximately constant in $\mathbf{y}$, while $\bar{\beta}(r^2, s)$ is close to its linear approximation in s. Solving this equation gives

$$\tilde{s}_t = \tilde{s}_0 e^{bt} + \frac{e^{bt}}{\sqrt{d}} \int_0^t e^{-b\tau} dB_\tau \sim \mathcal{N}\left(\tilde{s}_0 e^{bt}, \frac{e^{2bt} - 1}{bd}\right).$$

As such, no matter the value of $\tilde{s}_0$, we have that $\tilde{s}_{c \log d}$ is a Gaussian with variance bounded below by a constant, which implies the desired estimate (14).

2.7 Outline for rest of paper

In this section so far, we have sketched a heuristic proof for SGD convergence, where the main idea was to consider the continuous time limit of the state space Markov chain, and solve the resulting differential equation or stochastic differential equation. In our rigorous analysis, we will not adopt this approach, and instead solve finite difference equations coming from the Markov chain while obtaining non-asymptotic control over the fluctuation process.

For the rest of this paper, we work with a fixed initialization $\mathbf{y}^{(0)}$, and let $\mathbf{y}^{(1)}, \mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ be the sequence of iterates generated by repeated applying the Markov kernel (7). For each k, we will use r_k^2 and s_k to denote the coordinates of $\mathbf{y}^{(k)}$. We will continue to do a multi-phase analysis of convergence and, as such, define analogues of the stopping times $\bar{\tau}_1$ and $\bar{\tau}_2$. We set

$$\tau_1 \coloneqq \min\{k \colon |r_k^2 - 1| \le \log d / \sqrt{d}\} \tag{15}$$

$$\tau_{2a} \coloneqq \min\{k \ge \tau_1 \colon |s_k| \ge \gamma_1\} \tag{16}$$

$$\tau_{2b} \coloneqq \min\{k \ge \tau_{2a} \colon \Psi(\mathbf{y}^{(k)})\} \le \gamma_2. \tag{17}$$

Here, γ_1 and γ_2 are constants to be determined later.

We shall set $T = \tau_{2b}$ in Theorem 2. As such, we wish to prove that with high probability, $\tau_{2b} \lesssim d \log\big(d\|\mathbf{x}^{(0)}\| \vee 1\big)$, and that linear convergence in expectation occurs after τ_{2b}. The second statement follows easily from the theory we established in Tan and Vershynin (2018), while the first is, as mentioned before, the main result of this paper. Our strategy is to bound τ_{2b} by bounding τ_1, $\tau_{2a} - \tau_1$, and $\tau_{2b} - \tau_{2a}$ separately.

In Section 3, we bound τ_1, and also establish uniform control over $|r_k^2 - 1|$, which is needed for the rest of the proof. In Sections 4 and 5, we bound $\tau_{2a} - \tau_1$. This is the most difficult part of the proof, and relies on developing a nuanced notion of stochastic dominance with which we compare the sequence $\{s_k\}_k$ with a carefully constructed sequence $\{\hat{s}_k\}_k$. We then apply a small ball probability argument to the latter. In Section 6, we bound $\tau_{2b} - \tau_{2a}$ by approximating the sequence with that obtained by removing the fluctuations. Finally, we will complete the proof of Theorem 2 in Section 7, before concluding and discussing the broader implications of the result in Section 8.

3. Uniform control over r_k

In this section, we will show that the sequence of squared norms, $\{r_k^2\}_k$, quickly converges to a small interval of width $O(\log d/\sqrt{d})$ around the value 1, and thereafter remains within

this interval for at least $Cd\log d$ iterations. In subsequent sections, we will show that this is sufficient time for the $\{s_k\}_k$ sequence also to converge. The reason why such uniform control is necessary is because the formula for the horizontal update (9) taken from a point $\mathbf{y}$ depends on $r(\mathbf{y})$. By showing that $\{r_k^2\}_k$ concentrates uniformly, we thereby also obtain control over the drift and fluctuations of $\{s_k\}_k$, which enables us to do an essentially univariate analysis of this latter sequence.

Lemma 6 (Bound for duration of Phase 1) *There exists universal constants C_1 and C_2 such that $\tau_1 \le C_2 d\cdot(\log d + \log|r_0^2-1|)$ with probability at least $1 - C_2/\log d$.*

Proof We first compute $\mathbb{E}\{r_k^2 - 1\}$. Using (7) and (11), we have

$$\begin{aligned}\mathbb{E}\{r_{k+1}^2 - 1 \,\big|\, \mathcal{F}_k\} &= r_k^2 - 1 + \frac{1}{d}\mathbb{E}\{\alpha_{k+1}(\mathbf{y}^{(k)}) \,\Big|\, \mathcal{F}_k\} \\ &= \left(1 - \frac{1}{d}\right)(r_k^2 - 1).\end{aligned}$$

Iterating this gives

$$\mathbb{E}\{r_k^2 - 1\} = \left(1 - \frac{1}{d}\right)^k (r_0^2 - 1). \tag{18}$$

Hence, whenever

$$k \gtrsim \frac{\log d + \log(|r_0^2 - 1|)}{\log(1 - 1/d)} \approx d\big(\log d + \log|r_0^2 - 1|\big) \tag{19}$$

we have $|\mathbb{E}\{r_k^2 - 1\}| \le \frac{1}{\sqrt{d}}$.

Next, we may obtain a recursive bound for the variance of the iterates using the law of total variance. We have

$$\begin{aligned}\mathrm{Var}\big\{r_{k+1}^2 - 1\big\} &= \mathbb{E}\{\mathrm{Var}\{r_{k+1}^2 - 1 \,|\, \mathcal{F}_k\}\} + \mathrm{Var}\big\{\mathbb{E}\{r_{k+1}^2 - 1 \,|\, \mathcal{F}_k\}\big\} \\ &= \mathbb{E}\{\mathrm{Var}\{\alpha_{k+1}(\mathbf{y}^{(k)}) \,|\, \mathcal{F}_k\}\} + \mathrm{Var}\big\{(1 - 1/d)\big(r_k^2 - 1\big)\big\} \\ &\le \frac{C\mathbb{E}\{r_k^4\}}{d^2} + \left(1 - \frac{1}{d}\right)^2 \mathrm{Var}\big\{r_k^2 - 1\big\}. \end{aligned} \tag{20}$$

We may expand

$$\mathbb{E}\{r_k^4\} = \mathrm{Var}\{r_k^2 - 1\} + \mathbb{E}\{r_k^2\}^2,$$

and apply (18) to get

$$\mathbb{E}\{r_k^2\}^2 = \big(\mathbb{E}\{r_k^2 - 1\} + 1\big)^2 \lesssim \left(1 - \frac{1}{d}\right)^{2k} (r_0^2 - 1)^2 + 1.$$

Plugging these back into (20), we get

$$\mathrm{Var}\big\{r_{k+1}^2 - 1\big\} \le \left(1 - \frac{1}{2d}\right)\mathrm{Var}\big\{r_k^2 - 1\big\} + \frac{C}{d^2}\left(1 - \frac{1}{d}\right)^{2k}(r_0^2 - 1)^2 + \frac{C}{d^2},$$

and solving this recursion gives

$$\mathrm{Var}\{r_k^2 - 1\} \le \frac{C}{d^2} \cdot \left(k \cdot (1 - 1/2d)^k (r_0^2 - 1)^2 + \frac{1 - (1 - 1/2d)^{2k+2}}{1 - (1 - 1/2d)^2} \right).$$

It easy to check that this quantity is bounded by C/d whenever (19) holds. In this case, we may apply Chebyshev's inequality to conclude that

$$r_k^2 - 1 \le \mathbb{E}\{r_k^2 - 1\} + \frac{\mathrm{Var}\{r_k^2 - 1\}^{1/2}}{\delta} \le \frac{1}{\sqrt{d}} + \frac{C}{\delta\sqrt{d}}$$

with probability at least $1 - \delta$. We may also similarly bound $r_k^2 - 1$ from below. Choosing $\delta = C/\log d$ gives us the probability bound we want. ■

After r_k has contracted to a value close to 1, we need to show that it remains close to 1 throughout the time scale needed for the algorithm to converge. Although it is clear from the formula (8) that the increments are subexponential, a naive union bound is not tight enough for our purposes. To overcome this, we make use of a maximal Bernstein inequality.

Lemma 7 (Maximal Bernstein for subexponential martingale differences) *Let $X_1, X_2, \ldots, X_M$ be a martingale difference sequence that is adapted to a filtration $\{\mathcal{G}_t\}$. Suppose there is a constant $K > 0$ such that the following pointwise inequality holds almost surely for any time t and $0 < \lambda \le 1/2K$:*

$$\mathbb{E}\{e^{\lambda X_{t+1}} \,|\, \mathcal{G}_t\} \le e^{\lambda^2 K^2}.$$

Denote $S_t = \sum_{i=1}^t X_i$, for $t = 1, 2 \ldots$. Then for all $\epsilon > 0$, we have the uniform tail bound

$$\mathbb{P}\{\exists t \le M \,:\, |S_t| \ge \epsilon\} \le 2\exp\left(-\frac{1}{4} \left\{ \frac{\epsilon^2}{MK^2} \wedge \frac{\epsilon}{K} \right\} \right). \tag{21}$$

Proof This result is an easy consequence of combining two classical arguments: the supermartingale inequality and an exponential martingale inequality. This argument also appears with more sophistication in the exponential line-crossing method recently developed by Howard et al. (2018). The proof details are deferred to Appendix F. ■

Remark 8 *The tail bound on the right hand side of* (21) *is exactly the Bernstein tail at time M. See Vershynin (2018).*

Lemma 9 (Maximal Bernstein for process with contracting drift) *Let $W_1, W_2, \ldots, W_M$ be a real-valued stochastic process adapted to a filtration $\{\mathcal{G}_t\}$. Suppose that there is some $0 < \rho < 1$ such that for each time t, we have*

$$\mathbb{E}\{W_{t+1} \,|\, \mathcal{G}_t\} = \rho W_t. \tag{22}$$

Furthermore, assume that there is a constant $K > 0$ such that the following pointwise inequality holds almost surely for any time t and $0 < \lambda \le 1/2K$:

$$\mathbb{E}\{\exp(\lambda(W_{t+1} - \rho W_t)) \,\big|\, \mathcal{G}_t\} \le e^{\lambda^2 K^2}. \tag{23}$$

Then for all $\epsilon > 0$, we have the uniform tail bound

$$\mathbb{P}\{\exists t \le M \,:\, |W_t| \ge |W_0| + \epsilon\} \le 2\exp\left(-\frac{1}{4}\left\{\frac{\epsilon^2}{MK^2} \wedge \frac{\epsilon}{K}\right\}\right). \tag{24}$$

Proof For any t, we set $X_t = W_t - \mathbb{E}\{W_t \,|\, \mathcal{G}_t\}$ to obtain the recursive formula:

$$W_{t+1} = \rho W_t + X_{t+1}.$$

Since $X_1, X_2, \ldots$ form a martingale difference sequence satisfying the assumptions of Lemma 7, its partial sums are bounded, and we may apply an easy combinatorial result (Lemma 33) to bound the tails of W_t. ∎

We are now ready to state the guarantee that the radius remains uniformly close to 1.

Lemma 10 (Uniform bounds for r_k) *For any constant C_1, $M \le C_1 d \log d$, we have*

$$\sup_{0 \le k \le M} |r^2_{\tau_1 + k} - 1| \le \frac{C_2 \log d}{\sqrt{d}} \tag{25}$$

with probability at least $1 - \frac{1}{d}$, where C_2 is a universal constant depending only on C_1. Furthermore, for $M \le d^2/\log d$, there is a constant C_3 such that

$$\sup_{0 \le k \le M} r_{\tau_1 + k} \le C_3 \tag{26}$$

with probability at least $1 - 1/d$.

Proof For ease of notation, assume $\tau_1 = 0$. Set $W_k \coloneqq r_k^2 - 1$ for $k = 0, 1, 2, \ldots$. Recall also that $\{\mathcal{F}_k\}_k$ is the filtration generated by the stochastic updates. We want to show that these satisfy the assumptions of Lemma 9. By Lemma 4, we see that (22) is satisfied with $\rho = 1 - \frac{1}{d}$, and we would like to use Lemma 25 to verify (23). However, Lemma 25 is only valid when $\mathbf{y}$ lies in a bounded subset of $\mathcal{Y}$, and we don't assume this a priori.

In order to overcome this, we use a coupling argument. Define an auxiliary seqeuence $\{\tilde{r}_k\}_k$ with the same initialization $\tilde{r}_0 = r_0$, and coupled with r_k until the first time τ such that $|r_\tau - 1| \ge 1/2$. Thereafter, we evolve $\tilde{r}_k$ deterministically via the formula

$$\tilde{r}^2_{k+1} - 1 \coloneqq \left(1 - \frac{1}{d}\right)(\tilde{r}^2_k - 1).$$

In other words, we enforce (22) and (23) by fiat for the sequence $\{\tilde{r}_k^2 - 1\}_{k \ge 0}$. This allows us to invoke Lemma 9 for this sequence with $K = \frac{c}{d}$ for some universal constant $c > 0$. Plugging our choice of M and K into (24), and choosing $\epsilon = \frac{C \log d}{\sqrt{d}}$, we that for C large enough, the right hand side in (24) is bounded above by $\frac{1}{d}$.

As such we have proved that

$$\sup_{0\le k\le M} |\tilde{r}_k^2 - 1| \le \frac{C\log d}{\sqrt{d}}$$

with probability at least $1 - \frac{1}{d}$. Meanwhile, on the same event, we have $\tau > M$, so that the bound also holds for the original sequence, thereby giving us (25). The proof of (26) is similar and is hence omitted. ■

Let us denote $\tau_r := \min\{k\colon |r^2_{\tau_1+k} - 1| \ge C_2 \log d/\sqrt{d}\}$. The previous lemma tells us that $\tau_r \ge C_1 d\log d$ with probability at least $1 - 1/d$. In order to obtain the control over $\{s_k\}_k$ promised at the start of this section, we would ideally like to condition on this event, but doing so would destroy the Markov nature of the process $\mathbf{y}^{(k)}$ and invalidate the update formula (9).

