High-dimensional Limit of SGD for Diagonal Linear Networks

Begoña García Malaxechebarría
University of Washington
begogar9@uw.edu

Courtney Paquette
McGill University
courtney.paquette@mcgill.ca

Maryam Fazel
University of Washington & Amazon
mfazel@uw.edu

Dmitriy Drusvyatskiy
University of California, San Diego
ddrusvyatskiy@ucsd.edu

Abstract

Understanding the behavior of stochastic gradient methods is a central problem in modern machine learning. Recent work has highlighted diagonal linear networks as a simplified yet expressive setting for analyzing the optimization and generalization properties of neural models. In this work, we show that in the high-dimensional regime, stochastic gradient descent on diagonal linear networks is well-approximated by continuous dynamics governed by a stochastic differential equation (SDE), which explicitly decouples the drift from the gradient noise. We further derive a deterministic partial differential equation whose solution propagates the relevant state of the iterates and characterizes the time evolution of a broad class of observable statistics, including the risk, curvature, and other metrics for optimality. To the best of our knowledge, this is the first deterministic high-dimensional description of SGD for a genuinely nonlinearly parametrized model. Finally, we show that, under a suitable parametrization, the stochastic dynamics are globally well posed and converge exponentially fast to zero risk with high probability, yielding a fully explicit non-asymptotic description of their long-time behavior. Numerical simulations corroborate our theoretical findings.

1 Introduction

Diagonal linear networks provide a tractable model for studying phenomena in deep learning. For example, they exhibit implicit bias, where stochastic optimization favors low-complexity solutions even without explicit regularization [Pesme et al., 2021, Andriushchenko et al., 2023b, Even et al., 2023]. However, analyzing SGD in this setting remains challenging. Classical analyses are often too coarse, relying on worst-case guarantees valid only for very small stepsizes, whereas the relevant phenomena arise at larger stepsizes. The difficulty stems from the intrinsic randomness of SGD, due to stochastic data sampling at each iteration. This motivates sharper analytical frameworks that capture the evolution of quantities such as risk and curvature while remaining tractable.

One approach is to study gradient flow or diffusion approximations [Pesme et al., 2021, Vivien et al., 2022]. Gradient flow arises in the vanishing-stepsize limit, while diffusion adds noise to capture stochasticity. These methods are typically justified only for $\gamma \to 0$, a regime unrealistic in practice.

SGD in modern applications typically uses large stepsizes in high dimensions, where neither the stepsize nor the stochasticity is negligible. We instead exploit high dimensionality—the large number of parameters—as the simplifying mechanism. In this limit, concentration yields deterministic expressions for the risk (and other quantities of interest) at each iteration, even for large stepsizes.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

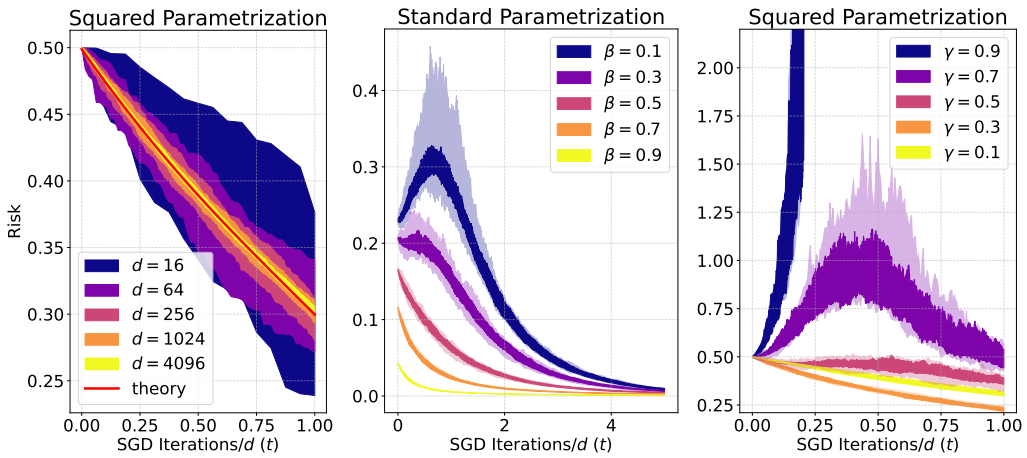

Figure 1: **Three views of empirical risk dynamics for SGD on a diagonal linear network.** *Left:* Covariance $K = I_d$. As d increases, the risk trajectory concentrates around a deterministic limit (red) described in Theorem 3.7. *Middle:* Power-law covariance spectrum. The homogenized SGD (transparent) from Theorem 3.7 closely tracks SGD (opaque) over a range of power-law exponents β in dimension $d = 10^3$. *Right:* Covariance $K = I_d$ in dimension $d = 10^3$. Varying the stepsize γ reveals distinct convergence/divergence regimes; the homogenized prediction remains accurate even for stepsizes above the convergence threshold. Left and right panels use the parametrization (F.4), whereas the middle one uses the parametrization (F.3). See Appendix G for simulation details.

Specifically, we derive a continuous-time stochastic differential equation (SDE) that tracks discrete-time SGD in this high-dimensional, large-stepsize regime for diagonal linear networks. From this SDE, we obtain a deterministic PDE that propagates the state of the iterates and predicts the evolution of a broad class of observable statistics. *To the best of our knowledge, our result provides the first deterministic description of the discrete-time SGD trajectory in high dimensions with fixed, non-vanishing stepsizes for diagonal linear networks, placing them among the first nonlinearly parametrized models admitting a description beyond gradient-flow or stochastic-gradient-flow limits.*

This effectively allows for the analysis of SGD *without running it* (Fig. 1): solving the deterministic dynamics yields learning curves without costly Monte Carlo simulations. While the PDE provides theoretical insight, the SDE remains a powerful practical tool: it is cheaper to simulate, and Itô calculus gives closed evolution equations for many statistics, which we validate numerically.

Setting the stage, our work targets a general class of problems where the risk exhibits the form:

$$\mathcal{R}(x) = \mathbb{E}_a \left[f\left(\frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x) \\ \beta^* \end{pmatrix}^\top a \right) \right] \qquad \text{for } a \sim \mathcal{N}(0, K). \tag{1}$$

Here $K \in \mathbb{R}^{d\times d}$ is the data covariance, $\beta^* \in \mathbb{R}^d$ is the signal, $\psi(x)$ maps $x \in \mathbb{R}^{2d}$ into $\mathbb{R}^d$, and f is a loss. The main example of (1) is the mean-squared error of a two-layer *diagonal linear network*,

$$\min_{u,v\in\mathbb{R}^d} \frac{1}{2d} \mathbb{E}_a \langle u \odot v - \beta^*, a\rangle^2, \tag{2}$$

where $x = (u, v)$ with $u, v \in \mathbb{R}^d$ and $u \odot v$ is the Hadamard product. Thus, although the model is linear in its output, the optimization variables enter the risk nonlinearly. This makes the high-dimensional closure problem more delicate than in linear or generalized linear models. It is standard that (2) admits a difference-of-squares reparametrization [Nacson et al., 2022, Pesme et al., 2021]:

$$\min_{u,v\in\mathbb{R}^d} \frac{1}{4d} \mathbb{E}_a \langle u^2 - v^2 - \beta^*, a\rangle^2. \tag{3}$$

Both formulations of diagonal linear networks have drawn attention for their ability to recover sparse solutions [Pesme et al., 2021, Vaskevicius et al., 2019, Kanade et al., 2023, HaoChen et al., 2021].

To solve (1), we run *streaming* stochastic gradient descent under a deterministic stepsize schedule γ_k, that is, at each iteration we sample $a_{k+1} \sim \mathcal{N}(0, K)$ and update the iterates according to the rule

$$x_{k+1} = x_k - \gamma_k \nabla_x \Psi(x_k; a_{k+1}),$$

where $\Psi(x, a)$ is the integrand in (1). To analyze SGD, we introduce a SDE, *homogenized SGD*:

$$\mathrm{d}\mathcal{X}_t = -\gamma(t)\, d\, \nabla \mathcal{R}(\mathcal{X}_t)\mathrm{d}t + \gamma(t)\sqrt{\mathbb{E}_a\left[\nabla f\left(\tfrac{1}{\sqrt{d}}\psi(\mathcal{X}_t)^\top a\right)^2\right]}\, \nabla\psi(\mathcal{X}_t)^\top \sqrt{K}\, \mathrm{d}\mathfrak{B}_t, \tag{4}$$

where the risk $\mathcal{R}$ and inner map ψ are defined in (1), the initial conditions are given by $\mathcal{X}_0 = x_0$, and $\mathfrak{B}_t$ is a standard Brownian motion in $\mathbb{R}^d$. In this work, we show that homogenized SGD matches SGD in high dimensions, even with stepsizes at or above the convergence threshold (see Fig. 1).

To enable a comparison between SGD and homogenized SGD, we extend the discrete-time iterate sequence $\{x_k\}$ to continuous time. The k-th iterate of SGD corresponds to the continuous time parameter t in homogenized SGD via $k = \lfloor td \rfloor$. Thus, when $t = 1$, SGD has completed d updates.

Main Contributions. Our first result shows that SGD and homogenized SGD are close: for a broad class of statistics φ, the values $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$ are uniformly close along the trajectories. Two key examples are the risk $\mathcal{R}(x)$ and the scaled Hessian-trace $\frac{1}{d}\operatorname{Tr}(\nabla^2 \mathcal{R}(x))$ (see Figures 1 and 2).

Theorem 1.1 (Informal)**.** *Under mild conditions of Theorem 3.7, for any φ satisfying Assumption 3.6, $\varepsilon \in (0, \frac{1}{2})$, and $T > 0$, there is a constant C (independent of d) such that w.o.p.*[1] *the estimate holds:*

$$\sup_{0 \leqslant t \leqslant T} \left|\varphi(x_{\lfloor td \rfloor}) - \varphi(\mathcal{X}_t)\right| \leqslant C d^{-\varepsilon}. \tag{5}$$

Thus, in the high-dimensional regime, SGD is well described by homogenized SGD. The SDE is both accurate and tractable, providing a practical tool to study SGD at fixed (possibly large) stepsizes. This contrasts with classical diffusion approximations, which instead require $\gamma \to 0$ at fixed dimension.