Instead, we will define a sequence $\{\tilde{\mathbf{y}}^{(k)}\}_k$ that we evolve according to

$$\tilde{\mathbf{y}}^{(k+1)} = \begin{cases} \mathbf{y}^{(k)} & \text{if } k \le \tau_1, \\ \Pi\big(\tilde{\mathbf{y}}^{(k)} + \frac{1}{d}\big(\alpha(\tilde{\mathbf{y}}^{(k)}), \beta(\tilde{\mathbf{y}}^{(k)})\big)\big) & \text{if } k > \tau_1. \end{cases} \tag{27}$$

Here, $\Pi\colon \mathcal{Y} \to \mathcal{Y}$ leaves the s component fixed, but projects the r^2 component to the interval $[1 - C_2\log d/\sqrt{d}, 1 + C_2 \log d/\sqrt{d}]$. We also continue to couple $\tilde{\mathbf{y}}^{(k)}$ and $\mathbf{y}^{(k)}$ for $k > \tau_1$ by using the same random seeds for their updates.

To summarize, we have $\mathbf{y}^{(k)} = \tilde{\mathbf{y}}^{(k)}$ for $k \le \tau_r$. Furthermore, the projection step means that the regularity of the horizontal updates for $\tilde{\mathbf{y}}^{(k)}$ is enforced by fiat. For the rest of this paper, we will work with this sequence instead, even going so far as to drop the tildes and use the same notation for both sequences, since the reason for the creation of this alternate sequence is purely technical and has no intuitive value.

4. Phase 2a: Stochastic dominance argument

We have broken up the analysis of "Phase 2" of the SGD process into two sub-phases. The first sub-phase concerns the portion of the process in which $|s_k| = o(1)$, so that the drift and fluctuations are of comparable magnitudes. The goal of this section and the next is to bound the length of this phase by proving the following theorem.

Theorem 11 (Bound for duration of Phase 2a) *There is some $0 < \gamma_1 < 1/2$ and success probability $p > 0$, such if we use this value of γ_1 in (16), for d large enough, $\tau_{2a} - \tau_1 \le Cd\log d$ with probability at least p.*

For ease of notation and making use of the strong Markov property, we may assume again that $\tau_1 = 0$. Our strategy for proving the theorem is to show that for iterations satisfying $\tau_1 \le k \le \tau_{2a}$, a subsequence of the horizontal marginals of the process, $\{s_{Bk}\}_k$, stochastically dominate another process $\{\hat{s}_k\}_k$ in magnitude. The second process will be constructed so that we will have precise control over its second and fourth moments, which will allow us to apply the Paley-Zygmund inequality to get a small ball probability argument.

In this section, we will focus on constructing the comparison process and fleshing out the stochastic dominance argument.

In order to get control over moments, the process $\{\hat{s}_k\}_k$ will be constructed to have approximately Gaussian increments. The comparison will thus be based on normal approximation using the Berry-Essen theory. It may be possible to obtain appropriate Berry-Essen bounds for martingale sequences, but the theory on this topic seems to be incomplete at the time of writing. We shall bypass this by showing that over epochs of an appropriate length B, the update $s_k \to s_{k+B}$ is well-approximated by the update we would get had we done a batch update, summing up B independent steps taken from $\mathbf{y}^{(k)}$. Since this is a sum of i.i.d. random variables, the classical Berry-Eseen bound can then be applied to this latter update.

More precisely, Let P denote the random mapping defined by (3). It will be helpful to use this notation in this section, as we will have to deal with a few different stochastic sequences. Fix the epoch length to be $B = d^{2/3}\log d$. We define a Markov kernel on the state space via the random mapping Q:

$$Q(\mathbf{y}) \coloneqq \mathbf{y} + \frac{1}{d}\sum_{k=1}^{B}(\alpha_k(\mathbf{y}), \beta_k(\mathbf{y})), \tag{28}$$

where $\alpha_k(\mathbf{y})$ and $\beta_k(\mathbf{y})$ for $k = 1, 2, \ldots, B$ are independent realizations of the random variables defined in (8) and (9). Denoting $\sigma(\mathbf{y})^2 \coloneqq \mathrm{Var}\{\beta(\mathbf{y})\}$, we note here that the drift and variance of this batch update satisfy the formulas

$$\mathbb{E}\{s(Q(\mathbf{y}))\} = s(\mathbf{y}) + \frac{B\bar{\beta}(\mathbf{y})}{d}, \qquad \mathrm{Var}\{s(Q(\mathbf{y}))\} = \frac{B\sigma(\mathbf{y})^2}{d^2}.$$

Lemma 12 (Distance moved in one epoch) *Fix $\boldsymbol{y}^{(m)}$ for $\tau_1 \le m \le \tau_1 + \tau_r$. With probability at least $1 - 1/d^2$, we have*

$$|s_{m+t} - s_m| \le \frac{CB}{d}\left(|s_m| + \sqrt{\frac{\log d}{B}}\right)$$

for $0 \le t \le B$.

Proof For ease of notation, re-index so that $m = 0$. First, observe that we have the decomposition

$$\begin{aligned} s_t - s_0 &= \sum_{k=0}^{t-1} s_{k+1} - s_k \\ &= \frac{1}{d}\sum_{k=0}^{t-1}\beta(\mathbf{y}^{(k)}) \\ &= \frac{1}{d}\sum_{k=0}^{t-1}\bar{\beta}(\mathbf{y}^{(k)}) + \frac{1}{d}\sum_{k=0}^{t-1}\Big(\beta_k(\mathbf{y}^{(k)}) - \bar{\beta}(\mathbf{y}^{(k)})\Big). \end{aligned} \tag{29}$$

Applying the maximal Bernstein bound for martingale sequences (Lemma 7) with $M = B$ and $\epsilon = \sqrt{B \log d}$, we get

$$\sup_{0 \le t \le B} \left| \sum_{k=0}^{t-1} \Big(\beta_k(\mathbf{y}^{(k)}) - \bar{\beta}(\mathbf{y}^{(k)}) \Big) \right| \le C\sqrt{B \log d}. \tag{30}$$

with probability at least $1 - 1/d^2$.

By Lemma 5, we have

$$|\bar{\beta}(\mathbf{y}^{(t)})| \le \bar{b}_{max} |s_t|.$$

Plugging (29) into the above equation, and using the bound (30), we get

$$\begin{aligned} |\bar{\beta}(\mathbf{y}^{(t)})| &\le \bar{b}_{max} \left(|s_0| + \frac{1}{d} \sum_{k=0}^{t-1} \left| \bar{\beta}(\mathbf{y}^{(k)}) \right| + \frac{1}{d} \left| \sum_{k=0}^{t-1} \Big(\beta_k(\mathbf{y}^{(k)}) - \bar{\beta}(\mathbf{y}^{(k)}) \Big) \right| \right) \\ &\le \bar{b}_{max} \left(|s_0| + \frac{1}{d} \sum_{k=0}^{t-1} \left| \bar{\beta}(\mathbf{y}^{(k)}) \right| + \frac{C\sqrt{B \log d}}{d} \right). \end{aligned}$$

We can simplify this recursive bound using some combinatorics. Applying Lemma 34 with $\rho = \bar{b}_{max}/d$, $x_t = |\bar{\beta}(\mathbf{y}^{(t)})|$, and $\xi = |s_0| + (C\sqrt{B \log d})/d$, we get

$$\begin{aligned} \frac{1}{d} \sum_{t=0}^{B-1} |\bar{\beta}(\mathbf{y}^{(t)})| &\le \left(|s_0| + \frac{C\sqrt{B \log d}}{d} \right) \left(\left(1 + \frac{\bar{b}_{max}}{d} \right)^B - 1 \right) \\ &\le \frac{1.1 \bar{b}_{max} B}{d} \left(|s_0| + \frac{C\sqrt{B \log d}}{d} \right). \end{aligned}$$

Plugging this back into (29), for any $t \le B$, we get

$$\begin{aligned} |s_t - s_0| &\le \frac{1.1 \bar{b}_{max} B}{d} \left(|s_0| + \frac{C\sqrt{B \log d}}{d} \right) + \frac{C\sqrt{B \log d}}{d} \\ &\le \frac{CB}{d} \left(|s_0| + \sqrt{\frac{\log d}{B}} \right) \end{aligned}$$

as we wanted. ■

Lemma 13 (Approximation error per epoch for Q kernel) *Fix $\boldsymbol{y}^{(m)}$ for $\tau_1 \le m \le \tau_1 + \tau_r$. For d large enough, with probability at least $1 - 1/d^2$, there is a coupling such that we have*

$$\left| s\Big(Q(\boldsymbol{y}^{(m)}) \Big) - s\Big(P^B(\boldsymbol{y}^{(m)}) \Big) \right| \le \frac{CB^2 \log d}{d^2} \left(|s_m| + \sqrt{\frac{\log d}{B}} \right).$$

Proof For ease of notation, we again re-index so that $m = 0$. Couple the random mappings Q and P^B by using the same draws for $u^{(t)}, v^{(t)}$, $t = 1, 2 \ldots$. First condition on the

probability $1 - 1/d^2$ event promised by Lemma 12. Now write

$$\begin{aligned} s\Big(Q(\mathbf{y}^{(0)})\Big) - s\Big(P^B(\mathbf{y}^{(0)})\Big) &= \frac{1}{d}\sum_{t=1}^{B}\Big(\beta_t(\mathbf{y}) - \beta_t(\mathbf{y}^{(t-1)})\Big) \\ &= \frac{1}{d}\sum_{t=1}^{B}(r_t\cos\theta_t - r_0\cos\theta_0)(u^{(t)})^2 + \frac{1}{d}\sum_{t=1}^{B}(r_t\sin\theta_t - r_0\sin\theta_0)u^{(t)}v^{(t)} \\ &\quad - \frac{2}{d}\sum_{t=1}^{B}\mathbf{1}_{\mathcal{A}_t}\cdot(u^{(t)})^2, \end{aligned}$$

where for each t, $\mathcal{A}_t$ is the event that

$$\operatorname{sign}(\cos\theta_t u^{(t)} + \sin\theta_t v^{(t)}) \neq \operatorname{sign}(\cos\theta_0 u^{(t)} + \sin\theta_0 v^{(t)}).$$

We further condition on the probability $1 - 1/d^2$ events in which we have

$$\sup_{1\le t\le B}|u^{(t)}|,\ \sup_{1\le t\le B}|v^{(t)}| \le C\sqrt{\log d}.$$

By Lemma 12, we see that the first term is bounded as follows:

$$\begin{aligned} \left|\frac{1}{d}\sum_{t=1}^{B}(r_t\cos\theta_t - r_0\cos\theta_0)(u^{(t)})^2\right| &\le \frac{B}{d}\sup_{0\le t\le B-1}|s_t - s_0|\cdot\sup_{1\le t\le B}(u^{(t)})^2 \\ &\le \frac{CB^2\log d}{d^2}\left(|s_0| + \sqrt{\frac{\log d}{B}}\right). \end{aligned}$$

The second term may be bounded similarly. To bound the third term, we use Lemma 35. Observe that $\{\mathbf{1}_{\mathcal{A}_t}\}$ is a sequence of Bernoulli variables. If we condition on the high probability event promised by Lemma 12, we also have

$$\mathbb{E}\{\mathbf{1}_{A_t} \mid \mathcal{F}_{k-1}\} \le C|\theta_t - \theta_0| \le \frac{CB}{d}\left(|s_0| + \sqrt{\frac{\log d}{B}}\right). \tag{31}$$

In Lemma 35, set $M = B$, $\epsilon = 1/2$, and θ to be the right hand side of (31). This gives

$$\sum_{t=1}^{B}1_{\mathcal{A}_t} \le \frac{CB^2}{d}\left(|s_0| + \sqrt{\frac{\log d}{B}}\right)$$

with probability at least

$$1 - \exp\Big(-CB^2/d\cdot\sqrt{\log d/B}\Big) \ge 1 - \exp\big(-C\log^2 d\big) \ge 1 - 1/d^2$$

for d large enough. On this event, the third term is therefore also bounded by $\frac{CB^2\log d}{d^2}\left(|s_0| + \sqrt{\frac{\log d}{B}}\right)$.

By adjusting constants if necessary, we can make sure that the total error probability is bounded by $1/d^2$. ∎

In order to compare $\{s_{Bk}\}_k$ with our (as yet undefined) reference sequence $\{\hat{s}_k\}_k$, we will use a more nuanced version of the usual notion of stochastic dominance.

Definition 14 (Stochastic dominance with error) *Given two real-valued random variables X and Y defined on the same probability space, we say that X* dominates *Y in place up to error δ, denoted $X \geq_\delta Y$, if $X \geq Y$ with probability at least $1-\delta$. In addition, given any two real-valued random variables X and Y, we say that X* stochastically dominates *Y up to error δ, if there is a coupling of X and Y for which $X \geq_\delta Y$. Denote this relation by $X \succeq_\delta Y$. Note that the usual notion of stochastic dominance is equivalent to $X \succeq_0 Y$, which we will also denote using $X \succeq Y$.*

We first establish stochastic dominance for the conditional distribution of $|s_{k+B}|$ given $\mathbf{y}^{(k)}$ over that obtained from an approximately Gaussian increment. Following this, we will argue that stochastic dominance of individual steps also implies stochastic dominance for the entire process.