Our second main result goes a step further. We show that, in high dimensions, randomness is negligible at the level of φ. Specifically, $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$ concentrate uniformly over $t \in [0, T]$ around a *deterministic function* $\phi(t)$ (Fig. 1), given by the deterministic PDE system (22)–(30). Thus, for large d, the function $\phi(t)$ predicts statistic curves without averaging over SGD noise.

Theorem 1.2 (Informal)**.** *Under mild conditions of Theorem 3.7, for any φ satisfying Assumption 3.6, $\varepsilon \in (0, \frac{1}{2})$, and $T > 0$, there is a constant C (independent of d) such that w.o.p. the estimate holds:*

$$\sup_{0 \leqslant t \leqslant T} \left(\left|\varphi(x_{\lfloor td \rfloor}) - \phi(t)\right| + \left|\varphi(\mathcal{X}_t) - \phi(t)\right| \right) \leqslant C d^{-\varepsilon} \tag{6}$$

where $\phi(t)$ is a deterministic function defined in (22) and (30).

Thus the function ϕ provides a deterministic description of a broad class of statistics $\varphi(x)$ along SGD in high dimension, even for *large* stepsizes. We expect this theorem to aid analysis of fine-grained SGD behavior such as instability, saddle avoidance/escape, progressive sharpening, and implicit bias.

Our final result complements this by establishing global linear convergence of the stochastic dynamics.

Theorem 1.3 (Informal)**.** *Consider homogenized SGD (4) under the squared parametrization (3), initialized at $\mathcal{U}_0 = \mathcal{V}_0 = \mathbb{1}_d$. Then a numerical constant $c > 0$ exists such that for any stepsize $\gamma \in (0, c)$, the risk converges to zero exponentially fast. Indeed, for any $\delta \in (0, 1)$, there are constants $C, \mu > 0$ such that with probability at least $1 - \delta$ we have $\mathcal{R}(\mathcal{X}_t) \leqslant C e^{-\mu t}$ for all $t \geqslant 0$.*

1.1 Literature Review

SGD is often analyzed in existing literature through continuous-time approximations, including stochastic gradient flow and diffusion-type SDE models [Pesme et al., 2021]. Related viewpoints include studying SGD augmented with label noise as a proxy for isolating the role of stochasticity

[1] We say an event holds *with overwhelming probability (w.o.p.)* if there is a function $\omega \colon \mathbb{N} \to \mathbb{R}$ with $\omega(d)/\log d \to \infty$ so that the event holds with probability at least $1 - e^{-\omega(d)}$.

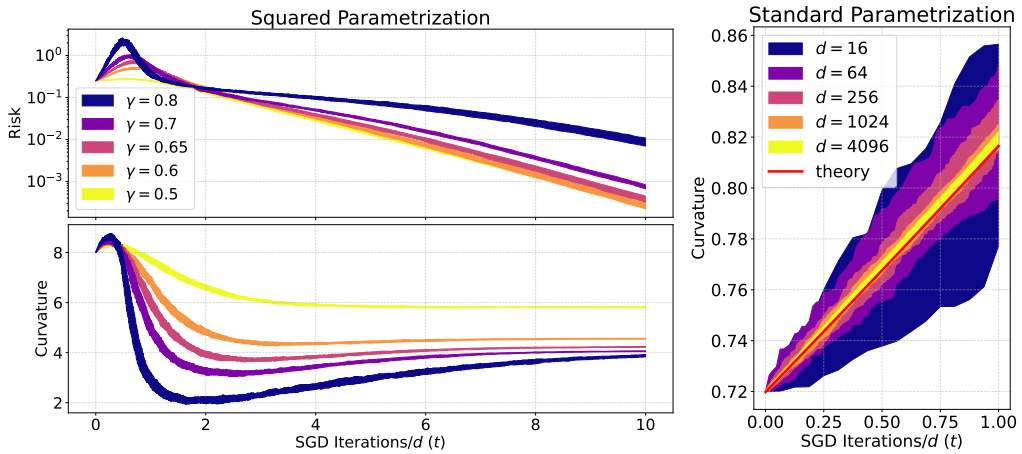

Figure 2: **Curvature dynamics for SGD on a diagonal linear network.** *Left:* Curvature measured by the scaled trace of the Hessian $\frac{1}{d}\operatorname{Tr}(\nabla^2\mathcal{R})$ alongside the empirical risk $\mathcal{R}$, showing "flat" progress in which the risk increases sharply accompanied by a drop in curvature as we vary the stepsize γ. *Right:* As the dimension d increases, the curvature dynamics of SGD concentrate around a deterministic limit (shown in red), as proven in Theorem 3.7. See Appendix G for simulation details.

during training [Vivien et al., 2022]. These diffusion approximations are typically justified in small-learning-rate limits at fixed dimension. Complementary recent work instead derives high-dimensional descriptions that remain informative at *finite* stepsizes under structured data models, including homogenization-type characterizations and analyses that explicitly capture the role of data covariance [Paquette et al., 2025, Collins–Woodfin and Paquette, 2024, Collins-Woodfin et al., 2024a].

There is a long tradition of deriving deterministic ODE descriptions for the risk dynamics of high-dimensional teacher–student and random-feature models trained with streaming SGD. Prior work has leveraged these ODEs to explicitly characterize how algorithmic choices—such as the stepsize—affect convergence and generalization behavior [Saad and Solla, 1995a, Biehl and Riegler, 1994, Biehl and Schwarze, 1995, Saad and Solla, 1995b]. SGD and its high-dimensional limit under isotropic data have been studied extensively (e.g., [Arous et al., 2022]). Beyond the isotropic setting, models with non-identity covariance have been analyzed to understand high-dimensional SGD dynamics [Goldt et al., 2020, 2022, Collins–Woodfin and Paquette, 2024, Collins-Woodfin et al., 2024a], with further extensions to adaptive algorithms [Collins-Woodfin et al., 2024b, Balasubramanian et al., 2025]. More recent work studies SGD and its variants in the multi-pass setting [Paquette et al., 2021, Lee et al., 2022, Paquette et al., 2025]. Additional related work is deferred to Appendix A.

1.2 High-dimensional Model Structure

We will consider objective functions $\mathcal{R}$ defined as the expectation of a composition of a simple outer function $f\colon \mathbb{R}^2 \to \mathbb{R}$ and an inner map $\psi\colon \mathbb{R}^{2d} \to \mathbb{R}^d$, given by

$$\mathcal{R}(x) := \mathbb{E}_a\, \Psi(x;a), \quad \text{for} \quad a \sim \mathcal{D} \subset \mathbb{R}^d, \quad \text{where} \quad \Psi(x;a) := f\left(\frac{1}{\sqrt{d}}\begin{pmatrix}\psi(x)\\ \beta^*\end{pmatrix}^\top a\right). \tag{7}$$

In the probabilistic analysis, f is treated as a function of both inputs $r = (r_1, r_2) \in \mathbb{R}^2$. In the optimization setting, we view f as a function of r_1, with r_2 as a fixed (possibly random) parameter.

We impose the following mild Lipschitzness assumption on f. Throughout, the symbol $\|\cdot\|$ will denote the standard ℓ_2-norm on vectors and the Frobenius norm on matrices.

Assumption 1.4 (Pseudo-Lipschitz continuity of f)**.** The outer function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$. That is, for all $r, \hat{r} \in \mathbb{R}^2$, it holds:

$$|f(r) - f(\hat{r})| \leqslant L(f)\,\|r - \hat{r}\|\,(1 + \|r\|^\alpha + \|\hat{r}\|^\alpha).$$

In words, pseudo-Lipschitzness stipulates polynomial growth of Lipschitz constants on balls with growing radii. Next we summarize our assumption on the data vector $a \sim \mathcal{D}$.

Assumption 1.5 (Data)**.** We consider data samples $a \sim \mathcal{D}$ drawn from a Gaussian distribution $\mathcal{N}(0, K)$, where the covariance $K \in \mathbb{R}^{d\times d}$ is diagonal. The entries of K are uniformly bounded, meaning its operator norm satisfies $\|K\|_{\mathrm{op}} \leqslant \bar{K}$, for some constant $\bar{K} > 0$ independent d. In addition, we assume that the trace of K scales linearly with the dimension d, that is, $\mathrm{Tr}(K) = \Theta(d)$.

Remark 1.6. The Gaussian assumption enables both concentration and conditional moment calculations. Conditioning a on $\mathrm{span}\{\psi(x), \beta^*\}$ yields explicit means and covariances identifying the drift and diffusion, and gives concentration for the linear and quadratic forms in the martingale estimates. While some steps extend to light-tailed data, these moment calculations explicitly rely on Gaussianity.

Remark 1.7. The condition $\mathrm{Tr}(K) = \Theta(d)$ assumes total variance of the input distribution scales with dimension. If it grows more slowly, different stepsize scalings are needed for nontrivial limits.

Our main target application is a diagonal linear network, under parametrizations (2) and (3). To treat both parametrizations simultaneously, we assume that the inner function ψ has a quadratic form.

Assumption 1.8 (Diagonal Linear Neural Networks)**.** The inner function $\psi\colon \mathbb{R}^{2d} \to \mathbb{R}^d$ is component-wise quadratic, meaning that each output coordinate for $i \in \{1, \dots, d\}$ is given by

$$\psi_i(u, v) = \begin{pmatrix} u_i & v_i \end{pmatrix} \mathcal{Q} \begin{pmatrix} u_i \\ v_i \end{pmatrix} + l^\top \begin{pmatrix} u_i \\ v_i \end{pmatrix} + c, \tag{8}$$

where $\mathcal{Q} = \begin{pmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$ is symmetric, $l = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \in \mathbb{R}^2$, and $c \in \mathbb{R}$.