Lemma 15 (One step stochastic dominance) *Define the region*

$$\mathcal{D} \coloneqq \left\{(r^2, s) \in \mathcal{Y} \,:\, |r^2 - 1| \leq \frac{C \log d}{\sqrt{d}}, |s| < \gamma_1\right\}, \tag{32}$$

and let $\boldsymbol{y}$ be any point in $\mathcal{D}$. Define the approximation error term

$$\epsilon(s) \coloneqq \frac{CB^2 \log d}{d^2}\left(|s| \vee \sqrt{\frac{\log d}{B}}\right). \tag{33}$$

For any $a > 0$, we also denote the soft-thresholding operator $\rho_a \colon \mathbb{R} \to \mathbb{R}$ via

$$\rho_a[x] = \operatorname{sign}(x)(|x| - a)_+.$$

Then with $\delta = 1/d^2 + C/\sqrt{B}$, we have

$$s(P^B(\boldsymbol{y}))^2 \succeq_\delta \rho_{\epsilon(s(\boldsymbol{y}))}\left[s(\boldsymbol{y}) + \frac{B\bar{\beta}(\boldsymbol{y})}{d} + \frac{\sqrt{B}\sigma(\boldsymbol{y})g}{d}\right]^2. \tag{34}$$

Proof Fix $\mathbf{y}$, and set $W_B \coloneqq \frac{1}{\sigma(\mathbf{y})\sqrt{B}} \sum_{k=1}^{B} \big(\beta_k(\mathbf{y}) - \bar{\beta}(\mathbf{y})\big)$, where $\sigma(\mathbf{y})^2 \coloneqq \operatorname{Var}\{\beta(\mathbf{y})\}$. The subexponential bound on $\beta(\mathbf{y})$ implies bounds on the 3rd moments, so by Berry-Esseen, its distribution function F_W satisfies the bound

$$\|F_W - \Phi\|_\infty \leq \frac{C}{\sqrt{B}}$$

where Φ is the distribution function of a standard normal random variable, and C is an absolute constant. Recall that

$$s(Q(\mathbf{y})) = s(\mathbf{y}) + \frac{B\bar{\beta}(\mathbf{y})}{d} + \frac{\sqrt{B}\sigma(\mathbf{y})}{d} \cdot W_B.$$

Using Lemma 27, we obtain the bound

$$\|F_{s(Q(\mathbf{y}))^2} - G\|_\infty \leq \frac{2C}{\sqrt{B}},$$

where G is the CDF of

$$\left(s(\mathbf{y}) + \frac{B\bar{\beta}(\mathbf{y})}{d} + \frac{\sqrt{B}\sigma(\mathbf{y})}{d} \cdot g\right)^2.$$

We next apply Lemma 28 and 30 with $\delta = 2C/\sqrt{B}$ to get

$$\rho_{\epsilon(s(\mathbf{y}))}[s(Q(\mathbf{y}))]^2 \succeq_\delta \rho_{\epsilon(s(\mathbf{y}))}\left[s(\mathbf{y}) + B\frac{\bar{\beta}(\mathbf{y})}{d} + \frac{\sqrt{B}\sigma(\mathbf{y})}{d} \cdot g\right]^2. \tag{35}$$

By Lemma 13, we also have

$$s(P^B(\mathbf{y}))^2 \succeq_{\delta'} \rho_{\epsilon(s(\mathbf{y}))}[s(Q(\mathbf{y}))]^2$$

for $\delta' = 1/d^2$. We may combine this with (35) using the transitivity of stochastic dominance (Lemma 29), to get the bound we want in (34). ■

We are now ready to formally define the process $\{\hat{s}_k\}_k$. Let $L\colon \mathbb{R}\times\mathcal{Y} \to \mathbb{R}$ be a transition kernel defined by

$$L(s, \mathbf{y}) \coloneqq \rho_{\epsilon(s)}\left[b(s) + \frac{\sqrt{B}\sigma(\mathbf{y})g}{d}\right],$$

where

$$b(s) \coloneqq (1 + \kappa B/d)s, \tag{36}$$

and $0 < \kappa < 1$ a small constant to be determined later, while $\sigma(\mathbf{y})^2 \coloneqq \mathrm{Var}\{\beta(\mathbf{y})\}$ as before. Again setting $\tau_1 = 0$ for notational convenience, we define

$$\hat{s}_{k+1} \coloneqq L(\hat{s}_k, \mathbf{y}^{(kB\wedge\tau_{2a})}). \tag{37}$$

Lemma 16 (Process-wise stochastic dominance) *Let $\hat{s}_0, \hat{s}_1, \hat{s}_2, \ldots$ be the Markov process defined by the update rule* (37)*. Then for any positive integer k, we have*

$$s^2_{kB\wedge\tau_{2a}} \succeq_{k\delta} \hat{s}^2_{k\wedge\lfloor\tau_{2a}/B\rfloor},$$

where $\delta = 1/d^2 + C/\sqrt{B}$.

Proof For convenience of notation, we define the auxiliary transitional kernel

$$K(s, \mathbf{y}) \coloneqq \rho_{\epsilon(s(\mathbf{y}))}\left[s(\mathbf{y}) + \frac{B\bar{\beta}(\mathbf{y})}{d} + \frac{\sqrt{B}\sigma(\mathbf{y})g}{d}\right].$$

Consider a fixed $\mathbf{y} \in \mathcal{D}$. In Lemma 15, we showed that $s(P^B(\mathbf{y}))^2 \succeq_\delta K(s(\mathbf{y}), \mathbf{y})^2$. If we choose κ small enough in (36), then by the drift lower bound in Lemma 5, replacing the kernel K with L simply reduces the magnitude of the drift while preserving the variance of the Gaussian increment and the magnitude of the soft-thresholding. Using the first part of Lemma 32 thus gives the bound $K(s(\mathbf{y}), \mathbf{y})^2 \succeq L(s(\mathbf{y}), \mathbf{y})^2$. This implies through transitivity (Lemma 29) that

$$s(P^B(\mathbf{y}))^2 \succeq_\delta L(s(\mathbf{y}), \mathbf{y})^2. \tag{38}$$

Treating each epoch of updates as a single step, what we have showed is the stochastic dominance of a single step over a Gaussian increment, conditioned on a fixed starting value for $\mathbf{y}$. It is not so easy, however, to conclude that stochastic dominance is preserved when composing multiple steps together. This is because we are working with transition kernels, and not sums of independent random variables. To overcome this, we need to use the second part of Lemma 32, which tells us that

$$L(s', \mathbf{y})^2 \succeq L(s, \mathbf{y})^2. \tag{39}$$

whenever $s^2 \leq (s')^2$.

Let us now prove the claim by induction. The case $k = 0$ is clear since $\hat{s}_0 = s_0$. Now assume that the statement holds for some k. Let $\mathcal{G}_k$ be the σ-algebra generated by $\mathbf{y}^{(0)}, \dots \mathbf{y}^{(k)}, \hat{s}_0^2, \dots \hat{s}_k^2$. Condition on $\mathcal{G}_k$ as well as the $1 - k\delta$ probability event for which

$$s_{kB \wedge \tau_{2a}}^2 \geq \hat{s}_{k \wedge \lfloor \tau_{2a}/B \rfloor}^2.$$

If $\tau_{2a} \leq kB$, then $s_{(k+1)B \wedge \tau_{2a}} = s_{kB \wedge \tau_{2a}}$ and $\hat{s}_{(k+1) \wedge \lfloor \tau_{2a}/B \rfloor} = \hat{s}_{k \wedge \lfloor \tau_{2a}/B \rfloor}$, so that

$$s_{(k+1)B \wedge \tau_{2a}}^2 \geq \hat{s}_{(k+1) \wedge \lfloor \tau_{2a}/B \rfloor}^2 \tag{40}$$

on the same event. Otherwise, since $\mathbf{y}^{(kB)} \in \mathcal{D}$, we have

$$\hat{s}_{k+1}^2 = L(\hat{s}_k, \mathbf{y}^{(kB)})^2 \preceq L(s_{kB}, \mathbf{y}^{(kB)})^2 \preceq_\delta s\Big(P^B(\mathbf{y}^{(kB)})\Big)^2 = s_{(k+1)B}^2.$$

Here, the first dominance bound follows from (39), while the second follows from (38). This means that we can construct a coupling of the update kernels such that $\hat{s}_{k+1}^2 \leq s_{(k+1)B}^2$ with conditional probability at least $1 - \delta$.

If $\tau_{2a} \geq (k+1)B$, then this immediately implies that (40) holds. On the other hand, if $kB < \tau_{2a} < (k+1)B$, we have $\lfloor \tau_{2a}/B \rfloor = k$ and $s_{kB}^2 < \gamma_1^2$, so that

$$\hat{s}_{(k+1) \wedge \lfloor \tau_{2a}/B \rfloor}^2 = \hat{s}_k^2 \leq s_{kB}^2 < \gamma_1^2 \leq s_{\tau_{2a}}^2 = s_{(k+1)B \wedge \tau_{2a}}^2.$$

As such, the statement also holds for $k + 1$. ■

5. Phase 2a: Small ball probability argument via Paley-Zygmund

Recall that our goal is to obtain a small ball probability bound for s_k^2, for $k \gtrsim d \log d$, thereby proving Theorem 11. By the stochastic dominance argument in the last section, it suffices to obtain such a bound for $\hat{s}_k^2$, with the appropriate time rescaling. This is the plan for this section, and we start by observing the following general recursive bounds for moments of adapted sequences.

Lemma 17 (Recursive moment bounds for adapted sequences) *Let $S_1, S_2, \dots$ be a process adapted to the filtration $\{\mathcal{G}_t\}$. Then we have*

$$\mathbb{E}\{S_{t+1}^2\} = \mathbb{E}\{\mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^2\} + \mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^2\}. \tag{41}$$

$$\mathbb{E}\{S_{t+1}^4\}^{1/2} \leq \mathbb{E}\{\mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}^4\}^{1/2} + 4\mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^4\}^{1/2}. \tag{42}$$

Proof The first equation is standard. To prove the second, we expand and use martingale orthogonality to write

$$\begin{aligned}\mathbb{E}\{S_{t+1}^4\} = \mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^4\} + \mathbb{E}\{\mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}^4\} \quad (43)\\ + 6\mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^2 \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}^2\}\\ + 4\mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^3 \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}\}. \quad (44)\end{aligned}$$

By Cauchy-Schwarz, we have

$$\mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^2 \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}^2\} \le \mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^4\}^{1/2} \cdot \mathbb{E}\{\mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}^4\}^{1/2}.$$

Furthermore, the third term can be bounded as follows:

$$\begin{aligned}\mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^3 \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}\} &= \mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})^2\\ &\qquad \cdot ((S_{t+1} - \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\}) \mathbb{E}\{S_{t+1} \,|\, \mathcal{G}_t\})\}\\ &\le \frac{1}{2}\mathbb{E}\{(S_{t+1} - \mathbb{E}\{S_{t+1} \mid \mathcal{G}_t\})^4\}\\ &\qquad + \frac{1}{2}\mathbb{E}\big\{(S_{t+1} - \mathbb{E}\{S_{t+1} \mid \mathcal{G}_t\})^2 \mathbb{E}\{S_{t+1} \mid \mathcal{G}_t\}^2\big\}.\end{aligned}$$

Plugging these into the original equation, we get

$$\mathbb{E}\{S_{t+1}^4\} \le \Big(\mathbb{E}\{\mathbb{E}\{S_{t+1} \mid \mathcal{G}_t\})^4\}^{1/2} + 4\mathbb{E}\{(S_{t+1} - \mathbb{E}\{X_{t+1} \mid \mathcal{G}_t\})^4\}^{1/2}\Big)^2$$

as we wanted. ■

In order to apply these bounds to our sequence $\{\hat{s}_k\}_k$, we will need estimates of the moments of each increment.

Lemma 18 (Moment bounds for $\hat{s}_k$ increments) *We have the following:*

$$\mathbb{E}\Big\{(\hat{s}_{k+1} - \mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\})^2 \,\Big|\, \mathcal{F}_{kB}\Big\} \ge \frac{C_1 B}{d^2}, \quad (45)$$

$$\mathbb{E}\Big\{(\hat{s}_{k+1} - \mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\})^4 \,\Big|\, \mathcal{F}_{kB}\Big\}^{1/2} \le \frac{C_2 B}{d^2}. \quad (46)$$

Proof In this proof, we shall for convenience denote $X \coloneqq b(\hat{s}_k) + \sqrt{B}\sigma(\mathbf{y}^{(kB)})g/d$ and $\epsilon \coloneqq \epsilon(\hat{s}_k)$. Writing out the definition (37), we then have $\hat{s}_{k+1} = \rho_\epsilon[X]$. Using the definition of X and the fact that $\mathbf{y}^{(kB)} \in \mathcal{D}$, we compute

$$\mathrm{Var}\{X \mid \mathcal{F}_{kB}\} = \mathrm{Var}\Big\{\sqrt{B}\sigma(\mathbf{y}^{(kB)})g/d\Big\} \ge \frac{CB}{d^2}. \quad (47)$$

We need to show that soft-thresholding X does not decrease its variance by too much, and will consider two cases depending on the value of $\hat{s}_k$. We shall suppose WLOG that $\hat{s}_k > 0$.

First, suppose $\hat{s}_k \le \frac{1}{\log^3 d}$. Then

$$\begin{aligned}\mathrm{Var}\{\rho_\epsilon[X] \mid \mathcal{F}_{kB}\}^{1/2} &\ge \mathbb{E}\{|\rho_\epsilon[X] - \mathbb{E}\rho_\epsilon[X]| \mid \mathcal{F}_{kB}\} \\ &\ge \mathbb{E}\{|X - \mathbb{E}X| - |X - \rho_\epsilon[X]| - |\mathbb{E}X - \mathbb{E}\rho_\epsilon[X]| \mid \mathcal{F}_{kB}\} \\ &\ge \mathbb{E}\{|X - \mathbb{E}X| \mid \mathcal{F}_{kB}\} - 2\epsilon \\ &= \sqrt{\frac{2}{\pi}}\mathrm{Var}\{X \mid \mathcal{F}_{kB}\}^{1/2} - 2\epsilon.\end{aligned}$$

Plugging our assumption on $\hat{s}_k$ into (33), we get

$$\epsilon \le \frac{B^2 \log d}{d^2}\frac{1}{\log^3 d} = \frac{1}{d^{2/3}} \ll \frac{\log^{1/2} d}{d^{2/3}} = \frac{\sqrt{B}}{d}.$$

By (47), the quantity on the right hand side is a lower bound for $\mathrm{Var}\{X \mid \mathcal{F}_{kB}\}$, giving us what we want.

Now assume instead that $\hat{s}_k \ge \frac{1}{\log^3 d}$. Let X' be an independent copy of X. Using a well-known formula for variance, we have

$$\begin{aligned}\mathrm{Var}\{\rho_\epsilon[X] \mid \mathcal{F}_{kB}\} &= \frac{1}{2}\mathbb{E}\Big\{\big(\rho_\epsilon[X] - \rho_\epsilon[X']\big)^2 \,\Big|\, \mathcal{F}_{kB}\Big\} \\ &\ge \frac{1}{2}\mathbb{E}\Big\{\big(\rho_\epsilon[X] - \rho_\epsilon[X']\big)^2 \cdot \mathbf{1}\{X, X' \ge \epsilon\} \,\Big|\, \mathcal{F}_{kB}\Big\} \\ &= \frac{1}{2}\mathbb{E}\Big\{\big(X - X'\big)^2 \cdot \mathbf{1}\{X, X' \ge \epsilon\} \,\Big|\, \mathcal{F}_{kB}\Big\} \\ &= \frac{B\sigma(\mathbf{y}^{(kB)})^2}{2d^2} \cdot \mathbb{E}\bigg\{(g - g')^2 \cdot \mathbf{1}\bigg\{g, g' \ge \frac{(-b(\hat{s}_k) + \epsilon(\hat{s}_k))d}{\sqrt{B}\sigma(\mathbf{y}^{(kB)})}\bigg\}\bigg\}, \qquad (48)\end{aligned}$$

where g and g' are independent standard normal random variables.

Observe that

$$\begin{aligned}\frac{(-b(\hat{s}_k) + \epsilon(\hat{s}_k))d}{\sqrt{B}\sigma(\mathbf{y}^{(kB)})} &= \frac{\big(-(1 + \kappa B/d) + B^2 \log d/d^2\big)d\hat{s}_k}{\sqrt{B}\sigma(\mathbf{y}^{(kB)})} \\ &\le -\frac{\hat{s}_k d}{\sqrt{B}\sigma(\mathbf{y}^{(kB)})} \\ &\lesssim -\frac{d}{\sqrt{B}\log^3 d},\end{aligned}$$

which tends to $-\infty$, so the expectation on the right hand side of (48) tends to $\mathbb{E}\{(g-g')^2\} = 2$. As such, the entire quantity on the right hand side is bounded from below by CB/d^2.

Next, to obtain (46), we simply apply Lemma 36 to remove the soft-thresholding:

$$\begin{aligned}&\mathbb{E}\Big\{(\hat{s}_{k+1} - \mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\})^4 \,\Big|\, \mathcal{F}_{kB}\Big\} \\ &\quad\le 8\mathbb{E}\bigg\{\Big(b(\hat{s}_k) + \sqrt{B}\sigma(\mathbf{y}^{(kB)})g/d - \mathbb{E}\{b(\hat{s}_k) + \sqrt{B}\sigma(\mathbf{y}^{(kB)})/d \,\Big|\, \mathcal{F}_{kB}\}\Big)^4 \Big| \mathcal{F}_{kB}\bigg\} \\ &\quad= \frac{8B^2\sigma(\mathbf{y}^{(kB)})^4\mathbb{E}\{g^4\}}{d^4}.\end{aligned}$$

■

Lemma 19 (Recursive moment bounds for $\hat{s}_k$ process) *We have the recursive formulas*

$$\mathbb{E}\{\hat{s}_{k+1}^2\} \geq \left(1 + \frac{\kappa B}{d}\left(1 - \frac{C_1}{\log^2 d}\right)\right)^2 \mathbb{E}\{\hat{s}_k^2\} + \frac{C_2 B}{d^2} \tag{49}$$

$$\mathbb{E}\{\hat{s}_{k+1}^4\}^{1/2} \leq \left(1 + \frac{\kappa B}{d}\right)^2 \mathbb{E}\{\hat{s}_k^4\}^{1/2} + \frac{C_3 B}{d^2}. \tag{50}$$

Proof The second inequality is easier to prove, so we shall start with this. We may use Lemma 36 to get

$$\begin{aligned}
|\mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\}| &= \left|\mathbb{E}\left\{\rho_{\epsilon(\hat{s}_k)}[b(\hat{s}_k) + \sqrt{B}\sigma(\mathbf{y}^{(kB)})g] \;\middle|\; \mathcal{F}_{kB}\right\}\right| \\
&\leq |b(\hat{s}_k)| \\
&= \left(1 + \frac{\kappa B}{d}\right)|\hat{s}_k|.
\end{aligned}$$

Plugging this bound together with (46) into (42) gives us (50).

Next, denoting $X \coloneqq b(\hat{s}_k) + \sqrt{B}\sigma(\mathbf{y}^{(kB)})g$ and $\epsilon \coloneqq \epsilon(\hat{s}_k)$ as in the previous lemma, observe that

$$\begin{aligned}
|\mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\}| &= |\mathbb{E}\{\rho_\epsilon[X] \mid \mathcal{F}_{kB}\}| \\
&\geq \left(|\mathbb{E}\{X \mid \mathcal{F}_{kB}\}| - |\mathbb{E}\{\rho_\epsilon[X] - X \mid \mathcal{F}_{kB}\}|\right)_+ \\
&\geq \left(\left(1 + \frac{\kappa B}{d}\right)|\hat{s}_k| - \epsilon(\hat{s}_k)\right)_+.
\end{aligned} \tag{51}$$

It remains to compare the relative magnitudes of the two terms. For convenience, we reproduce the definition of $\epsilon(\hat{s}_k)$ here:

$$\epsilon(\hat{s}_k) \coloneqq \frac{CB^2 \log d}{d^2}\left(|\hat{s}_k| \vee \sqrt{\frac{\log d}{B}}\right).$$

We will divide the proof into three cases, depending on the magnitude of $\hat{s}_k$. WLOG, assume that $\hat{s}_k \geq 0$.

When $\hat{s}_k \geq \frac{1}{d^{1/3}}$, then

$$\epsilon(\hat{s}_k) = \frac{CB^2 \log d}{d^2}\hat{s}_k \leq \frac{1}{\log^2 d} \cdot \frac{\kappa B}{d}\hat{s}_k$$

for d large enough, and when $\frac{\log^4 d}{d^{2/3}} \leq \hat{s}_k \leq \frac{1}{d^{1/3}}$, one may check that

$$\epsilon(\hat{s}_k) = \frac{CB^2 \log d}{d^2}\sqrt{\frac{\log d}{B}} \leq \frac{C}{\log^2 d} \cdot \frac{\kappa B}{d}\hat{s}_k.$$

In either case, we may plug the bound for $\epsilon(\hat{s}_k)$ into (51) to get

$$(\mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\})^2 \geq \left(1 + \frac{\kappa B}{d}\left(1 - \frac{C}{\log^2 d}\right)\right)^2 \hat{s}_k^2.$$

Next, when $0 \leq \hat{s}_k \leq \frac{\log^4 d}{d^{2/3}}$, we have

$$\frac{CB^2 \log d}{d^2}\left(\hat{s}_k \vee \sqrt{\frac{\log d}{B}}\right) \leq \frac{C \log^3 d}{d},$$

so that we have

$$\left(\left(1 + \frac{\kappa B}{d}\right)\hat{s}_k - \frac{CB^2 \log d}{d^2}\left(\hat{s}_k \vee \frac{\log d}{\sqrt{B}}\right)\right)_+^2 \geq \left(1 + \frac{\kappa B}{d}\right)^2 \hat{s}_k^2 - \frac{C \log^3 d}{d}\hat{s}_k.$$

Furthermore, we can bound the second term on the right hand side via

$$\frac{\log^3 d}{d}\hat{s}_k \leq \frac{\log^7 d}{d^{5/3}} \lesssim \frac{B}{d^2 \log d}.$$

As such, we may write

$$\mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\}^2 \geq \left(1 + \frac{\kappa B}{d}\right)^2 \hat{s}_k^2 - \frac{cB}{d^2}\left(\frac{1}{\log d}\right).$$

We may thus take expectations to get

$$\mathbb{E}\{\mathbb{E}\{\hat{s}_{k+1} \mid \mathcal{F}_{kB}\})^2\} \geq \left(1 + \frac{\kappa B}{d}\left(1 - \frac{1}{\log^2 d}\right)\right)^2 \mathbb{E}\{\hat{s}_k^2\} - \frac{cB}{d^2}\left(\frac{1}{\log d}\right).$$

Combining this with (41) and (45) gives (49). ■

Proof (of Theorem 11) For convenience, denote $A = 1 - 1/\log^2 d$. We solve the recursion in (49) to get

$$\mathbb{E}\{\hat{s}_k^2\} \geq \frac{(1 + \kappa AB/d)^{2k+2} - 1}{(1 + \kappa AB/d)^2 - 1} \cdot \frac{C_2 B}{d^2} + \left(1 + \frac{\kappa AB}{d}\right)^{2k} \hat{s}_0^2. \tag{52}$$

Note that

$$\frac{1}{(1 + \kappa AB/d)^2 - 1} \cdot \frac{C_2 B}{d^2} \asymp \frac{d}{\kappa AB} \cdot \frac{B}{d^2} \asymp \frac{1}{\kappa d},$$

so that we can simplify (52) to

$$\mathbb{E}\{\hat{s}_k^2\} \gtrsim \left(\left(1 + \frac{\kappa AB}{d}\right)^{2k+2} - 1\right)\frac{C}{\kappa d} + \left(1 + \frac{\kappa AB}{d}\right)^{2k} \hat{s}_0^2. \tag{53}$$

Meanwhile, for any $T \asymp \frac{Cd}{B} \log d \asymp \frac{\log d}{\log(1 + \kappa AB/d)}$, we have

$$\left(1 + \frac{\kappa AB}{d}\right)^{2T+2} = \exp(C \log d). \tag{54}$$

Putting these together, we see that regardless of the value of $\hat{s}_0$, there is an absolute constant γ such that $\mathbb{E}\{\hat{s}_T^2\} \geq \gamma$ if we choose the constant C in the definition of T to be large enough.

On the other hand, we may solve the second recursion (50) and apply a similar computation as before to get

$$\mathbb{E}\{\hat{s}_k^4\}^{1/2} \lesssim \left(\left(1+\frac{\kappa B}{d}\right)^{2k+2} - 1\right)\frac{C}{\kappa d} + \left(1+\frac{\kappa B}{d}\right)^{2k}\hat{s}_0^2. \tag{55}$$

Taking the ratio of (53) and (55), plugging in the time point $k = T$, we get

$$\begin{aligned}
\frac{\mathbb{E}\{\hat{s}_T^2\}}{\mathbb{E}\{\hat{s}_T^4\}^{1/2}} &\gtrsim \frac{\Big((1+\kappa AB/d)^{2T+2} - 1\Big)/\kappa d + (1+\kappa AB/d)^{2T}\hat{s}_0^2}{\Big((1+\kappa B/d)^{2T+2} - 1\Big)/\kappa d + (1+\kappa B/d)^{2T}\hat{s}_0^2} \\
&= \frac{(1+\kappa AB/d)^{2T}}{(1+\kappa B/d)^{2T}} \cdot \frac{(1+\kappa AB/d)^2 + \kappa d\hat{s}_0^2 - (1+\kappa AB/d)^{-2T}}{(1+\kappa B/d)^2 + \kappa d\hat{s}_0^2 - (1+\kappa B/d)^{-2T}} \\
&\asymp \left(\frac{1+\kappa AB/d}{1+\kappa B/d}\right)^{2T}.
\end{aligned}$$

Note that in the last equation, we used the definition of A and (54).

Since

$$\frac{1+\kappa AB/d}{1+\kappa B/d} = \frac{d+\kappa B - \kappa B/\log^2 d}{d+\kappa B} = 1 - \frac{\kappa B}{(d+\kappa B)\log^2 d},$$

we may take logs and observe

$$\begin{aligned}
\log\left(\frac{1+\kappa AB/d}{1+\kappa B/d}\right)^{2T} &\asymp \frac{d}{B}\log d \cdot \log\left(1 - \frac{\kappa B}{(d+\kappa B)\log^2 d}\right) \\
&\asymp -\frac{d}{B}\log d \cdot \frac{B}{d\log^2 d} \\
&= -\frac{1}{\log d}.
\end{aligned}$$

Putting everything together, we have $\frac{\mathbb{E}\{\hat{s}_T^2\}}{\mathbb{E}\{\hat{s}_T^4\}^{1/2}} \gtrsim 1$. As such, for d large enough, we may apply the Paley-Zygmund inequality to see that we may pick some $0 < \gamma_1 < \gamma$ such that $|\hat{s}_T| \geq \gamma_1$, with probability at least some constant $p' > 0$. Let us now condition on the intersection of this event and that promised to us by Lemma 16 when we choose $k = T$. The probability of the intersection is at least

$$p' - T\cdot\left(\frac{1}{d^2} + \frac{C}{\sqrt{B}}\right) = p' - \frac{C\log d}{Bd} - \frac{Cd\log d}{B^{3/2}} \geq p - \frac{C}{\sqrt{\log d}} =: p$$

where the inequality holds for d large enough. On this event, we have $\tau_{2a} \leq TB \leq Cd\log d$ as we wanted. ∎

6. Phase 2b: Approximation by drift process

In the previous two sections, we have bounded the duration of Phase 2a of the SGD process, that is, the time it takes for $|s_k|$ to increase to a constant value. The goal of this section is to bound the duration of Phase 2b in which the iterates converge to the "basin of convexity" around $\mathbf{y}^*$ or $-\mathbf{y}^*$. We will prove the following theorem.

Theorem 20 (Bound for duration of Phase 2b) *We have $\tau_{2b} - \tau_{2a} \le Cd$ with probability at least $1 - 1/d$.*

As mentioned in the overall proof outline, the idea is to understand the trajectory of the drift process, and then show that the fluctuations do not affect the trajectory by too much. For convenience, we condition on the event $\tau_{2a} < \infty$, and then use the strong Markov property to re-index, setting $\tau_{2a} = 0$.

The drift process $\{\bar{\mathbf{y}}^{(k)}\}_k$ is defined via the deterministic update

$$\bar{\mathbf{y}}^{(k+1)} \coloneqq \bar{\mathbf{y}}^{(k)} + \frac{1}{d}(\bar{\alpha}(\bar{\mathbf{y}}^{(k)}), \bar{\beta}(\bar{\mathbf{y}}^{(k)})),$$

with the initialization $\bar{\mathbf{y}}^{(0)} = \mathbf{y}^{(0)}$. For simplicity, we denote $\bar{r}_k \coloneqq r(\bar{\mathbf{y}}^{(k)})$ and $\bar{s}_k \coloneqq s(\bar{\mathbf{y}}^{(k)})$.

Lemma 21 (Behavior of drift sequence) *Let $\tau = \min\{k : \bar{s}_k \ge 1 - (\gamma_2 - \epsilon)/2\}$ for some small $\epsilon > 0$. Then for d large enough, we have $\tau \le Cd$, where C is a universal constant depending only on γ_1, γ_2, and ϵ.*

Proof First, choose ϵ in Lemma 5 to be equal to $\gamma_2/4$. Let d be large enough so that $|r_0^2 - 1| \le C\frac{\log d}{\sqrt{d}} \le \eta$, where $\eta = \eta(\epsilon)$ is the required value in Lemma 5. By Lemma 4, we see that $|\bar{r}_k^2 - 1| \le \eta$ for all $k \le \tau$. This allows us to use Lemma 4 to observe that the recursive inequality $\bar{s}_{k+1} \ge (1 + c/d)\bar{s}_k$ applies whenever $k \le \tau$, where c is a constant only depending on γ_2. By the definition of τ, we have

$$(1 + c/d)^{\tau - 1}\gamma_1 < 1 - (\gamma_2 - \epsilon)/2,$$

which we can solve to get

$$\tau \le \frac{\log((2 - \gamma_2 + \epsilon)/2\gamma_1)}{\log(1 + c/d)} + 1 \le Cd.$$

■

Lemma 22 (Approximation error) *We have*

$$|s_\tau - \bar{s}_\tau| \le \frac{C \log d}{\sqrt{d}} \tag{56}$$

with probability at least $1 - 2/d$.