Concretely, both (2) and (3) correspond to $\psi(u, v) = u \odot v$ for $\mathcal{Q} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$ and $\psi(u, v) = u^2 - v^2$ for $\mathcal{Q} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, with $l = c = 0$ in both cases. Next, we assume the signal has bounded entries.

Assumption 1.9 (Parameter Scaling)**.** The entries of the signal $\beta^* \in \mathbb{R}^d$ are uniformly bounded, that is, $\|\beta^*\|_\infty \leqslant C$ for some $C > 0$ independent of d.

In what follows, define the block matrix $W(x) := [\,\psi(x) \,|\, \beta^* \,|\, \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3}$ for $x = (u, v) \in \mathbb{R}^{2d}$, and let $B(x)$ be the covariance matrix of the random vector $\frac{1}{\sqrt{d}} W(x)^\top a \in \mathbb{R}^3$, or more explicitly

$$B(x) \;:=\; \frac{1}{d}\, W(x)^\top K\, W(x) \in \mathbb{R}^{3\times 3}. \tag{9}$$

Since $\mathcal{R}(x)$ depends on x only through the Gaussian $\frac{1}{\sqrt{d}} W(x)^\top a$, the dynamics are captured by $B(x)$: there exists $h\colon \mathbb{R}^{3\times 3} \to \mathbb{R}$ such that $\mathcal{R}(x) = h\,(B(x))$. Explicitly, for both (2) and (3), we obtain the same function $h\,(B) = \frac{1}{2}\,(B_{11} - B_{12} - B_{21} + B_{22})$.

The next two assumptions simply require that (i) the risk is a smooth function of this summary (so we may differentiate through the expectation to obtain the gradient) and (ii) the corresponding second-moment quantity controlling the size of the stochastic gradient noise is equally well-behaved.

Assumption 1.10 (Risk Representation)**.** There is an open set $\mathcal{U} \subseteq \mathbb{R}^{3\times 3}$ such that $B(x_0) \in \mathcal{U}$, and provided that $B(x) \in \mathcal{U}$, the map $x \to \mathcal{R}(x) := h\,(B(x))$ is differentiable and satisfies

$$\nabla \mathcal{R}(x) = \mathbb{E}_a\, \nabla_x \Psi(x; a).$$

Furthermore, h is continuously differentiable on $\mathcal{U}$ and its derivative ∇h is α-pseudo-Lipschitz, that is, there is a constant $L(h) > 0$, so that for all $B, \hat{B} \in \mathcal{U}$, it holds:

$$\left\|\nabla h(B) - \nabla h(\hat{B})\right\| \leqslant L \left\|B - \hat{B}\right\| (1 + \|B\|^\alpha + \left\|\hat{B}\right\|^\alpha). \tag{10}$$

We note that the commutation of expectation and gradient holds trivially on $\mathcal{U} = \mathbb{R}^{3\times 3}$ once Ψ is continuously differentiable. Moreover, the choice of the Frobenius norm in (10) is essentially arbitrary, and it can be replaced by the operator norm due to equivalence of norms on $\mathbb{R}^{3\times 3}$.

Our arguments will rely on the evolution of the second moment of the gradient $\mathbb{E}_a\left[\nabla_{r_1} f\,(r)^2\right]$, where we set $r := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x) \\ \beta^* \end{pmatrix}^\top a$. Since this quantity depends on x only through the Gaussian $\frac{1}{\sqrt{d}} W(x)^\top a$,

analogous to the risk, there exists a function $I\colon\ \mathbb{R}^{3\times 3} \to \mathbb{R}$ satisfying $\mathbb{E}_a[\nabla_{r_1} f(r)^2] = I\left(B(x)\right)$. In particular, for the two formulations (2) and (3), the function I is simply $I(B) = 2h\left(B\right)$.

Assumption 1.11 (Pseudo-Lipschitzness of Square Gradients)**.** The function I is α-pseudo-Lipschitz with constant $L(I) > 0$, that is, for all $B, \hat{B} \in \mathcal{U}$,

$$\left|I(B) - I(\hat{B})\right| \leqslant L \left\|B - \hat{B}\right\| (1 + \|B\|^{\alpha} + \left\|\hat{B}\right\|^{\alpha}).$$

Remark 1.12. All the typical losses satisfy the requisite assumptions, including logistic regression and the mean square error (see Appendix F). We also note that Assumptions 1.10 and 1.11 are nearly satisfied for L-smooth objectives f, and a version of the main theorem holds under just this assumption (albeit with a weaker conclusion) (see [Collins-Woodfin et al. [2024a], Appendix B]).

1.3 Algorithm Formulation

We consider the widely used *streaming stochastic gradient descent* (SGD) algorithm. At each iteration k, the algorithm generates a fresh data point $a_{k+1} \sim \mathcal{N}(0, K)$ and updates the iterates $x_k \in \mathbb{R}^{2d}$ using a stepsize γ_k according to the recurrence

$$x_{k+1} = x_k - \gamma_k \nabla_x \Psi(x_k; a_{k+1}), \tag{11}$$

where ∇_x is the usual gradient operator with respect to the x variable. We impose the following assumption on the stepsize, which stipulates that γ_k is a bounded deterministic function of k/d.

Assumption 1.13 (Stepsize)**.** There exists a constant $\bar{\gamma} < \infty$ and a deterministic scalar function $\gamma\colon [0,\infty) \to [0,\infty)$ which is bounded by $\bar{\gamma}$ and satisfies the equation $\gamma_k = \gamma\left(\frac{k}{d}\right)$.

We work in a formulation where entries of the iterates $\{x_k\}$ are bounded, independent of dimension. In high dimensions, the random variable $\frac{1}{\sqrt{d}} \left(\psi(x)\right)^{\top} a$ should not carry dimension dependence, as otherwise the outer function f (which can very well be non-linear) degenerates at infinity.

The functions h and I will allow us to construct closed, deterministic dynamics that describe the high-dimensional limit of stochastic gradient descent. To condense the notation, we use

$$r_k := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix}^{\top} a_{k+1} \in \mathbb{R}^2,$$

so that the SGD update (11) simplifies as follows:

$$x_{k+1} = x_k - \frac{\gamma_k}{\sqrt{d}} \nabla_{r_1} f(r_k) \left(\nabla \psi(x_k)\right)^{\top} a_{k+1}. \tag{12}$$

The stepsize γ_k is scaled so SGD behaves well across dimensions; without the factor of $\sqrt{d}$ in (12), the algorithm would degenerate to pure noise or gradient flow as dimension increases. For fixed dimension, note however that γ_k can be arbitrarily large (if bounded uniformly across dimensions).

2 High-dimensional Diffusion Approximation for SGD

We begin by introducing our stochastic differential equation (SDE), called *homogenized SGD*. Specifically, we approximate SGD by the diffusion $(\mathcal{X}_t)_{t\geqslant 0}$ solving the SDE:

$$\mathrm{d}\mathcal{X}_t = -\gamma(t) d \nabla \mathcal{R}(\mathcal{X}_t)\mathrm{d}t + \gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)} \left(\nabla \psi(\mathcal{X}_t)\right)^{\top} \sqrt{K}\mathrm{d}\mathfrak{B}_t, \tag{13}$$

with initial condition $\mathcal{X}_0 = x_0$ and $\mathfrak{B}_t$ a standard Brownian motion in $\mathbb{R}^d$. In the diffusion coefficient in (13), the quantity $I(B(x))$ captures the effective scalar magnitude of the stochastic gradient noise, while $\left(\nabla \psi(x)\right)^{\top} \sqrt{K}$ describes how this noise propagates through the nonlinear parametrization and the data covariance. This structure is not imposed by covariance matching; it arises from the high-dimensional concentration argument used to identify the limiting drift and diffusion.

Remark 2.1 (Non-diagonal Covariance)**.** The homogenized SDE (13) extends to general covariance K in the mean-squared error setting of Section F.1, with diffusion given by the full stochastic gradient covariance. While the last two terms in the conditional covariance decomposition of the Hessian

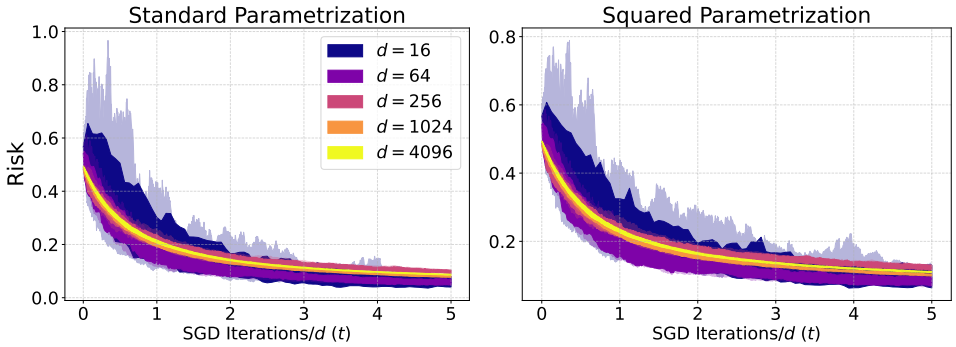

Figure 3: **Risk concentration of SGD and the homogenized SDE under non-diagonal covariance on a diagonal linear network.** As d increases, SGD risk trajectories (opaque) concentrate around the non-diagonal homogenized SDE prediction (14) (transparent), suggesting that the same high-dimensional concentration phenomenon persists beyond diagonal covariance. The covariance matrix K is sampled from a Marchenko–Pastur ensemble. See Appendix G for simulation details.

contribution are negligible for our statistics when K is diagonal, they appear non-negligible for general K, adding a rank-one fluctuation in the high-dimensional limit. Thus, Wick's formula yields

$$\mathbb{E}_a\left[\langle\psi(\mathcal{X}_t)-\beta^\star,a\rangle^2 aa^\top\right]=((\psi(\mathcal{X}_t)-\beta^\star)^\top K(\psi(\mathcal{X}_t)-\beta^\star))K+(K(\psi(\mathcal{X}_t)-\beta^\star))(K(\psi(\mathcal{X}_t)-\beta^\star))^\top.$$

This formally yields the SDE

$$\begin{aligned}\mathrm{d}\mathcal{X}_t =& -\gamma(t)\,d\,\nabla\mathcal{R}(\mathcal{X}_t)\,\mathrm{d}t\\ &+\gamma(t)\left(I\left(B(\mathcal{X}_t)\right)\left(\nabla\psi(\mathcal{X}_t)\right)^\top K\nabla\psi(\mathcal{X}_t)+d\,\nabla\mathcal{R}(\mathcal{X}_t)\left(\nabla\mathcal{R}(\mathcal{X}_t)\right)^\top\right)^{1/2}\mathrm{d}\mathfrak{B}_t,\end{aligned}\tag{14}$$

where $\mathrm{d}\mathfrak{B}_t$ is the differential of a standard Brownian motion in $\mathbb{R}^{2d}$.

The main obstacle to extending our deterministic equivalent to general K is closing the deterministic equations, not the SDE approximation: non-diagonal K couples coordinates through non-commuting resolvent terms, preventing the finite-dimensional PDE closure used here. Nevertheless, numerical experiments suggest that high-dimensional concentration persists beyond the diagonal setting; see 3.

The SDE (13) is useful because it describes the evolution of many quantities of interest by direct Itô calculus. Concretely, given any sufficiently regular statistic φ, Itô's formula gives a decomposition:

$$\mathrm{d}\varphi(\mathcal{X}_t)=\underbrace{\mathcal{L}\varphi(t,\mathcal{X}_t)\,\mathrm{d}t}_{\text{drift}}+\underbrace{\nabla\varphi(\mathcal{X}_t)^\top\sigma(t,\mathcal{X}_t)\,\mathrm{d}\mathfrak{B}_t}_{\text{Brownian / martingale term}},\tag{15}$$

where $\mathcal{L}$ is the drift–diffusion operator (first-order drift contribution plus second-order Itô correction) and σ is the diffusion matrix coefficient of the SDE.

2.1 Relation to Prior SDE and Homogenization Approximations

A key distinction from classical diffusion approaches [Kushner and Clark, 1978, Li et al., 2019, Malladi et al., 2022, Compagnoni et al., 2025] is the regime in which the continuous-time limit becomes accurate. These approaches consider fixed dimension and take $\gamma \to 0$ (after time rescaling), yielding weak convergence to an ODE/SDE as the drift dominates and stochastic fluctuations vanish.

Our goal is a closed finite-stepsize, high-dimensional description of a nonlinear overparameterized model, rather than a general homogenization theorem. The diagonal linear network is a natural minimal target: it admits an exact resolvent closure while capturing the interplay between nonconvex parametrization, stochastic gradients, and finite-stepsize effects. Existing diffusion approximations and general homogenization approaches do not yield a closed deterministic description for the nonlinear feature map $\psi(x)$; this requires tracking a multi-resolvent statistic, leading to the PDE (30).

In the diagonal linear network setting defined in (3), Pesme et al. [2021] follows this diffusion viewpoint and studies SGD via stochastic gradient flow (SGF), a perturbed-gradient-flow SDE whose

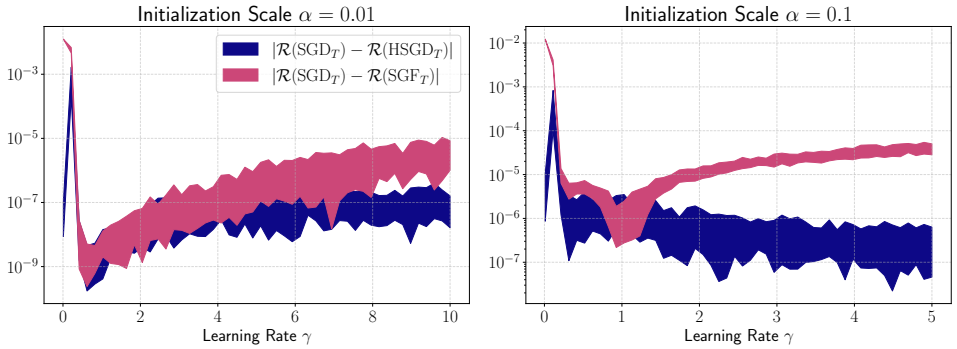

Figure 4: **Risk discrepancy between SGD and its continuous-time approximations on a diagonal linear network.** For each stepsize γ, we report the absolute difference between the empirical risk of SGD after $T \cdot d$ iterations (with $T = 20$) and two approximations: (i) homogenized SGD (13) (blue), and (ii) stochastic gradient flow (16) (pink). As γ increases, HSGD is a more accurate approximation of SGD, whereas SGF degrades. Initialization scale α controls proximity to the saddle point $x = 0$: smaller α yields longer transients before acceleration. See Appendix G for simulation details.

diffusion term is calibrated so that an Euler discretization matches the covariance of the SGD noise:

$$\mathrm{d}\mathcal{X}_t = -d\nabla\mathcal{R}(\mathcal{X}_t)\mathrm{d}t + \sqrt{\gamma(t) I\left(B(\mathcal{X}_t)\right)}\,\left(\nabla\psi(\mathcal{X}_t)\right)^\top \sqrt{K}\,\mathrm{d}\mathfrak{B}_t. \tag{16}$$

The two SDEs are similar in form but describe different limiting regimes. Homogenized SGD (13) has drift and noise proportional to γ, while the covariance-matched diffusion (16) has drift independent of γ and noise of order $\sqrt{\gamma}$. A time change in (13) recovers the scaling of (16), but the approximations differ: the former holds as $d \to \infty$ at fixed γ, while the latter holds as $\gamma \to 0$ at fixed d. Our approach departs from covariance matching: the diffusion coefficient follows from high-dimensional limit theory [Collins-Woodfin et al., 2024a], with accuracy improving as $d \to \infty$. The SDE increasingly matches discrete-time risk and outperforms the covariance-matched model at large stepsizes (Figure 4). Shrinking γ recovers (16), while increasing d sharpens the approximation.

3 High-dimensional Concentration of SGD and its Diffusion Approximation

This section formalizes the main concentration phenomenon: in the high-dimensional regime, both streaming SGD and the homogenized SGD in Section 2 concentrate around the same deterministic dynamics. Under our assumptions, the risk and update terms are captured by a small set of empirical matrix statistics. We encode these in a 3×3 resolvent matrix $S(x, z)$ (with z on a fixed contour Γ). The map $S(x, z)$ determines $B(x)$ via a contour integral, and hence all other statistics of interest.

Let $z = (z_1, z_2, z_3, z_4) \in \mathbb{C}^4$ be such that all resolvents below are well-defined. In Remark 3.3, we specify a product domain on which this condition holds. Define the matrix

$$\Omega(x, z) := R\big(z_1; \mathrm{diag}(u)\big) R\big(z_2; \mathrm{diag}(v)\big) R\big(z_3; \mathrm{diag}(\beta^*)\big) R\big(z_4; K\big) \in \mathbb{C}^{d\times d}, \tag{17}$$

where $R(z; A) := (zI_d - A)^{-1}$ denotes the resolvent of A. We then set

$$S(x, z) := \frac{1}{d} W(x)^\top \Omega(x, z) W(x) \in \mathbb{C}^{3\times 3}, \tag{18}$$

where $W(x) := [\,\psi(x) \,|\, \beta^* \,|\, \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3}$. In particular, we evaluate these quantities either along SGD, writing $S(x_k, z)$ and $B(x_k)$, or along the diffusion, writing $S(\mathcal{X}_t, z)$ and $B(\mathcal{X}_t)$.

To control resolvents uniformly in time (and justify the fixed contour), we work on the high-probability event that iterates remain bounded on the time interval of interest. This non-explosiveness condition is standard in high-dimensional limits and consistent with our numerics, where no explosions are observed. Moreover, Theorem 3.9 shows it holds for all stepsizes γ below a small explicit constant.

Assumption 3.1 (Non-explosiveness)**.** There exists $M > 0$ independent of d such that the initialization satisfies $\|x_0\|_\infty \leqslant M$, and such that, for every fixed $T > 0$, with overwhelming probability:

$$\sup_{0\leqslant k\leqslant \lfloor Td\rfloor} \|x_k\|_\infty \leqslant M \qquad \text{and} \qquad \sup_{0\leqslant t\leqslant T} \|\mathcal{X}_t\|_\infty \leqslant M.$$

Remark 3.2. Assumption 3.1 is stated in a general form. In Appendix E, we verify it in the isotropic squared-parametrization setting for constant stepsize γ. Indeed, for sufficiently small γ, we establish high-probability exponential decay of the risk. Risk integrability then yields uniform-in-time bounds on the coordinates, which by the standard SDE continuation criterion rule out finite-time explosion.

Remark 3.3. Under Assumption 3.1, the spectra of $\mathrm{diag}(u)$ and $\mathrm{diag}(v)$ remain inside the circle $C_{2M}(0) := \{z \in \mathbb{C} : \ |z| = 2M\}$ over the time intervals considered. Throughout the paper, we fix the product contour $\Gamma = \Gamma_1 \times \Gamma_2 \times \Gamma_3 \times \Gamma_4 \subset \mathbb{C}^4$ given by $\Gamma_1 = \Gamma_2 = C_{2M}(0)$ and

$$\Gamma_3 = \{z_3 : |z_3| = \max\{1, 2\|\beta^*\|_\infty\}\}, \qquad \Gamma_4 = \{z_4 : |z_4| = \max\{1, 2\|K\|_{\mathrm{op}}\}\}.$$

The contours are chosen in such a way that each Γ_i encloses the relevant spectrum with a fixed positive margin. Thus, by the Cauchy integral formula, B can be recovered from S, and hence

$$B(x_k) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \, S(x_k, z) \, \mathrm{d}z, \quad \text{and} \quad B(\mathfrak{X}_t) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \, S(\mathfrak{X}_t, z) \, \mathrm{d}z. \tag{19}$$

In order to derive deterministic dynamics, we pass to a rescaled continuous-time parameter:

$$k \text{ iterations of SGD } = \lfloor td \rfloor, \quad \text{where } t \geqslant 0.$$

This time change is natural: as problem size grows, more SGD iterations are needed to make equivalent progress. This scaling law ensures training dynamics are comparable across problem sizes.