Proof Recall that by the discussion at the end of Section 3, we have

$$\sup_{0\le k\le\tau}|r_k^2-1|\le\frac{C\log d}{\sqrt{d}}. \tag{57}$$

Using Lemma 7, there is also a probability $1-1/d$ event over which we have

$$\sup_{0\le t\le\tau}\left|\sum_{k=0}^{t-1}\Big(\beta_k(\mathbf{y}^{(k)})-\bar\beta(\mathbf{y}^{(k)})\Big)\right|\le C\sqrt{d}\log d. \tag{58}$$

We will show that (56) holds when conditioned on both of these events. First, for any $t\le\tau$, we may write

$$\bar s_t-\bar s_0=\sum_{k=0}^{t-1}(\bar s_{k+1}-\bar s_k)=\frac{1}{d}\sum_{k=0}^{t-1}\bar\beta(\bar{\mathbf{y}}^{(k)}),$$

and similarly we have (29), which we reproduce here:

$$s_t-s_0=\frac{1}{d}\sum_{k=0}^{t-1}\bar\beta(\mathbf{y}^{(k)})+\frac{1}{d}\sum_{k=0}^{t-1}\Big(\beta_k(\mathbf{y}^{(k)})-\bar\beta(\mathbf{y}^{(k)})\Big).$$

Subtracting these two equations, and recalling that $\bar s_0=s_0$, we get

$$\begin{aligned}|s_t-\bar s_t|&\le\frac{1}{d}\left|\sum_{k=0}^{t-1}\Big(\beta_k(\mathbf{y}^{(k)})-\bar\beta(\mathbf{y}^{(k)})\Big)\right|+\frac{1}{d}\left|\sum_{k=0}^{t-1}\Big(\bar\beta(\mathbf{y}^{(k)})-\bar\beta(\bar{\mathbf{y}}^{(k)})\Big)\right|\\&\le\frac{C\log d}{\sqrt{d}}+\frac{1}{d}\sum_{k=0}^{t-1}\Big|\bar\beta(\mathbf{y}^{(k)})-\bar\beta(\bar{\mathbf{y}}^{(k)})\Big|.\end{aligned} \tag{59}$$

where the bound for the first term on the right hand side comes from (58)

Meanwhile, by Lemma 26, we also have

$$\Big|\bar\beta(\mathbf{y}^{(t)})-\bar\beta(\bar{\mathbf{y}}^{(t)})\Big|\le L(|r_t-\bar r_t|+|s_t-\bar s_t|) \tag{60}$$

and plugging in (59) and (57) gives

$$\Big|\bar\beta(\mathbf{y}^{(t)})-\bar\beta(\bar{\mathbf{y}}^{(t)})\Big|\le L\left(\frac{C\log d}{\sqrt{d}}+\frac{1}{d}\sum_{k=0}^{t-1}\Big|\bar\beta(\mathbf{y}^{(k)})-\bar\beta(\bar{\mathbf{y}}^{(k)})\Big|\right).$$

We are now in a position to apply Lemma 34 with $x_t=\frac{1}{d}\big|\bar\beta(\mathbf{y}^{(t)})-\bar\beta(\bar{\mathbf{y}}^{(t)})\big|$, $\rho=L/d$, and $\xi=C\log d/\sqrt{d}$. Doing so, we get

$$\begin{aligned}\frac{1}{d}\sum_{k=0}^{\tau-1}\Big|\bar\beta(\mathbf{y}^{(k)})-\bar\beta(\bar{\mathbf{y}}^{(k)})\Big|&\le\frac{C\log d}{\sqrt{d}}\cdot((1+L/d)^\tau-1)\\&\le\frac{C\log d}{\sqrt{d}}\end{aligned}$$

as we wanted. ■

Proof (of Theorem 20) Set $\epsilon = \gamma_2/2$. Combining the previous two lemmas, we have

$$\begin{aligned} |s_\tau| &\geq |\bar{s}_\tau| - |s_\tau - \bar{s}_\tau| \\ &\geq 1 - \frac{\gamma_2 - \epsilon}{2} - \frac{C \log d}{\sqrt{d}}, \end{aligned}$$

and combined with (57) gives

$$\begin{aligned} \Psi(\mathbf{y}^{(\tau)}) &= r_\tau^2 - 2|s_\tau| + 1 \\ &= r_\tau^2 - 1 + 2(1 - |s_\tau|) \\ &\leq \frac{C \log d}{\sqrt{d}} + \gamma_2 - \epsilon \\ &\leq \gamma_2, \end{aligned}$$

where the last inequality holds for d large enough. As such, we have $\tau_{2b} \leq \tau \leq Cd$. ■

7. Linear convergence in Phase 3 and proof of Theorem 2

Proof (of Theorem 2) To summarize, we have showed that $\tau_1 \leq Cd \cdot (\log d + \log|r_0^2 - 1|)$ with probability at least $1 - C/\log d$ (Lemma 6), $\tau_{2a} - \tau_1 \leq Cd \log d$ with probability at least p (Theorem 11), and $\tau_{2b} - \tau_{2a} \leq Cd$ with probability at least $1 - 1/d$ (Lemma 20). Putting all of these together gives

$$\tau_{2b} \leq Cd \log\Big(d \cdot \|\mathbf{x}^{(0)}\| \vee 1\Big)$$

with probability at least $p - C/\log d - C/d$, which is larger than $p/2$ for d large enough. Unfortunately, this is not good enough for our purposes and we need to do a bit more work to bring down the error probability.

Let us condition on the $1 - 1/d$ probability event for which (26) holds so that we have uniform control over $\{r_k^2\}$ over an appropriate timescale (more precisely, we use the coupling argument explained in the discussion after Lemma 10). Define $A \coloneqq Cd \log(d \cdot C_3)$, where C_3 is the same constant used in (26). Then by the strong Markov property, for any $k_0 > 0$, we have

$$\begin{aligned} \mathbb{P}\{\tau_{2b} - \tau_1 \geq A + k_0\} &= \mathbb{P}\Big\{\inf_{0 \leq k < k_0 + A} \Psi(\mathbf{y}^{(\tau_1 + k)}) > \gamma_2\Big\} \\ &= \mathbb{E}\Big\{\mathbb{P}\Big\{\inf_{k_0 \leq k < k_0 + A} \Psi(\mathbf{y}^{(\tau_1 + k)}) > \gamma_2 \,\Big|\, \mathcal{F}_{\tau_1 + k_0}\Big\} \mathbf{1}\Big\{\inf_{0 \leq k < k_0} \Psi(\mathbf{y}^{(\tau_1 + k)}) > \gamma_2\Big\}\Big\} \\ &\leq (1 - p/2) \cdot \mathbb{P}\{\tau_{2b} - \tau_1 \geq k_0\}, \end{aligned}$$

which implies that

$$\mathbb{P}\{\tau_{2b} - \tau_1 \geq A + k_0 \,|\, \tau_{2b} - \tau_1 \geq k_0\} \leq 1 - p/2.$$

As such, we have

$$\begin{aligned}\mathbb{P}\{\tau_{2b} - \tau_1 \geq tA\} &= \prod_{k=0}^{t-1} \mathbb{P}\{\tau_{2b} - \tau_1 \geq (k+1)A \,|\, \tau_{2b} - \tau_1 \geq kA\} \\ &\leq (1 - p/2)^t.\end{aligned}$$

If we set $t = \frac{\log(10)}{\log(1-p/2)}$, we see that with probability at least $0.9 - C/\log d - 1/d$,

$$\begin{aligned}\tau_{2b} &\leq tA + \tau_1 \\ &\lesssim d\log d + d\log\Big(d \cdot \|\mathbf{x}^{(0)}\| \vee 1\Big) \\ &\leq d \cdot \Big(\log d + \log\Big(\|\mathbf{x}^{(0)}\| \vee 1\Big)\Big).\end{aligned}$$

As mentioned earlier, we will take $T = \tau_{2b}$.

It remains to establish linear convergence of the process to $\pm\mathbf{y}^*$ during the iterations following T. To do this, we will follow the ideas of Section 3 in Tan and Vershynin (2018). First, we choose γ_2 in (17) to be equal to $\pi^2/320$, and define a stopping time $\tau := \min\{k \geq \tau_{2b} : \Psi(\mathbf{y}^{(k)}) \geq \pi^2/16\}$. The stopping time argument in Tan and Vershynin (2018) then implies that $\mathbb{P}\{\tau = \infty\} \geq 0.95$. Also, we may use Lemma 2.2 therein to guarantee conditional contraction in each step. Namely, if $\Psi(\mathbf{y}^{(k)}) \leq \pi^2/16$ for any k, then

$$\mathbb{E}\Big\{\Psi(\mathbf{y}^{(k+1)}) \,\Big|\, \mathcal{F}_k\Big\} \leq \left(1 - \frac{1}{2d}\right)\Psi(\mathbf{y}^{(k)}).$$

If we write $Y_k := (1 - 1/2d)^{-k}\Psi(\mathbf{y}^{(T+k)}) \cdot \mathbf{1}_{\tau > T+k}$, then the above bound allows us to compute

$$\begin{aligned}\mathbb{E}\{Y_{k+1} \,|\, \mathcal{F}_{T+k}\} &= \left(1 - \frac{1}{2d}\right)^{-k-1} \mathbb{E}\Big\{\Psi(\mathbf{y}^{(T+k+1)}) \,\Big|\, \mathcal{F}_{T+k}\Big\}\mathbf{1}_{\tau > T+k} \\ &\leq \left(1 - \frac{1}{2d}\right)^{-k-1} \Big(1 - \frac{c}{d}\Big)\Psi(\mathbf{y}^{(T+k)})\mathbf{1}_{\tau > T+k} \\ &\leq \left(1 - \frac{1}{2d}\right)^{-k} \Psi(\mathbf{y}^{(T+k)})\mathbf{1}_{\tau > T+k-1} \\ &= Y_k,\end{aligned}$$

and we see that $\{Y_k\}_k$ is a supermartingale with respect to the filtration $\{\mathcal{F}_{T+k}\}_k$.

By the supermartingale inequality, there is a probability 0.95 event over which

$$\sup_{k \geq 0} Y_k \leq 20\Psi(\mathbf{y}^{(\tau_{2b})}) \leq 20\gamma_2.$$

On the intersection of this event, and that on which $\tau = \infty$, we have that

$$\Psi(\mathbf{y}^{(T+k)}) \leq 20\gamma_2\left(1 - \frac{1}{2d}\right)^k \leq \left(1 - \frac{1}{2d}\right)^k$$

for all $k \geq 0$ as we wanted. If we total the measure of the excluded bad events, the final success probability is at least $0.8 - C/\log d - 1/d$ as promised. ■

We now discuss some straightforward extensions of this result. First, the lack of a high probability guarantee is a little unfortunate, and results from having to apply the supermartingale inequality. We are unsure whether this can be overcome theoretically, but we can easily modify the algorithm so that convergence holds with high probability. To do this, we use the "majority vote" procedure described in Tan and Vershynin (2018). The price we have to pay is an additional $\log(1/\delta)$ factor on the number of iterations, where δ is the total error probability we can tolerate.

Second, in the algorithm we presented, we required fresh samples to be used in every update step. Once we are in the linear convergence regime, however, samples can actually be reused so long as we choose uniformly from $\Omega(d)$ of them (Tan and Vershynin, 2018).

8. Conclusion and discussion

In this paper, we have analyzed the convergence of constant step-size stochastic gradient descent for the non-convex, non-smooth phase retrieval objective (1), for which we assume Gaussian sampling vectors, and use an arbitrary initialization. The main idea was to view the SGD sequence as a Markov chain on a summary state space, and then use the natural $1/d$ step size scaling to argue that as d tends to infinity, the process trajectory converges to something we understand. We believe that our proof framework and techniques will have applications beyond the vanilla phase retrieval model, and indeed inform the theory of non-convex optimization in general.

8.1 Extensions of the analysis

Noisy measurements. We have analyzed phase retrieval in the noiseless setting, as is customary in the literature. It is easy to see that the arguments still go through for an additive noise model, except that we may now need to either use batch updates, or reduce the step size further, in order to get convergence. On the other hand, just like in noisy linear regression, we can no longer hope for exact recovery.

Complex measurements. While optimization theory traditionally works with real-valued objects, the physical experiments that inspired the mathematical phase retrieval problem deal with complex measurements. Extending our analysis to this setting is thus important for practical relevance. To do this, we would first need to define gradient descent for functions of complex variables. This can be done naively by decomposing a vector in $\mathbb{C}^d$ into real and imaginary parts, thereby viewing it as a vector in $\mathbb{R}^{2d}$. However, Candès et al. (2015) has argued that it is more convenient to make use of Wirtinger derivatives, which rewrite the objective function F to be a function $\tilde{F}$ on $\mathbb{C}^d \times \mathbb{C}^d$, with $F(\mathbf{z}) = \tilde{F}(\mathbf{z}, \bar{\mathbf{z}})$. $\tilde{F}$ is now holomorphic as a function of $\mathbf{z}$, and so we can directly take complex derivatives, with many of the original formulas from real phase retrieval carrying over formally.[2] The natural

2. Candès et al. (2015) work with the alternate loss function (1), but the arguments also apply to our loss function (2) which is differentiable almost everywhere.

extension of our sampling scheme to the complex setting is to assume independent complex Gaussian sampling vectors. Under this setting, we believe that it is still possible to show that we have a two-dimensional *real* state space, again comprising r^2, the squared norm of the iterate, and s, the absolute value of its projection onto the signal direction. We expect most of the old calculations going through, but leave this to future work.

Non-Gaussian sampling vectors. Our state space argument relies on the rotational symmetry of the sampling vectors, and this clearly will not hold in the non-Gaussian setting. Nonetheless, we conjecture that the phenomenon of SGD convergence from a random initialization should hold more generally for sub-Gaussian sampling vectors, which are straightforward generalizations of the Gaussian distribution. A reasonable approach would be to consider central limits for projections of sub-Gaussian distributions, which may add an additional benign approximation to our argument.

Other single index models. Phase retrieval is an example of a single index model with the link function $f(t) = |t|$. One can easily check that the state space argument generalizes to models with other link functions. Less clear is how to generalize the other arguments in the paper. Nonetheless, we expect this to be not too difficult, so long as we assume some natural regularity conditions on the link function. Finally, we conjecture that similar ideas can work for analyzing SGD for low-rank matrix sensing, since heuristically what makes everything work is the underlying low-dimensional structure in the problem.

8.2 Implications for non-convex optimization

In the introduction to this paper, we have already talked about the growing interest in a theoretical understanding of first-order methods applied to non-convex problems. Here, we reiterate that the main contribution of our paper should be seen as not simply providing a convergence guarantee for SGD, but also one that has close to optimal sample and computational complexity. We are able to achieve this through a careful analysis of the "essential dynamics" of the SGD process, as represented by the summary state space.

Standard proofs of convergence for first order methods applied to convex problems proceed by tracking one of the following three quantities: $\|\mathbf{x}_k - \mathbf{x}^*\|^2$, $f(\mathbf{x}_k) - f(\mathbf{x}^*)$, or $\|\nabla f(\mathbf{x}^*)\|$. Under our framework, this can be seen as implicitly using a one-dimensional state space. In non-convex optimization, however, it makes sense to use a multi-dimensional state space, whereby we measure "progress" in terms of multiple quantities. The number of such quantities one needs to track for a given problem can then perhaps be used to define a notion of "complexity" for that problem.

Acknowledgements

Y.T. was partially supported by NSF CCF-1740855, and completed part of this manuscript while visiting the Simons Institute for the Theory of Computing. R.V. was partially supported by U.S. Air Force Grant FA9550-18-1-0031. Y.T. would like to thank Xiang Cheng, Jelena Diakonikolas, Michael Jordan, and Yian Ma for helpful discussions. We would also like to thank the AE and the two anonymous reviewers for their helpful comments.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. *arXiv preprint arXiv:1811.03962*, nov 2018. URL http://arxiv.org/abs/1811.03962.

Sohail Bahmani and Justin Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2017. ISSN 1938-7228. URL http://arxiv.org/abs/1610.04210.

Tamir Bendory, Robert Beinert, and Yonina C. Eldar. Fourier Phase Retrieval: Uniqueness and Algorithms. *Applied and Numerical Harmonic Analysis*, pages 55–91, 2017. ISSN 22965017. doi: 10.1007/978-3-319-69802-1_2.

Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013. ISSN 00103640. doi: 10.1002/cpa.21432.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. ISSN 00189448. doi: 10.1109/TIT.2015.2399924. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7029630{%}5Cnpapers3://publication/doi/10.1109/TIT.2015.2399924.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval. *arXiv preprint arXiv:1803.07726*, mar 2018. URL http://arxiv.org/abs/1803.07726.

Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The Nonsmooth Landscape of Phase Retrieval. *arXiv preprint arXiv:1711.03247*, 2017. URL http://arxiv.org/abs/1711.03247.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *arXiv preprint arXiv:1707.06386*, 2017. ISSN 07300301. doi: 10.1145/3130800.3130801. URL http://arxiv.org/abs/1707.06386.

John Duchi and Feng Ruan. Solving (Most) of a Set of Quadratic Equalities: Composite Optimization for Robust Phase Retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 09 2018. ISSN 2049-8772. doi: 10.1093/imaiai/iay015. URL https://doi.org/10.1093/imaiai/iay015.

Yonina C. Eldar and Shahar Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014. ISSN 1096603X. doi: 10.1016/j.acha.2013.08.003. URL http://dx.doi.org/10.1016/j.acha.2013.08.003.

James R Fienup. Phase Retrieval Algorithms: A Comparison. *Appl. Opt.*, 21(15):2758, 1982.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points — online stochastic gradient for tensor decomposition. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 797–842, Paris, France, 03–06 Jul 2015. PMLR. URL http://proceedings.mlr.press/v40/Ge15.html.

Tom Goldstein and Christoph Studer. PhaseMax: Convex Phase Retrieval via Basis Pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, Apr 2018. ISSN 0018-9448. doi: 10.1109/TIT.2018.2800768. URL http://arxiv.org/abs/1610.07531http://ieeexplore.ieee.org/document/8278279/.

Paul Hand and Vladislav Voroninski. An Elementary Proof of Convex Phase Retrieval in the Natural Parameter Space via the Linear Program PhaseMax. *arXiv preprint arXiv:1611.03935*, nov 2016. URL http://arxiv.org/abs/1611.03935.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Exponential Line-crossing Inequalities. *arXiv preprint arXiv:1808.03204*, 5(1975), 2018. URL http://arxiv.org/abs/1808.03204.

Halyun Jeong and C. Sinan Güntürk. Convergence of the Randomized Kaczmarz Method for Phase Retrieval. *arXiv preprint arXiv:1706.10291*, jun 2017. URL http://arxiv.org/abs/1706.10291.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to Escape Saddle Points Efficiently. *Journal of Geometric Analysis*, 26(1):231–251, Mar 2017. ISSN 10506926. doi: 10.1007/s12220-014-9547-5. URL http://arxiv.org/abs/1703.00887.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. Stochastic Gradient Descent Escapes Saddle Points Efficiently. *arXiv preprint arXiv:1902.04811*, 2019. URL http://arxiv.org/abs/1902.04811.

Chris Junchi Li, Zhaoran Wang, and Han Liu. Online ICA: Understanding Global Dynamics of Nonconvex Optimization via Diffusion Processes. *Advances in Neural Information Processing Systems 29*, pages 4967–4975, 2016. ISSN 10495258. doi: 10.1055/s-2004-822831.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic Gradient Descent as Approximate Bayesian Inference. *Journal of Machine Learning Research*, 18:1–35, 2017. ISSN 15337928.

Song Mei, Andrea Montanari, and Phan Minh Nguyen. A Mean Field View of the Landscape of Two-layer Neural Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33):E7665–E7671, Apr 2018. ISSN 10916490. doi: 10.1073/pnas.1806579115. URL http://arxiv.org/abs/1804.06561.

Deanna Needell, Nathan Srebro, and Rachel Ward. Stochastic Gradient Descent, Weighted Sampling, and the Randomized Kaczmarz Algorithm. *Mathematical Programming*, 155 (1-2):549–573, 2016. ISSN 14364646. doi: 10.1007/s10107-015-0864-7.

Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry Nicholas Chapman, Jianwei Miao, and Mordechai Segev. Phase Retrieval with Application to Optical Imaging: A Contemporary Overview. *IEEE Signal Processing Magazine*, 32(3):87–109, May 2015. ISSN 1053-5888. doi: 10.1109/MSP.2014.2352673. URL http://arxiv.org/abs/1402.7350http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7078985.

Thomas Strohmer and Roman Vershynin. A Randomized Kaczmarz Algorithm with Exponential Convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009. ISSN 10695869. doi: 10.1007/s00041-008-9030-4.

Ju Sun, Qing Qu, and John Wright. A Geometric Analysis of Phase Retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, Aug 2018. ISSN 16153383. doi: 10.1007/s10208-017-9365-9. URL http://arxiv.org/abs/1602.06664http://dx.doi.org/10.1007/s10208-017-9365-9http://link.springer.com/10.1007/s10208-017-9365-9.

Yan Shuo Tan and Roman Vershynin. Phase Retrieval via Randomized Kaczmarz: Theoretical Guarantees. *Information and Inference: A Journal of the IMA*, Apr 2018. ISSN 2049-8764. doi: 10.1093/imaiai/iay005. URL http://arxiv.org/abs/1706.09993https://academic.oup.com/imaiai/advance-article/doi/10.1093/imaiai/iay005/4959059.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018. ISBN 978-1108415194. URL https://books.google.com.sg/books?id=J-VjswEACAAJ.

Gang Wang, Georgios B. Giannakis, and Jie Chen. Scalable Solvers of Random Quadratic Equations via Stochastic Truncated Amplitude Flow. *IEEE Transactions on Signal Processing*, 65(8):1961–1974, may 2017. ISSN 1053587X. doi: 10.1109/TSP.2017.2652392. URL http://arxiv.org/abs/1605.08285.

Ke Wei. Solving Systems of Phaseless Equations via Kaczmarz Methods: A Proof of Concept Study. *Inverse Problems*, 31(12):125008, Dec 2015. ISSN 13616420. doi: 10.1088/0266-5611/31/12/125008. URL http://arxiv.org/abs/1502.01822http://stacks.iop.org/0266-5611/31/i=12/a=125008?key=crossref.c102af53b008e09d07a576b02fef01a0.

Huishuai Zhang, Yi Zhou, Yingbin Liang, and Yuejie Chi. Reshaped Wirtinger Flow and Incremental Algorithm for Solving Quadratic System of Equations. *NIPS Proceedings*, pages 2622–2630, 2016. ISSN 10495258. URL http://arxiv.org/abs/1605.07719.

Appendix A. State space calculations

Recall that we let $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ denote the sequence obtained by iteratively performing the SGD update (3) with constant step size $\eta = \frac{1}{d}$.

Lemma 23 *The sequence $\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots$ is a Markov chain whose transition kernel has the random mapping representation*

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \frac{1}{d}\Delta(\boldsymbol{x}^{(k)}), \tag{61}$$

where

$$\Delta(\boldsymbol{x}) = (\langle \boldsymbol{a}, \boldsymbol{x}^* - \boldsymbol{x}\rangle - 2\mathbf{1}_{\mathcal{A}} \cdot \langle \boldsymbol{a}, \boldsymbol{x}^*\rangle)\boldsymbol{a}, \tag{62}$$

$\boldsymbol{a} \sim \mathrm{Unif}(\sqrt{d}S^{d-1})$, and $\mathcal{A}$ is the event that $\mathrm{sign}(\langle \boldsymbol{a}, \boldsymbol{x}\rangle) \neq \mathrm{sign}(\langle \boldsymbol{a}, \boldsymbol{x}^*\rangle)$.

Proof The fact that the sequence is a Markov chain is clear. After dropping indices, we may write the update step (3) as

$$\begin{aligned}\Delta(\mathbf{x}) &\coloneqq (\mathrm{sign}(\langle \mathbf{a}, \mathbf{x}\rangle)|\langle \mathbf{a}, \mathbf{x}^*\rangle| - \langle \mathbf{a}, \mathbf{x}\rangle)\mathbf{a} \\ &= (\mathrm{sign}(\langle \mathbf{a}, \mathbf{x}\rangle)\mathrm{sign}(\langle \mathbf{a}, \mathbf{x}^*\rangle)\langle \mathbf{a}, \mathbf{x}^*\rangle - \langle \mathbf{a}, \mathbf{x}\rangle)\mathbf{a} \\ &= (\langle \mathbf{a}, \mathbf{x}^* - \mathbf{x}\rangle - 2\mathbf{1}\{\langle \mathbf{a}, \mathbf{x}\rangle\langle \mathbf{a}, \mathbf{x}^*\rangle < 0\}\langle \mathbf{a}, \mathbf{x}^*\rangle)\mathbf{a}\end{aligned}$$

■

Proof (of Theorem 3) Abusing notation slightly, let us define the state space updates

$$\alpha(\mathbf{x}) \coloneqq d \cdot \left(r^2(\mathbf{x} + \frac{1}{d}\Delta(\mathbf{x})) - r^2(\mathbf{x})\right)$$

$$\beta(\mathbf{x}) \coloneqq d \cdot \left(s(\mathbf{x} + \frac{1}{d}\Delta(\mathbf{x})) - s(\mathbf{x})\right).$$

Our goal is to compute formulas for the distributions for $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$, in particular showing that they depend on $\mathbf{x}$ only through $\mathbf{y}$. To this end, we simplify notation, denoting $r = r(\mathbf{x})$, $s = s(\mathbf{x})$, and $\theta = \theta(\mathbf{x})$. Furthermore, define $\mathbf{x}^\perp \coloneqq \frac{\mathbf{P}^\perp_{\mathbf{x}^*}\mathbf{x}}{\|\mathbf{P}^\perp_{\mathbf{x}^*}\mathbf{x}\|}$.

We can decompose $\mathbf{x}$ into its components parallel and perpendicular to $\mathbf{x}^*$, writing

$$\mathbf{x} = r\cos\theta\mathbf{x}^* + r\sin\theta\mathbf{x}^\perp. \tag{63}$$

Let $\mathbf{a} \sim \mathrm{Unif}(\sqrt{d}S^{d-1})$ be the random vector used to generate $\Delta(\mathbf{x})$. We also have the orthogonal decomposition

$$\mathbf{a} = a_1\mathbf{x}^* + a_2\mathbf{x}^\perp + \mathbf{r} \tag{64}$$

where a_1, a_2 are the marginals of $\mathbf{a}$ along $\mathbf{x}^*$ and $\mathbf{x}^\perp$ respectively, and $\mathbf{r}$ is defined as the remainder in the decomposition above.