Deterministic Limit Dynamics (PDE). Next, for each $z \in \Gamma \subset \mathbb{C}^4$, let $\mathscr{S}(t, z) \in \mathbb{C}^{3\times 3}$ denote the (local) solution of the partial integro-differential equation

$$\partial_t \mathscr{S}(t, \cdot) = \mathcal{F}\big(\cdot, \mathscr{S}(t, \cdot)\big), \qquad \mathscr{S}(0, z) = S(x_0, z), \tag{20}$$

where the function $\mathcal{F}$ is defined in (29) and involves contour integrals around Γ and derivatives in z of $\mathscr{S}$. Note that the full expression of $\mathcal{F}$ is not needed for the stability and net arguments (C.4 and D.1). In contrast, for the concentration results (proven via Propositions D.2 and D.4), the explicit formula for $\mathcal{F}$, which can be found in the displayed equations following Remark D.3, plays an essential role. We define the associated deterministic quantities

$$\mathscr{B}(t) := \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \, \mathscr{S}(t, z) \, \mathrm{d}z, \qquad \mathscr{R}(t) := h\left(\mathscr{B}(t)\right), \qquad \mathscr{I}(t) := I\left(\mathscr{B}(t)\right). \tag{21}$$

Since (20) is a fully nonlinear, nonlocal parabolic PDE, we do not pursue global well-posedness. We denote by $\mathscr{S}(t, z)$ any solution defined up to the first exit time of $\mathscr{B}(t)$ from the domain $\mathcal{U}$ (or blow-up), when it exists. In the simulations presented in this paper, we solve (20) numerically.

The next theorem states the concentration result: both sequences $B(x_{\lfloor td \rfloor})$ (SGD) and $B(\mathfrak{X}_t)$ (homogenized SGD) track the same deterministic curve $\mathscr{B}(t)$ while the dynamics remain in $\mathcal{U}$.

Theorem 3.4 (Learning Curves)**.** *Suppose Assumptions 1.4, 1.5, 1.8, 1.9, 1.10, 1.11, 1.13, and 3.1 hold. Let $\mathscr{S}(t, z)$ solve (20) and define $\mathscr{B}(t)$ by (21).*

(i) Streaming SGD. *Let ϑ be the first exit time of $\mathscr{B}(t)$ or $B(x_{\lfloor td \rfloor})$ from $\mathcal{U}$. Then there exists an $\varepsilon > 0$ such that for any $T > 0$, it holds:*

$$\sup_{0 \leqslant t \leqslant T \wedge \vartheta} \left\| \mathscr{B}(t) - B(x_{\lfloor td \rfloor}) \right\| \leqslant d^{-\varepsilon} \quad \textit{w.o.p.}$$

(ii) Homogenized SDE. *For $\eta > 0$, let $\mathcal{U}_\eta := \left\{ B \in \mathcal{U} : \ \inf_{\widehat{B} \notin \mathcal{U}} \|B - \widehat{B}\| \geqslant \eta \right\}$, and let ϑ be the first exit time of $\mathscr{B}(t)$ or $B(\mathfrak{X}_t)$ from $\mathcal{U}_\eta$. Then for some $\varepsilon > 0$, for any $T > 0$, it holds:*

$$\sup_{0 \leq t \leq T \wedge \vartheta} \|\mathscr{B}(t) - B(\mathfrak{X}_t)\| \leqslant d^{-\varepsilon} \quad \textit{w.o.p.}$$

Remark 3.5. Theorem 3.4 extends the concentration framework in Collins-Woodfin et al. [2024a] to a broader class of risks, including multi-class logistic regression and mean-squared error over diagonal linear networks. The main difficulty comes from the nonlinear function ψ, through which the data enter as $\langle a, \psi(x) \rangle$ rather than $\langle a, x \rangle$. Applying Itô's formula produces products of resolvents involving u, v, β^*, and K, breaking the finite-dimensional ODE closure available in least-squares. The deterministic closure takes the form of a PDE for $\mathscr{S}(t, z)$. Commutativity allows these products to be reordered into a fixed multi-resolvent statistic; without it, noncommuting resolvent terms fall outside the closure provided by $S(x, z)$. In our setting, this assumption requires K to be diagonal.

3.1 Other Statistics

We have shown that $B(x_{\lfloor td \rfloor})$ and $B(\mathcal{X}_t)$ concentrate around a deterministic limit. We now track statistics $\varphi(x)$ along SGD and homogenized SGD and show they follow the same deterministic dynamics in the high-dimensional limit. The relevant statistics satisfy the following assumption:

Assumption 3.6. The statistic $\varphi\colon \mathbb{R}^{2d} \to \mathbb{R}$ satisfies a composite structure,

$$\varphi(x) = g\left(\frac{1}{d} W^\top q_1(\mathrm{diag}(u)) q_2(\mathrm{diag}(v)) q_4(K) W\right),$$

where $g\colon \mathbb{R}^{3\times 3} \to \mathbb{R}$ is α-pseudo-Lipschitz on $\mathcal{U}$ and q_1, q_2, q_4 are polynomials.

This assumption covers typical statistics, including risk, curvature, squared norms, and estimation errors not involving the covariance K, as well as ℓ^2-regularized objectives and higher-order derivatives of the risk. To describe the deterministic equivalent of φ, we require its continuous-time analog:

$$\phi(t) := g\left(\frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4)\, \mathscr{S}(t,z)\, \mathrm{d}z\right). \tag{22}$$

The following theorem shows that $\phi(t)$–a deterministic function of a single variable–governs the evolution of φ both along the SGD and the homogenized SGD trajectories.

Theorem 3.7. *Suppose Assumptions 1.4, 1.5, 1.8, 1.9, 1.10, 1.11, 1.13, and 3.1 hold and $\mathcal{U} = \mathbb{R}^{3\times 3}$. For any φ satisfying Assumption 3.6, $\varepsilon \in (0, 1/2)$, and $T > 0$, there is a constant C so that it holds:*

$$\sup_{0 \leqslant t \leqslant T} \left(\left|\varphi(x_{\lfloor td \rfloor}) - \phi(t)\right| + \left|\varphi(\mathcal{X}_t) - \phi(t)\right|\right) \leqslant C d^{-\varepsilon} \quad \textit{w.o.p.} \tag{23}$$

Intuition of the proof. We introduce S in (18) because the resolvents in Ω express the admissible polynomial statistics through contour integrals via Cauchy's integral formula. It thus suffices to prove uniform concentration of S on the fixed product contour from Remark 3.3. The main step is to show that $S(x_{\lfloor td \rfloor}, z)$ and $S(\mathcal{X}_t, z)$ approximately solve the same deterministic equation, with solution $\mathscr{S}(t,z)$. Commutativity closes the Doob/Itô expansions in terms of S, yielding the PDE (30); stability then gives convergence to $\mathscr{S}$. B and any admissible statistic φ follow by contour integration.

Remark 3.8. While the PDE (20) gives the most complete description of SGD, the SDE (13) is often more practical: it is cheaper to simulate (e.g., via Euler–Maruyama) and yields the evolution of admissible statistics φ. Applying Itô's formula to $\varphi(\mathcal{X}_t)$ gives a drift, including the Itô correction, and a martingale term. The key is that $\mathcal{L}\varphi(\mathcal{X}_t)$ can be rewritten in terms of the same class of statistics, possibly enlarged for closure, yielding a closed deterministic system. In the high-dimensional limit, concentration makes the martingale term negligible, so $\phi(t)$ evolves according to the full Itô drift.

As an application of our result, we show that for stepsizes below a threshold, the SDE iterates remain bounded and the risk decays exponentially fast to zero. By Theorem 3.7 the same holds for SGD.

Theorem 3.9 (Exponential Decay)**.** *Consider homogenized SGD (4) under squared parametrization (3), initialized at $\mathcal{U}_0 = \mathcal{V}_0 = \mathbb{1}_d$. Then for sufficiently small stepsize $\gamma > 0$, and for any $\delta \in (0,1)$, there are constants $C, \mu > 0$ (independent of t or d) such that with probability at least $1-\delta$, it holds:*

$$\mathcal{R}(\mathcal{X}_t) \leqslant C e^{-\mu t} \qquad \textit{for all } t \geqslant 0. \tag{24}$$

In particular, Assumption 3.1 holds.

Potential further applications. Beyond convergence, our framework can analyze finer statistics of the dynamics. One example is $\varphi(x) = \frac{1}{d}\mathrm{Tr}(\nabla^2 \mathcal{R}(x))$, a proxy for the *average local curvature* (or "sharpness") [Ding et al., 2024, Wen et al., 2023]. Tracking $\mathcal{R}(x)$ and $\varphi(x)$ through their deterministic limits enables the study of *progressive sharpening* [MacDonald et al., 2023, Yoo et al., 2025, Liu et al., 2026, Kalra et al., 2026, Cohen et al., 2021, Jastrzębski et al., 2019], where risk may temporarily increase while curvature decreases, suggesting movement toward flatter regions (Figure 2). This allows us to quantify how these phases depend on the stepsize and identify the corresponding regimes from the deterministic equations. The framework also enables the study of *implicit bias* and how problem-dependent quantities, such as the data covariance, influence the effective stepsize of SGD.

Acknowledgments and Disclosure of Funding

Research of B. García Malaxechebarría was in part funded by NSF DMS 2023166 (NSF TRIPODS II). C. Paquette was supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada, NSERC CREATE grant Interdisciplinary Math and Artificial Intelligence Program (INTER-MATH-AI), Google x Mila research grant, Fonds de recherche du Quebec Nature et technologies (FRQNT) NOVA Grant, and CIFAR AI Catalyst Grant. Additionally C. Paquette has a 20% part-time employment at Google DeepMind. The work of M. Fazel is supported in part by awards NSF TRIPODS II DMS 2023166, NSF CCF 2212261, NSF CCF 2312775, and the Moorthy Family Professorship at UW. She also works part-time at Amazon Inc. as an Amazon Scholar. Research of D. Drusvyatskiy was supported by NSF DMS-2306322, NSF DMS-2023166, and AFOSR FA9550-24-1-0092 awards.

References

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A Modern Look at the Relationship between Sharpness and Generalization. In *Proceedings of the 40th International Conference on Machine Learning*, pages 840–902. PMLR, 2023a.

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. SGD with Large Step Sizes Learns Sparse Features. In *Proceedings of the 40th International Conference on Machine Learning*, pages 903–925. PMLR, 2023b.

Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Escaping mediocrity: How two-layer networks learn hard generalized linear models with SGD. In *OPT2023: 15th Annual Workshop on Optimization for Machine Learning*, 2023a. arXiv preprint.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 1199–1227. PMLR, 2023b.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent. In *Proceedings of the 38th International Conference on Machine Learning*, pages 468–477. PMLR, 2021.

Krishnakumar Balasubramanian, Promit Ghosal, and Ye He. High-dimensional scaling limits and fluctuations of online least-squares SGD with smooth covariance. *The Annals of Applied Probability*, 35(5), 2025. ISSN 1050-5164. doi: 10.1214/24-AAP2123.

M Biehl and P Riegler. On-Line Learning with a Perceptron. *Europhysics Letters (EPL)*, 28(7):525–530, 1994. ISSN 0295-5075, 1286-4854. doi: 10.1209/0295-5075/28/7/012.

M Biehl and H Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643–656, 1995. ISSN 0305-4470, 1361-6447. doi: 10.1088/0305-4470/28/3/018.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data, 2021.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*, 2021.

Elizabeth Collins–Woodfin and Elliot Paquette. High-dimensional limit of one-pass SGD on least squares. *Electronic Communications in Probability*, 29(none), 2024. ISSN 1083-589X. doi: 10.1214/23-ECP571.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the High-dimensional notes: An ODE for SGD learning dynamics on GLMs and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024a. ISSN 2049-8772. doi: 10.1093/imaiai/iaae028.

Elizabeth Collins-Woodfin, Inbar Seroussi, Begoña García Malaxechebarría, Andrew Mackenzie, Elliot Paquette, and Courtney Paquette. The High Line: Exact Risk and Learning Rate Curves of Stochastic Adaptive Learning Rate Algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b.

Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive Methods through the Lens of SDEs: Theoretical Insights on the Role of Noise. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the Landscape Boosts the Signal for SGD: Optimal Sample Complexity for Learning Single Index Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 752–784. Curran Associates, Inc., 2023.

Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 2024. ISSN 2049-8772. doi: 10.1093/imaiai/iaae009.

Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics. Wiley-Interscience, Hoboken, NJ, 1986. ISBN 978-0-470-31665-8 978-0-470-31732-7. doi: 10.1002/9780470316658.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over Diagonal Linear Networks: Implicit bias, Large Stepsizes and Edge of Stability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29406–29448. Curran Associates, Inc., 2023.

Cédric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous Dynamical Mean-Field Theory for Stochastic Gradient Descent Methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024. doi: 10.1137/23M1594388.

Sebastian Goldt, Madhu Advani, Andrew Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model. *Physical Review X*, 10(4):041044, 2020. ISSN 2160-3308. doi: 10.1103/PhysRevX.10.041044.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pages 426–471. PMLR, 2022.

Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape Matters: Understanding the Implicit Bias of the Noise Covariance. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2315–2357. PMLR, 2021.

Fengxiang He, Tongliang Liu, and Dacheng Tao. Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amost Storkey. On the Relation Between the Sharpest Directions of DNN Loss and the SGD Step Length. In *International Conference on Learning Representations*, 2019.

Dayal Singh Kalra, Jean-Christophe Gagnon-Audet, Andrey Gromov, Ishita Mediratta, Kelvin Niu, Alexander H. Miller, and Michael Shvartsman. A Scalable Measure of Loss Landscape Curvature for Analyzing the Training Dynamics of LLMs, 2026.

Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. The statistical complexity of early-stopped mirror descent. *Information and Inference: A Journal of the IMA*, 12(4):3010–3041, 2023. ISSN 2049-8772. doi: 10.1093/imaiai/iaad047.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*, 2017.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An Alternative View: When Does SGD Escape Local Minima? In *Proceedings of the 35th International Conference on Machine Learning*, pages 2698–2707. PMLR, 2018.

Harold J. Kushner and Dean S. Clark. *Weak Convergence for Unconstrained Systems*, volume 26, pages 106–157. Springer New York, New York, NY, 1978. ISBN 978-0-387-90341-5 978-1-4684-9352-8. doi: 10.1007/978-1-4684-9352-8_4.

Kiwon Lee, Andrew Nicholas Cheng, Elliot Paquette, and Courtney Paquette. Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

Liming Liu, Zixuan Zhang, Simon Shaolei Du, and Tuo Zhao. A Minimalist Example of Edge-of-Stability and Progressive Sharpening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Lachlan Ewen MacDonald, Jack Valmadre, and Simon Lucey. On progressive sharpening, flat minima and generalisation, 2023.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and Scaling Rules for Adaptive Gradient Algorithms. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 9540–9550. Curran Associates, Inc., 2020.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A. Erdogdu. Neural Networks Efficiently Learn Low-Dimensional Representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023.

Mor Shpigel Nacson, Kavya Ravichandran, Nathan Srebro, and Daniel Soudry. Implicit Bias of the Step Size in Linear Diagonal Neural Networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16270–16295. PMLR, 2022.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3548–3626. PMLR, 2021.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, 214(1-2):1–90, 2025. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-024-02171-3.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit Bias of SGD for Diagonal Linear Networks: A Provable Benefit of Stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pages 29218–29230. Curran Associates, Inc., 2021.

David Saad and Sara Solla. Dynamics of On-Line Gradient Descent Learning for Multilayer Neural Networks. In *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995a.

David Saad and Sara A. Solla. Exact Solution for On-Line Learning in Multilayer Neural Networks. *Physical Review Letters*, 74(21):4337–4340, 1995b. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.74.4337.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Aditya Varre, Margarita Sagitova, and Nicolas Flammarion. SGD vs GD: Rank Deficiency in Linear Networks. In *High-Dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit Regularization for Optimal Sparse Recovery. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Loucas Pillaud Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2127–2159. PMLR, 2022.

Chuang Wang, Jonathan Mattingly, and Yue M. Lu. Scaling Limit: Exact and Tractable Analysis of Online Learning Algorithms with Applications to Regularized Regression and PCA, 2017.

Chuang Wang, Hong Hu, and Yue Lu. A Solvable High-Dimensional Model of GAN. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How Sharpness-Aware Minimization Minimizes Sharpness? In *The Eleventh International Conference on Learning Representations*, 2023.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

Geonhui Yoo, Minhak Song, and Chulhee Yun. Understanding Sharpness Dynamics in NN Training with a Minimalist Example: The Effects of Dataset Difficulty, Depth, Stochasticity, and More. In *Forty-Second International Conference on Machine Learning*, 2025.

Yuki Yoshida and Masato Okada. Data-Dependence of Plateau Phenomenon in Learning with Neural Network — Statistical Mechanical Analysis. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: Spikes in the training loss and their impact on generalization through feature learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, Vienna, Austria, 2024. JMLR.org.

Supplementary Material: High-dimensional Limit of SGD for Diagonal Linear Networks

Broader Impact Statement. The work presented in this paper is foundational research and it is not tied to any particular application. The set-up is on a simple well-studied high-dimensional diagonal neural network with synthetic data and solved using known algorithms such as SGD. We present deterministic dynamics for the training loss. The results are theoretical and we do not anticipate any direct ethical and societal issues. We believe the results will be used by machine learning practitioners and we encourage them to use it to build a more just, prosperous world.

Outline of the paper. The remainder of the article is structured as follows:

1. Appendix B collects notation and auxiliary tools used throughout the proofs. It fixes our conventions for complex-valued tensor products, coordinate contractions, and tensor norms; records derivative computations for the special functions ψ, Ψ, and S appearing in the proof of Theorem 3.7; and recalls the concentration and pseudo-Lipschitz estimates used in the probabilistic arguments.
2. Appendix C develops the main dynamical argument. It introduces the partial integro-differential equation (30) and the notion of approximate solutions, proves a stability principle for these solutions, and applies it to the resolvent statistic S along SGD and homogenized SGD. This yields Theorem C.7; the result is then transferred to general statistics satisfying Assumption 3.6, yielding Theorem C.9 and its corollaries.
3. Appendix D proves that the resolvent statistics $t \mapsto S(x_{\lfloor td \rfloor}, \cdot)$ and $t \mapsto S(\mathcal{X}_t, \cdot)$, associated respectively with SGD and homogenized SGD (13), are approximate solutions of the partial integro-differential equation (30). The proof uses Doob/Itô decompositions, a net argument over the fixed contour, and martingale and Taylor-error bounds.
4. Appendix E studies the homogenized SDE in the isotropic squared-parameterization setting. It introduces an empirical entropy adapted to the coordinatewise dynamics, proves an exact entropy SDE and barrier estimates, and uses an exponential supermartingale argument to obtain high-probability global existence and exponential decay of the risk. The section also records consequences such as risk integrability and uniform separation from the saddle.
5. Appendix F presents key examples illustrating our concentration risk framework.
6. Appendix G provides additional details on the numerical simulations used to produce the figures in the main text.