Combining these formulas, we immediately get

$$\langle \mathbf{a}, \mathbf{x}\rangle = r\cos\theta a_1 + r\sin\theta a_2. \tag{65}$$

To compute the formula for $\alpha(\mathbf{x})$, we first expand

$$\begin{aligned}\alpha(\mathbf{x}) &= d \cdot \left(r^2(\mathbf{x} + \Delta(\mathbf{x})/d) - r^2(\mathbf{x})\right) \\ &= d \cdot \left(\|\mathbf{x} + \Delta(\mathbf{x})/d\|^2 - \|\mathbf{x}\|^2\right) \\ &= 2\langle \mathbf{x}, \Delta(\mathbf{x})\rangle + \|\Delta(\mathbf{x})\|^2/d.\end{aligned} \tag{66}$$

Now use equations (62), (64) and (65) to write

$$\langle \Delta(\mathbf{x}), \mathbf{x} \rangle = (\langle \mathbf{a}, \mathbf{x}^* - \mathbf{x} \rangle - 2\mathbf{1}_{\mathcal{A}} \cdot \langle \mathbf{a}, \mathbf{x}^* \rangle)\langle \mathbf{a}, \mathbf{x} \rangle,$$

and

$$\begin{aligned} \frac{\|\Delta(\mathbf{x})\|^2}{d} &= \frac{\|\mathbf{a}\|^2}{d} \cdot \left(\langle \mathbf{a}, \mathbf{x}^* - \mathbf{x} \rangle^2 + 4\langle \mathbf{a}, \mathbf{x}^* \rangle^2 \mathbf{1}_{\mathcal{A}} - 4\langle \mathbf{a}, \mathbf{x}^* - \mathbf{x} \rangle \langle \mathbf{a}, \mathbf{x}^* \rangle \mathbf{1}_{\mathcal{A}} \right) \\ &= \langle \mathbf{a}, \mathbf{x}^* - \mathbf{x} \rangle^2 + 4\langle \mathbf{a}, \mathbf{x} \rangle \langle \mathbf{a}, \mathbf{x}^* \rangle \mathbf{1}_{\mathcal{A}} \end{aligned} \tag{67}$$

Plugging these into (66), we have

$$\begin{aligned} \alpha(\mathbf{x}) &= 2\langle \mathbf{a}, \mathbf{x}^* - \mathbf{x} \rangle \langle \mathbf{a}, \mathbf{x} \rangle + \langle \mathbf{a}, \mathbf{x}^* - \mathbf{x} \rangle^2 \\ &= \langle \mathbf{a}, \mathbf{x}^* \rangle^2 - \langle \mathbf{a}, \mathbf{x} \rangle^2 \\ &= a_1^2 - (r\cos\theta a_1 + r\sin\theta a_2)^2 \\ &= (1 - r^2\cos^2\theta)a_1^2 - r^2\sin^2\theta a_2^2 + r^2\sin\theta\cos\theta a_1 a_2 \end{aligned}$$

For the second formula, we note that $\beta(\mathbf{x}) = \langle \Delta(\mathbf{x}), \mathbf{x}^* \rangle$, and write

$$\begin{aligned} \beta(\mathbf{x}) &= (\langle \mathbf{a}, \mathbf{x}^* - \mathbf{x} \rangle - 2\mathbf{1}_{\mathcal{A}} \cdot \langle \mathbf{a}, \mathbf{x}^* \rangle)\langle \mathbf{a}, \mathbf{x}^* \rangle \\ &= (a_1(1 - r\cos\theta) - a_2 r\sin\theta - 2\mathbf{1}_{\mathcal{A}} a_1)a_1 \\ &= (1 - r\cos\theta)a_1^2 - r\sin\theta a_1 a_2 - 2\mathbf{1}_{\mathcal{A}} a_1^2. \end{aligned}$$

Here, the second equality follows from a combination of (64), (63) and (65). ■

Lemma 24 (Expectations involving $\mathcal{A}$) *The distribution of the event $\mathcal{A} = \mathcal{A}(\theta)$ depends only on the angle θ between $\boldsymbol{x}$ and $\boldsymbol{x}^*$. Furthermore, using the formula* (64)*, we have the following identities.*

$$\mathbb{P}\{\mathcal{A}(\theta)\} = \frac{\theta}{\pi} \tag{68}$$

$$\mathbb{E}\{a_1^2 \mathbf{1}_{\mathcal{A}(\theta)}\} = \frac{1}{2\pi}(2\theta - \sin(2\theta)) \tag{69}$$

$$\mathbb{E}\{a_1 a_2 \mathbf{1}_{\mathcal{A}(\theta)}\} = \frac{1}{2\pi}(\cos(2\theta) - 1) \tag{70}$$

Proof For the first identity, let $\tilde{\mathbf{a}} \coloneqq (a_1, a_2)$. We may write $a_1 = \|\tilde{\mathbf{a}}\|\cos t$ and $a_2 = \|\tilde{\mathbf{a}}\|\sin t$. Event $\mathcal{A}$ occurs precisely when $\frac{\pi}{2} \le t \le \frac{\pi}{2} + t$.

To prove (69), observe that t is independent of $\|\tilde{\mathbf{a}}\|$. This allows us to compute:

$$\begin{aligned}
\mathbb{E}\{a_1^2 \mathbf{1}_{\mathcal{A}}\} &= \mathbb{E}\{(a_1^2 + a_2^2)\}\mathbb{E}\{\cos^2(t)\mathbf{1}_{\mathcal{A}}\} \\
&= 2 \cdot \frac{1}{\pi} \int_{\frac{\pi}{2}}^{\frac{\pi}{2}+\theta} \cos^2(t) dt \\
&= \frac{2}{\pi} \int_0^{\theta} \sin^2(t) dt \\
&= \frac{1}{\pi} \int_0^{\theta} 1 - \cos(2t) dt \\
&= \frac{1}{\pi} \left[t - \frac{\sin(2t)}{2} \right]_0^{\theta} \\
&= \frac{1}{\pi} \left(\theta - \frac{\sin(2\theta)}{2} \right).
\end{aligned}$$

A similar calculation yields (70). ■

Proof (of Lemma 4) It is easy to compute $\mathbb{E}u^2 = 1$, and $\mathbb{E}uv = 0$. Applying this and (69) in the previous lemma to (8) and (9) yields (11) and (12). ■

Appendix B. Properties of subexponential random variables

Subexponential random variables are defined in terms of tail bounds, and can also be equivalently defined as elements of an Orlicz space with Orlicz norm

$$\|X\|_{\psi_1} \coloneqq \inf\{C \colon \mathbb{E} \exp(X/C) \leq 2\}.$$

One may easily check that this is a norm, which allows for easy tail bounds for random variables that are sums of subexponential random variables. We will only state the propreties of subexponential variables needed in our paper, and refer the interested reader to the textbook by Vershynin (2018).

Lemma 25 (ψ_1-norm for $\alpha(\mathbf{y})$ and $\beta(\mathbf{y})$) *Let $\mathcal{D} \subset \mathcal{Y}$ be a compact domain. Then the subexponential norms of $\alpha(\boldsymbol{y})$ and $\beta(\boldsymbol{y})$ are uniformly bounded for $\boldsymbol{y} \in \mathcal{Y}$.*

Proof Suppose we have $r(\mathbf{y}) \leq R$ for all $\mathbf{y} \in \mathcal{Y}$. Then

$$\begin{aligned}
\|\alpha(\mathbf{y})\|_{\psi_1} &= \|(1 - r^2\cos^2\theta) \cdot u^2 - r^2 \sin^2\theta \cdot v^2 + r^2 \sin\theta\cos\theta \cdot uv\|_{\psi_1} \\
&\leq |1 - r^2\cos^2\theta| \cdot \|u^2\|_{\psi_1} + r^2\sin^2\theta \cdot \|v^2\|_{\psi_1} + |r^2 \sin\theta\cos\theta| \cdot \|uv\|_{\psi_1} \\
&\leq (R^2 + 1) \cdot \left(\|u^2\|_{\psi_1} + \|v^2\|_{\psi_1} + \cdot \|uv\|_{\psi_1} \right) \\
&\leq (R^2 + 1) \cdot \left(\|u\|_{\psi_2}^2 + \|v\|_{\psi_2}^2 + \cdot \|u\|_{\psi_2}\|v\|_{\psi_2} \right) \\
&\lesssim R^2 + 1.
\end{aligned}$$

The first inequality is an application of the triangle inequality, the third follows from the following basic property for subexponential random variables: $\|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$. Finally, it is easy to check that $\|u\|_{\psi_2} \lesssim 1$. ■

Appendix C. Lemmas for drift

Proof (of Lemma 5) For each fixed r, $\bar{\beta}$ is an odd function with respect to s, so it suffices to prove the statement for $s > 0$. Let us now compute its derivative with respect to s. Differentiating (12) first with respect to θ, we get

$$\begin{aligned}\frac{d}{d\theta}\bar{\beta}(r^2, s) &= r\sin\theta - \frac{1}{\pi}(2 - 2\cos(2\theta)) \\ &= r\sin\theta - \frac{4}{\pi}\sin^2\theta.\end{aligned}$$

Next, note that since $s = r\cos\theta$, we have

$$\frac{d\theta}{ds} = \left(\frac{ds}{d\theta}\right)^{-1} = -\frac{1}{r\sin\theta}$$

Using the chain rule and then simplifying, we thereby get

$$\begin{aligned}\frac{d}{ds}\bar{\beta}(r^2, s) &= \left(r\sin\theta - \frac{4}{\pi}\sin^2\theta\right)\cdot\left(-\frac{1}{r\sin\theta}\right) \\ &= \frac{4}{\pi}\frac{\sqrt{r^2 - s^2}}{r^2} - 1.\end{aligned}$$

From this expression, we can see that for any fixed $r > 0$, the function $s \mapsto \bar{\beta}(r^2, s)$ is concave downwards on $[0, r]$, with $\bar{\beta}(r^2, 0) = 0$, and $\partial_s\bar{\beta}(r^2, s)\big|_{s=0} > 0$. This implies that the graph of $\bar{\beta}(r^2, s)$ as a function of s lies beneath the line passing through the origin with slope $\partial_s\bar{\beta}(r^2, s)\big|_{s=0}$. When $r \geq \frac{1}{2}$, we have

$$\partial_s\bar{\beta}(r^2, s)\Big|_{s=0} = \frac{4}{\pi r} - 1 < \frac{8}{\pi} - 1$$

thereby giving the upper bound.

For the lower bound, first note that concavity also implies that $\bar{\beta}(r^2, s) \geq \frac{s}{s'}\bar{\beta}(r^2, s')$ for any $0 < s < s' < r$. Now, one may easily check that $\bar{\beta}(1, 1) = 0$, so that $\bar{\beta}(1, 1 - \epsilon) > 0$ for ϵ small enough. By continuity, there is some $\eta > 0$ for which

$$\inf_{|r^2-1|\leq\eta}\bar{\beta}(r^2, 1 - \epsilon) \geq \frac{\bar{\beta}(1, 1 - \epsilon)}{2}.$$

Indeed, since $\partial_r\bar{\beta}(r^2, s) = -\cos\theta$, one may even provide a precise formula if one wishes. Set $\underline{b} \coloneqq \frac{\bar{\beta}(1,1-\epsilon)}{2(1-\epsilon)}$. This is the universal constant we want. ■

Lemma 26 (Lipschitz continuity) *Fix $\epsilon > 0$ in the previous lemma, and by making ϵ and η smaller if necessary, assume that $|s| < r - \epsilon/2$ for all $(r^2, s) \in \mathcal{D}$, where $\mathcal{D} := \{(r^2, s) \in \mathcal{Y} : \ |s| \le 1-\epsilon, |r^2 - 1| \le \eta\}$. Then $\bar{\beta}$ is Lipschitz continuous on $\mathcal{D}$ with Lipschitz constant bounded by a universal constant L depending only on ϵ.*

Proof We have

$$\begin{aligned}\big|\bar{\beta}(\mathbf{y}) - \bar{\beta}(\mathbf{y}')\big| &\le \big|s - s'\big| + \frac{2}{\pi}|\theta - \theta'| + \frac{1}{\pi}\big|\sin(2\theta) - \sin(2\theta')\big| \\ &\le \big|s - s'\big| + \frac{4}{\pi}|\theta - \theta'|.\end{aligned}$$

The first term is trivially bounded by $\|\mathbf{y} - \mathbf{y}'\|$. Next, observe that $\theta = \arccos(s/\sqrt{r^2})$, which is jointly differentiable in r^2 and s, and so is Lipschitz continuous with respect to these coordinates on a compact set bounded away from $r = s$. By assumption, $\mathcal{D}$ is such a compact set. ∎

Appendix D. Facts about Kolmogorov distance and stochastic dominance

Lemma 27 (Properties of Kolmogorov distance) *For any real-valued random variables X and Y, we have*

$$\|F_{X^2} - F_{Y^2}\|_\infty \le 2\|F_X - F_Y\|_\infty. \tag{71}$$

In addition, for any constant $c \in \mathbb{R}$, we have

$$\|F_{X+c} - F_{Y+c}\|_\infty = \|F_X - F_Y\|_\infty \tag{72}$$

Proof For any $t \ge 0$, we have

$$\begin{aligned}|F_{X^2}(t) - F_{Y^2}(t)| &= \mathbb{P}\{X^2 \le t\} - \mathbb{P}\{Y^2 \le t\} \\ &= \mathbb{P}\{-t \le X \le t\} - \mathbb{P}\{-t \le Y \le t\} \\ &= F_X(t) - F_X(-t) + F_Y(t) - F_Y(-t) \\ &\le |F_X(t) - F_Y(t)| + |F_X(-t) - F_Y(-t)| \\ &\le 2\|F_X - F_Y\|_\infty\end{aligned}$$

as we wanted. The second identity can be obtained similarly. ∎

Lemma 28 (Characterization of stochastic dominance in terms of CDFs) *Let X and Y be real-valued random variables. Then for any $0 < \delta < 1$, X stochastically dominates Y up to error δ if and only if their CDFs satisfy $F_X \le F_Y + \delta$.*

Proof The forward implication is trivial, so we only need to show the backward implication. Let q_X and q_Y denote the quantile functions for X and Y respectively. We claim that

for any $0 \le t \le 1-\delta$, we have $q_X(t+\delta) \ge q_Y(t)$. To see this, first recall the definition $q_Y(t) = \inf\{a : F_Y(a) \ge t\}$. By definition, we thus have a decreasing sequence $x_n \downarrow q_X(t+\delta)$ such that $F_X(x_n) \ge t+\delta$ for each n. Then by assumption, $F_X(a) \le F_Y(a) + \delta$ for any $a \in \mathbb{R}$, so

$$F_Y(x_n) \ge F_X(x_n) - \delta \ge (t+\delta) - \delta = t,$$

which implies that

$$q_Y(t) \le \liminf_{n\to\infty} x_n = q_X(t+\delta)$$

as claimed.