Contents

1 Introduction **1**

1.1 Literature Review 3

1.2 High-dimensional Model Structure 4

1.3 Algorithm Formulation 6

2 High-dimensional Diffusion Approximation for SGD **6**

2.1 Relation to Prior SDE and Homogenization Approximations 7

3 High-dimensional Concentration of SGD and its Diffusion Approximation **8**

3.1 Other Statistics 10

A Additional Related Work **17**

A.1 Continuous-time and Stochastic-process Approximations of SGD 17

A.2 High-dimensional Deterministic Limits for SGD Dynamics 17

A.3 Implicit Bias, Flatness, and the Role of Stepsize 17

A.4 Diagonal Linear Neural Networks . . . 18

B Notation and Preliminaries **18**

B.1 Tensor Products and Contractions . . . 18

B.2 Norms on Tensors . . . 19

B.3 Derivative Identities for Special Statistics . . . 20

B.4 Concentration and Pseudo-Lipschitz Estimates . . . 23

C Dynamics of the Resolvent Statistic **24**

C.1 Approximate Solutions and Stability . . . 25

C.2 Main Argument of the Proof: Concentration of the Resolvent Statistic S . . . 31

C.3 Concentration of General Statistics . . . 34

D SGD and Homogenized SGD are Approximate Solutions **35**

D.1 Homogenized SGD Under the Resolvent Statistic S . . . 39

D.1.1 Doob Decomposition for Homogenized SGD . . . 39

D.1.2 $S(\mathcal{X}_t, z)$ is an Approximate Solution, Proof of Proposition D.2 . . . 40

D.2 SGD Under the Resolvent Statistic S . . . 44

D.2.1 Doob Decomposition for SGD . . . 45

D.2.2 $S\left(x_{\lfloor td \rfloor}, z\right)$ is an Approximate Solution, Proof of Proposition D.4 . . . 48

D.3 Error Bounds . . . 49

D.3.1 Homogenized SGD Martingale Error . . . 50

D.3.2 Bounds on the Martingales $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Hess}}$ . . . 52

D.3.3 Bounds on the Higher Order Error Term in the Taylor Expansion, $\mathcal{E}_t^{\mathrm{High}}$ . . . 57

D.3.4 Bounds on the Lower Order Terms in the Hessian, $\mathcal{E}_t^{\mathrm{Hess}}$ . . . 59

E Entropy Barrier and High-probability Exponential Decay **60**

E.1 Exact Product Identity and Logarithmic Coordinates . . . 61

E.2 Exact Dynamics of the Empirical Entropy . . . 63

E.3 Barrier Estimates . . . 64

E.4 Exponential Decay with Explicit Confidence . . . 66

E.5 High-probability Decay at Prescribed Confidence . . . 70

E.6 Risk Integrability and Dynamical Consequences . . . 71

E.6.1 Non-explosion via Risk Integrability . . . 71

E.6.2 Uniform Separation from the Saddle via Risk Integrability . . . 72

F Examples **73**

F.1 Mean-squared Error . . . 73

F.2 Binary Logistic Regression . . . 74

G Numerical Simulation Details **74**

A Additional Related Work

A.1 Continuous-time and Stochastic-process Approximations of SGD

A common analytical approach is to replace discrete-time SGD with continuous-time dynamics. Gradient flow arises as the infinitesimal step-size limit, but it suppresses both finite-step effects and stochasticity. Stochastic gradient flow and diffusion approximations incorporate noise and, in some regimes, can predict solution-selection phenomena not captured by gradient flow [Pesme et al., 2021]. Related modeling ideas include studying SGD with label noise as a proxy for stochasticity during the central phase of training, when the risk is still significantly above zero [Vivien et al., 2022]. When stepsizes act as an implicit regularizer, this effect has been related to the magnitude and structure of the parameter-dependent noise in SGD, including biases toward low-rank structure in the hidden layers [Andriushchenko et al., 2023b]. In high-dimensional settings, SDE-based analyses can also reveal additional regularization mechanisms, such as repulsive interactions in eigenvalue dynamics that promote rank deficiency [Varre et al., 2024].

Beyond diffusion approximations, dynamical mean field theory (DMFT) provides a powerful (though often technically involved) route to tracking training dynamics in high dimensions. For instance, DMFT has been used to describe SGD in Gaussian mixture classification [Mignacco et al., 2020]. Related work develops rigorous analyses for closely related dynamics and batching regimes [Celentano et al., 2021, Gerbelot et al., 2024].

A.2 High-dimensional Deterministic Limits for SGD Dynamics

There is a long line of work deriving deterministic dynamical descriptions of learning in high-dimensional teacher–student and random-feature models, particularly under isotropic Gaussian assumptions. Early studies analyzed multi-index/soft-committee models and derived ODE characterizations that expose the role of stepsize in convergence [Saad and Solla, 1995a, Biehl and Riegler, 1994, Biehl and Schwarze, 1995, Saad and Solla, 1995b]. More recent work emphasizes that generalization depends on an interplay between algorithm, architecture, and data distribution [Goldt et al., 2019], and develops general techniques (e.g., martingale arguments) for proving high-dimensional limits [Wang et al., 2017, 2019]. Comparisons across limiting regimes and refinements of these ODE descriptions are investigated in Arnaboldi et al. [2023b]. Other variants highlight how modifying the geometry or constraints of the dynamics can lead to signal-locking and dimension-robust behavior [Arous et al., 2022, Arnaboldi et al., 2023a, Damian et al., 2023], and how streaming or multi-pass SGD can admit deterministic characterizations via integral equations or homogenization methods [Mousavi-Hosseini et al., 2023, Paquette et al., 2021, 2025, Lee et al., 2022].

While many results assume identity covariance, non-identity covariance can qualitatively change training dynamics and is important when the population covariance is unknown. Prior work derives equations of motion under Gaussian-equivalence principles for multi-index/random-feature models [Goldt et al., 2020] and extends them to deeper architectures [Goldt et al., 2022]. Covariance structure has also been linked to plateau phenomena and slow phases in learning [Yoshida and Okada, 2019]. Recent work derives exact high-dimensional limits for linear regression and analyzes how non-identity covariance reshapes the optimization landscape and convergence [Collins-Woodfin et al., 2024a,b]; extensions to adaptive algorithms and related effects are studied in [Balasubramanian et al., 2025].

A.3 Implicit Bias, Flatness, and the Role of Stepsize

A related theme is the implicit bias of gradient-based optimization. In separable logistic regression, gradient descent converges in direction to the max-margin classifier [Soudry et al., 2018], and in modern overparameterized settings SGD is often viewed as providing an implicit regularization mechanism that can improve generalization with little or no explicit regularization [Zhang et al., 2017]. One influential hypothesis links this behavior to SGD's preference for flatter minima [He et al., 2019] and to empirical correlations between flatness and generalization [Keskar et al., 2017]. Related evidence appears in structured recovery problems, where flat minima can coincide with exact recovery under suitable conditions [Ding et al., 2024]. Stepsize and batch size are key levers in this picture: larger stepsizes and smaller batch sizes have been argued to encourage exploration of flatter regions [Jastrzębski et al., 2019], and the batch-size-to-learning-rate ratio can affect generalization [He et al., 2019]. Initialization also plays an important role in determining the implicit bias: the

initialization scale can control the transition between kernel/lazy and rich/active regimes [Woodworth et al., 2020], while the relative shape of the initialization can further affect the limiting solution selected by gradient methods [Azulay et al., 2021]. Other proposed mechanisms include landscape-smoothing interpretations [Kleinberg et al., 2018], random-walk models on random landscapes [Hoffer et al., 2017], and explanations based on batch variability and "catapult" effects [Zhu et al., 2024]. At the same time, recent evidence cautions that sharpness and flatness can be tightly entangled with optimization hyperparameters such as the stepsize, and need not correlate monotonically with generalization across settings [Andriushchenko et al., 2023a].

A.4 Diagonal Linear Neural Networks

Diagonal linear networks have emerged as a simplified yet expressive proxy for studying implicit bias and solution selection in overparameterized models. They enable precise comparisons between SGD and gradient descent/flow, including settings where SGD exhibits favorable implicit bias [Pesme et al., 2021]. In sparse feature learning, empirical work highlights that large stepsizes alone may be insufficient for sparsification without stochasticity [Andriushchenko et al., 2023b], and complementary theory shows that large stepsizes can systematically improve SGD while harming gradient descent in sparse regression with diagonal linear networks [Even et al., 2023].

B Notation and Preliminaries

This appendix collects notation and auxiliary facts used throughout the proofs. We first fix basic notation and conventions for tensor products, coordinate contractions, and tensor norms for complex-valued tensors. We then record derivative identities for ψ, Ψ, and S, which are used in the derivation of the limiting dynamics. Finally, we recall the concentration and pseudo-Lipschitz estimates needed for the probabilistic bounds.

Basic notation. Throughout the paper, e_i denotes the i-th canonical coordinate vector, with its dimension inferred from context. Additionally, the matrix E_{ij} is defined as the outer product $e_i \cdot e_j^\top$.

For $x = (u, v)$, recall that each coordinate of the inner function ψ is given by

$$\psi_i(u, v) = \begin{pmatrix} u_i & v_i \end{pmatrix} \mathcal{Q} \begin{pmatrix} u_i \\ v_i \end{pmatrix} + l^\top \begin{pmatrix} u_i \\ v_i \end{pmatrix} + c,$$

where $\mathcal{Q} = \begin{pmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}$ is symmetric, $l = (l_1, l_2)^\top \in \mathbb{R}^2$, and $c \in \mathbb{R}$.

Equivalently, we have

$$\psi_i(u, v) = q_{11}u_i^2 + 2q_{12}u_iv_i + q_{22}v_i^2 + l_1u_i + l_2v_i + c.$$

B.1 Tensor Products and Contractions

We briefly fix the tensor-contraction conventions used throughout the proofs. The optimization variables are real, but several quantities, such as $S(x, z)$, are complex-valued because of the spectral parameter z. Consequently, we will use complex-valued tensors whose derivative directions are indexed by real coordinate spaces such as $\mathbb{R}^d$.

We order tensor indices so that derivative or ambient coordinates appear first, and observable coordinates appear last. For example, since $S(x, z) \in \mathbb{C}^{3\times 3}$, we identify

$$\nabla_u S(x, z) \in \mathbb{C}^{d\times 3\times 3} \cong \mathbb{C}^d \otimes \mathbb{C}^{3\times 3},$$

where the first index corresponds to the derivative with respect to u.