Let U be uniformly distributed on $[0,1]$, and on the same probability space, define $U'(\omega) = U(\omega) + \delta \mod 1$. Then U' is also uniformly distributed on $[0,1]$. Next, it easy to check that $q_X(U') \sim F_X$ and $q_Y(U) \sim F_Y$ (this is a standard construction in probability theory). With probability $1-\delta$, we have $\delta \le U' = U + \delta \le 1$. Conditioning on the event in which this occurs, we have

$$q_X(U') = q_X(U+\delta) \ge q_Y(U),$$

which gives us the coupling we want. ■

Lemma 29 (Transitivity of stochastic dominance) *Let X, Y, and Z be random variables, $\delta, \delta' > 0$ such that $X \succeq_\delta Y$ and $Y \succeq_{\delta'} Z$. Then $X \succeq_{\delta+\delta'} Z$.*

Proof This follows from the characterization in the previous lemma. ■

Lemma 30 (Stochastic dominance preserved under monotone transformations) *Let X and Y, and Z be random variables, $\delta > 0$ such that $X \succeq_\delta Y$. Suppose X and Y have range $R \subset \mathbb{R}$, and $\tau\colon R \to R$ is a non-decreasing transformation. Then $\tau(X) \succeq_\delta \tau(Y)$.*

Proof This is obvious. ■

Corollary 31 (Kolmogorov distance and coupling) *Let X and Y be real-valued random variables such that their CDFs satisfy $\|F_X - F_Y\|_\infty \le \delta$ for some $0 < \delta < 1$. Then X stochastically dominates Y up to error δ.*

Lemma 32 (Stochastic dominance ordering for truncated Gaussians) *The following hold:*

1. *Fix $\sigma^2, \epsilon > 0$. For any $b > 0$, define the random variable $X_b \coloneqq \rho_\epsilon[b + \sigma g]^2$. Then whenever $b' > b$, we have $X_{b'} \succeq X_b$.*
2. *Fix σ_0^2, set $\sigma^2 = \frac{B\sigma_0^2}{d^2}$. Let $\epsilon(s)$ and $b(s)$ be defined as in (33) and (36) respectively . Consider the collection of random variables $Y_s \coloneqq \rho_{\epsilon(s)}[b(s)+\sigma g]^2$ for $-1/2 < s < 1/2$. For d large enough, whenever $1/2 > s' > s > 0$, we have $Y_{s'} \succeq Y_s$.*

Proof Throughout, we let g denote a standard normal random variable. We start by proving the first statement. Let F_b denote the CDFs of X_b and $X_{b'}$ respectively. For any given $a > 0$, we wish to show that $F_b(a) \geq F_{b'}(a)$, and it suffices to show that $\frac{d}{db}F_b(a) \leq 0$ for $b > 0$. In order to do this, we write out F_s in terms of the Gaussian CDF. We have

$$\begin{aligned} F_b(a) &= \mathbb{P}\{X_s \leq a\} \\ &= \mathbb{P}\{-\sqrt{a} - \epsilon \leq \sigma g + b \leq \sqrt{a} + \epsilon\} \\ &= \mathbb{P}\left\{g \leq \frac{\sqrt{a} + \epsilon - b(s)}{\sigma}\right\} - \mathbb{P}\left\{g \leq \frac{-\sqrt{a} - \epsilon - b}{\sigma}\right\} \\ &= \Phi\left(\frac{\sqrt{a} + \epsilon - b}{\sigma}\right) - \Phi\left(\frac{-\sqrt{a} - \epsilon - b}{\sigma}\right). \end{aligned}$$

Hence, we have

$$\frac{d}{db}F_b(a) = \frac{1}{\sigma}\left(\phi\left(\frac{-\sqrt{a} - \epsilon - b}{\sigma}\right) - \phi\left(\frac{\sqrt{a} + \epsilon - b}{\sigma}\right)\right).$$

Since ϕ is even, and decreases away from 0, and $|-\sqrt{a} - \epsilon - b| > |\sqrt{a} + \epsilon - b|$ for $b > 0$, the quantity on the right hand side is negative as we wanted.

The second statement is proved similarly. We let F_s denote the CDF of Y_s and will show that $\frac{d}{ds}F_s(a) \leq 0$. As before we compute

$$F_s(a) = \Phi\left(\frac{\sqrt{a} + \epsilon(s) - b(s)}{\sigma}\right) - \Phi\left(\frac{-\sqrt{a} - \epsilon(s) - b(s)}{\sigma}\right).$$

Hence, we have

$$\frac{d}{ds}F_s(a) = \frac{\epsilon'(s) - b'(s)}{\sigma}\phi\left(\frac{\sqrt{a} + \epsilon(s) - b(s)}{\sigma}\right) - \frac{-\epsilon'(s) - b'(s)}{\sigma}\phi\left(\frac{-\sqrt{a} - \epsilon(s) - b(s)}{\sigma}\right).$$

For $0 < s < \sqrt{\frac{\log d}{B}}$, we have $\epsilon'(s) = 0$, and it is clear that this quantity is nonpositive. For $s > \sqrt{\frac{\log d}{B}}$, first observe that for any x, y, we have $e^{-(x-y)^2}/e^{-(x+y)^2} = e^{4xy}$. As such, we compute

$$\begin{aligned} \phi\left(\frac{\sqrt{a} + \epsilon(s) - b(s)}{\sigma}\right)\Big/\phi\left(\frac{-\sqrt{a} - \epsilon(s) - b(s)}{\sigma}\right) &= \exp\left(\frac{2(\sqrt{a} + \epsilon(s))b(s)}{\sigma^2}\right) \\ &\geq \exp\left(\frac{d^2}{B\sigma_0^2} \cdot \frac{CB^2 \log d}{d^2} \cdot s^2\right) \\ &\geq \exp\left(\frac{d^2}{B\sigma_0^2} \cdot \frac{CB^2 \log d}{d^2} \cdot \frac{\log d}{B}\right) \\ &= \exp\left(\frac{C \log^2 d}{\sigma_0^2}\right). \end{aligned}$$

For d large enough, this ratio is greater than

$$\frac{b'(s) + \epsilon'(s)}{b'(s) - \epsilon'(s)} = \frac{1 + \kappa B/d + CB^2 \log d/d^2}{1 + \kappa B/d - CB^2 \log d/d^2},$$

which converges to 1 as d tends to infinity. This concludes the proof of the claim. ■

Appendix E. Combinatorial lemmas

Lemma 33 *Let $x_1, x_2, \dots$ be a sequence of real numbers. Denote the partial sums by $s_t = \sum_{i=1}^{t} x_i$. For any $0 < \rho < 1$, define the sequence $w_1, w_2, \dots$ via the recursive formula*

$$w_{t+1} \coloneqq \rho w_t + x_{t+1}.$$

If there is some positive integer M and some $C > 0$ such that $|s_t| \le C$ for all $t \le M$, then we also have $|w_t| \le 2C + |w_0|$ for all $t \le M$.

Proof We first prove by induction that the following representation for w_t holds:

$$w_t = s_t - (1-\rho)\sum_{i=1}^{t-1} \rho^{t-i-1} s_i + \rho^t w_0. \tag{73}$$

First assume that the formula holds for some t. Then starting from the definition, we use the inductive hypothesis to write

$$\begin{aligned} w_{t+1} &= \rho w_t + x_{t+1} \\ &= \rho \cdot \left(s_t - (1-\rho)\sum_{i=1}^{t-1} \rho^{t-1-i} s_i + \rho^t w_0 \right) + x_{t+1} \\ &= \rho s_t - (1-\rho)\sum_{i=1}^{t-1} \rho^{t-i} s_i + \rho^{t+1} w_0 + (s_{t+1} - s_t) \\ &= s_{t+1} - (1-\rho)\sum_{i=1}^{t} \rho^{t-i} s_i + \rho^{t+1} w_0. \end{aligned}$$

Applying the triangle inequality to (73), we get

$$\begin{aligned} |w_t| &\le |s_t| + (1-\rho)\sum_{i=1}^{t} \rho^{i-1} |s_{t-i}| + \rho^t |w_0| \\ &\le C + (1-\rho)\sum_{i=1}^{\infty} \rho^{i-1} C + |w_0| \\ &\le 2C + |w_0| \end{aligned}$$

as we wanted. ■

Lemma 34 *Let $x_1, x_2, \dots$ be a sequence of non-negative real numbers. Denote the partial sums by $s_t = \sum_{i=1}^{t} x_i$. Let $\rho > 0$ and $\xi > 0$ be such that we have the recursive inequality*

$$x_t \le \rho(\xi + s_{t-1}).$$

Then for all t, we have

$$s_t \le \xi \cdot \big((1+\rho)^{t+1} - 1 \big).$$

Proof We have

$$\begin{aligned} s_t &= s_{t-1} + x_t \\ &\leq s_{t-1} + \rho(\xi + s_{t-1}) \\ &\leq (1+\rho)s_{t-1} + \rho\xi. \end{aligned}$$

Solving this recursion, we get

$$s_t \leq \rho\xi \cdot \sum_{i=0}^{t-1} (1+\rho)^i = \xi \cdot \big((1+\rho)^{t+1} - 1\big).$$

■

Appendix F. Concentration inequalities

Proof (of Lemma 7) This essentially follows from Theorem 1 in Howard et al. (2018). For completeness and clarity, however, we give a direct proof of this here using the same technique.

First fix $\epsilon > 0$. Let $\lambda \leq \frac{1}{2K}$, and for each positive integer t, we write

$$L_t(\lambda) \coloneqq \exp(\lambda S_t - \lambda^2 K^2 t).$$

Observe that

$$\mathbb{E}\{L_{t+1}(\lambda) \,\big|\, \mathcal{G}_t\} = \mathbb{E}\{e^{\lambda X_{t+1} - \lambda^2 K^2} \,\big|\, \mathcal{G}_t\} L_t(\lambda) \leq L_t(\lambda),$$

so that the sequence $L_t(\lambda)$ forms a supermartingale, and remains so if we extend this to time 0 by setting $L_0(\lambda) = 1$. We may now apply the supermartingale inequality. We have

$$\begin{aligned} \mathbb{P}\{\exists t \leq M \,:\, S_t \geq \lambda K^2 M + \epsilon/2\} &\leq \mathbb{P}\{\exists t \leq M \,:\, S_t \geq \lambda K^2 t + \epsilon/2\} \\ &= \mathbb{P}\{\exists t \leq M \,:\, L_t(\lambda) \geq e^{\lambda\epsilon/2}\} \\ &\leq \mathbb{E}\{L_0(\lambda)\} \cdot e^{-\lambda\epsilon/2} \\ &\leq e^{-\lambda\epsilon/2}. \end{aligned} \tag{74}$$

It remains to choose λ appropriately in order to get the tail bound we want. If $\epsilon \leq MK$, then we set $\lambda = \epsilon/2K^2M$, observing that for this choice of λ,

$$\lambda \leq \frac{MK}{2K^2M} \leq \frac{1}{2K},$$

and $L_t(\lambda)$ is indeed a supermartingale, and (74) holds. Plugging our choice of λ into the left hand and right hand sides, this yields the bound

$$\mathbb{P}\{\exists t \leq M \,:\, S_t \geq \epsilon\} \leq e^{-\epsilon^2/4mK^2}.$$

On the other hand, if $\epsilon > MK$, then we pick $\lambda = 1/2K$. Once again plugging this into (74), we get

$$\mathbb{P}\{\exists t \le M \,:\, S_t \ge \epsilon\} \le e^{-\epsilon/4K}.$$

Putting these two bounds together gives the upper tail in (21), and considering the negative of the sequence gives the lower tail. ■

Lemma 35 (Chernoff for non-independent Bernoullis) *Let $X_1, X_2, \dots, X_M$ be a sequence of Bernoulli random variables adapted to a filtration $\{\mathcal{G}_t\}$, and let θ be such that for $1 \le t \le M$, we have*

$$\mathbb{E}\{X_t \,|\, \mathcal{G}_{t-1}\} \le \theta$$

Then for any $0 < \epsilon < 1$, we have

$$\mathbb{P}\left\{\sum_{i=1}^{M} X_i \ge (1+\epsilon)M\theta\right\} \le e^{-M\theta\epsilon^2/3}$$

Proof Denote $S_t \coloneqq \sum_{i=1}^{t} X_i$, observe that for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\{e^{\lambda S_t}\} &= \mathbb{E}\{\mathbb{E}\{e^{\lambda X_t} \,|\, \mathcal{G}_{t-1}\} e^{\lambda S_{t-1}}\} \\
&\le \mathbb{E}\Big\{\exp\Big((e^\lambda - 1)\mathbb{E}\{X_t \,\Big|\, \mathcal{G}_{t-1}\}\Big) e^{\lambda S_{t-1}}\Big\} \\
&\le \exp\Big(\theta(e^\lambda - 1)\Big)\mathbb{E}\{e^{\lambda S_{t-1}}\},
\end{aligned}$$

so that

$$\mathbb{E}\{e^{\lambda S_M}\} \le \exp\Big(M\theta(e^\lambda - 1)\Big).$$

The rest of the proof is exactly the same as that of the regular Chernoff's inequality (Vershynin, 2018). ■

Lemma 36 (Moment bounds for contractions of random variables) *Let X be a random variable, $\rho\colon \mathbb{R} \to \mathbb{R}$ a contraction (i.e. a map such that $|\rho(x) - \rho(y)| \le |x - y|$ for all x, y.) Then $\mathbb{E}\{(\rho(X) - \mathbb{E}\rho(X))^4\} \le 8\mathbb{E}\{(X - \mathbb{E}X)^4\}$. If in addition, ρ is an odd function and X is symmetric, then for any $b \in \mathbb{R}$, we also have $|\mathbb{E}\{\rho(X + b)\}| \le |b|$.*

Proof Let X' be an independent copy of X. By Jensen's inequality, followed by applying the contraction inequality pointwise, we have

$$\mathbb{E}\{(\rho(X) - \mathbb{E}\rho(X))^4\} \le \mathbb{E}\{\big(\rho(X) - \rho(X')\big)^4\} \le \mathbb{E}\Big\{\big(X - X'\big)^4\Big\}.$$

Writing $X - X' = X - \mathbb{E}X + \mathbb{E}X' - X'$ and expanding, we can bound the right hand side via

$$\begin{aligned}
\mathbb{E}\Big\{\big(X - X'\big)^4\Big\} &= \mathbb{E}\Big\{(X - \mathbb{E}X)^4\Big\} + 6\mathbb{E}\{(X - \mathbb{E}X)^2\big(X' - \mathbb{E}X'\big)^2\} + \mathbb{E}\Big\{\big(X' - \mathbb{E}X'\big)^4\Big\} \\
&\le 8\mathbb{E}\{(X - \mathbb{E}X)^4\}.
\end{aligned}$$

Here, the inequality follows from Cauchy-Schwarz.

For the second claim, assume $b > 0$ and write

$$\begin{aligned}\mathbb{E}\{\rho(X+b)\} &= \mathbb{E}\{\rho(X+b) - \rho(X)\} \\ &\leq \mathbb{E}\{|\rho(X+b) - \rho(X)|\} \\ &\leq \mathbb{E}\{|(X+b) - X|\} \\ &= b.\end{aligned}$$

The first equality follows from the oddness of ρ and the symmetry of X, while the second follows from contraction. If $b < 0$, the statement may be proved similarly. ■