We use $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$ to denote contraction over the indicated d-dimensional coordinate. This contraction is bilinear and is defined by summing over the contracted coordinate, with no complex conjugation. Thus, if

$$A \in \mathbb{C}^{d\times m_1\times\cdots\times m_k}, \qquad B \in \mathbb{C}^{d\times n_1\times\cdots\times n_\ell},$$

then the expression

$$\langle A, B\rangle_{\mathbb{R}^d} \in \mathbb{C}^{m_1\times\cdots\times m_k\times n_1\times\cdots\times n_\ell}$$

is defined by

$$\big(\langle A, B\rangle_{\mathbb{R}^d}\big)_{\alpha_1,\dots,\alpha_k,\beta_1,\dots,\beta_\ell} := \sum_{i=1}^{d} A_{i,\alpha_1,\dots,\alpha_k} B_{i,\beta_1,\dots,\beta_\ell}.$$

In words, the contracted axes are summed over, and the uncontracted axes of the first tensor are listed before the uncontracted axes of the second tensor.

For instance, if

$$\nabla_u S(x,z) \in \mathbb{C}^{d\times 3\times 3}, \qquad \nabla_u \mathcal{R}(x) \in \mathbb{R}^d,$$

then the matrix

$$\langle \nabla_u S(x,z), \nabla_u \mathcal{R}(x)\rangle_{\mathbb{R}^d} \in \mathbb{C}^{3\times 3}$$

has entries

$$[\langle \nabla_u S(x,z), \nabla_u \mathcal{R}(x)\rangle_{\mathbb{R}^d}]_{ab} = \sum_{i=1}^{d} \partial_{u_i} S_{ab}(x,z)\, \partial_{u_i} \mathcal{R}(x).$$

Similarly, if

$$\nabla_u^2 S(x,z) \in \mathbb{C}^{d\times d\times 3\times 3}, \qquad M \in \mathbb{R}^{d\times d},$$

then contraction over the two derivative coordinates gives

$$\langle \nabla_u^2 S(x,z), M\rangle_{\mathbb{R}^{d\times d}} \in \mathbb{C}^{3\times 3},$$

with entries

$$\big[\langle \nabla_u^2 S(x,z), M\rangle_{\mathbb{R}^{d\times d}}\big]_{ab} = \sum_{i,j=1}^{d} \partial_{u_i}\partial_{u_j} S_{ab}(x,z)\, M_{ij}.$$

We reserve the unsubscripted notation $\langle\cdot,\cdot\rangle$ for the full Hilbert–Schmidt tensor inner product. Thus, for complex matrices or tensors of the same shape,

$$\langle A, B\rangle := \sum_{\alpha} \overline{A_\alpha} B_\alpha,$$

where α ranges over all tensor indices. In particular, for matrices,

$$\langle A, B\rangle = \mathrm{Tr}(A^* B),$$

where $A^* = \overline{A}^{\top}$ denotes the conjugate transpose. The corresponding unsubscripted norm is

$$\|A\| := \sqrt{\langle A, A\rangle}.$$

In contrast, a subscript on the bracket indicates a partial contraction over the specified coordinate space. For example, $\langle A, B\rangle_{\mathbb{R}^d}$ denotes contraction over the $\mathbb{R}^d$ coordinate. These partial contractions are bilinear coordinate contractions used in derivative and generator computations, and no complex conjugation is applied unless explicitly stated. Thus $\langle\cdot,\cdot\rangle_{\mathbb{R}^d}$ should be read as a contraction label, not as a Hermitian inner product on $\mathbb{C}^d$.

B.2 Norms on Tensors

We now define the tensor norms used throughout the paper. Unless otherwise specified, $\|\cdot\|$ denotes the Hilbert-space norm: the Euclidean norm for vectors and the Hilbert–Schmidt/Frobenius norm for matrices and tensors. As above, $\langle A, B\rangle$ denotes the full Hilbert–Schmidt inner product, and

$$\|A\| := \sqrt{\langle A, A\rangle}$$

denotes the associated Hilbert-space norm. Thus, for matrices, $\|A\|$ is the Frobenius norm unless another subscript, such as $\|\cdot\|_{\mathrm{op}}$, is specified. We use subscripts such as $\|\cdot\|_{\mathrm{op}}$, $\|\cdot\|_\sigma$, and $\|\cdot\|_*$ for the operator, injective, and nuclear norms, respectively.

For a matrix $A \in \mathbb{C}^{d\times d}$, the operator norm admits the variational representation

$$\|A\|_{\mathrm{op}} = \sup_{\substack{\|y\|_2=1\\ \|z\|_2=1}} \big|y^* A z\big|.$$

The injective tensor norm is the natural higher-order analogue of this formula. Let

$$A \in V_1 \otimes V_2 \otimes \cdots \otimes V_k.$$

We define

$$\|A\|_\sigma := \sup_{\substack{\|y_i\|_{V_i}=1 \\ i=1,\dots,k}} |\langle A,\, y_1 \otimes y_2 \otimes \cdots \otimes y_k \rangle| \,.$$

Equivalently, $\|A\|_\sigma$ is the largest correlation of A with a unit simple tensor. This norm is also known as the *injective tensor norm*. In the case $k = 2$, it reduces to the usual operator norm, up to the standard identification of matrices with order-two tensors.

By Cauchy–Schwarz, the Hilbert-space norm also has the variational representation

$$\|A\| = \sup_{\|B\| \leqslant 1} |\langle A, B \rangle| \,,$$

where the supremum is over tensors B of the same shape as A, and $\|B\| = \sqrt{\langle B, B \rangle}$ is the Hilbert–Schmidt norm.

Finally, we define the nuclear norm as the dual norm of the injective norm:

$$\|A\|_* := \sup_{\|B\|_\sigma \leqslant 1} |\langle A, B \rangle| \,.$$

For order-two tensors this agrees with the usual matrix nuclear norm. For higher-order tensors, this is the tensor nuclear norm, equivalently the projective tensor norm.

These norms satisfy the chain of inequalities

$$\|A\|_\sigma \leqslant \|A\| \leqslant \|A\|_*. \tag{25}$$

Indeed, the first inequality follows because unit simple tensors have Hilbert–Schmidt norm one. The second follows from the variational formula above and the fact that $\|B\|_\sigma \leq \|B\|$ for every tensor B.

B.3 Derivative Identities for Special Statistics

We record the derivative identities used in the derivation of the limiting dynamics. All tensor-valued derivatives are interpreted using the contraction conventions from the previous subsection.

Lemma B.1 (Jacobian of ψ for diagonal linear networks)**.** *Let ψ be as in Assumption 1.8. Then its Jacobian with respect to $x = (u, v)$ is*

$$\nabla\psi(u, v) = [\nabla_u \psi(u, v) \quad \nabla_v \psi(u, v)] \in \mathbb{R}^{d \times 2d},$$

where

$$\nabla_u \psi(u, v) = 2q_{11}\,\mathrm{diag}(u) + 2q_{12}\,\mathrm{diag}(v) + l_1 I_d,$$

and

$$\nabla_v \psi(u, v) = 2q_{12}\,\mathrm{diag}(u) + 2q_{22}\,\mathrm{diag}(v) + l_2 I_d.$$

Lemma B.2 (Gradient of Ψ)**.** *Let*

$$r := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x) \\ \beta^* \end{pmatrix}^\top a \in \mathbb{R}^2,$$

and define $\Psi(x; a) := f(r)$. Then we hav the expression

$$\nabla_x \Psi(x; a) = \frac{1}{\sqrt{d}}\, \nabla_{r_1} f(r)\, (\nabla\psi(x))^\top a \in \mathbb{R}^{2d},$$

where $\nabla_{r_1} f$ denotes the partial derivative of f with respect to its first coordinate.

Lemma B.3 (Derivative identities for S)**.** *Fix the product contour $\Gamma \subset \mathbb{C}^4$ from Remark 3.3, and let $z = (z_1, z_2, z_3, z_4) \in \Gamma$. We define*

$$\Omega(x, z) := R\big(z_1; \mathrm{diag}(u)\big) R\big(z_2; \mathrm{diag}(v)\big) R\big(z_3; \mathrm{diag}(\beta^*)\big) R\big(z_4; K\big) \in \mathbb{C}^{d \times d},$$

where

$$R(z; A) := (z I_d - A)^{-1}$$

denotes the resolvent of A. Since K is diagonal in our setting, all factors in $\Omega(x,z)$ are diagonal; in particular, $\Omega(x,z)^\top = \Omega(x,z)$.

Let us consider

$$W(x) := [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3},$$

and set the matrix

$$S(x,z) := \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3}.$$

Then

$$\nabla_u S(x,z), \nabla_v S(x,z) \in \mathbb{C}^{d\times 3\times 3},$$

and

$$\nabla_u^2 S(x,z), \nabla_{uv}^2 S(x,z), \nabla_v^2 S(x,z) \in \mathbb{C}^{d\times d\times 3\times 3}.$$

In the following displays, we write Ω for $\Omega(x,z)$. Products involving block matrices are interpreted as tensor contractions over the d-dimensional coordinate: a 3×3 block matrix with vector-valued entries is identified with an element of $\mathbb{C}^{d\times 3\times 3}$, and a 3×3 block matrix with matrix-valued entries is identified with an element of $\mathbb{C}^{d\times d\times 3\times 3}$.

The first derivatives of S have the following form

$$\begin{aligned}
\nabla_u S(x,z) = {}& \frac{1}{d}\nabla_u\psi(x)\Omega \begin{bmatrix} \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{d}\nabla_u\psi(x)\Omega \begin{bmatrix} \psi(x) & 0 & 0 \\ \beta^* & 0 & 0 \\ \mathbb{1}_d & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d}\operatorname{diag}(\psi(x)) R\big(z_1; \operatorname{diag}(u)\big)\Omega \begin{bmatrix} \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d}\operatorname{diag}(\beta^*) R\big(z_1; \operatorname{diag}(u)\big)\Omega \begin{bmatrix} 0 & 0 & 0 \\ \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d} R\big(z_1; \operatorname{diag}(u)\big)\Omega \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \psi(x) & \beta^* & \mathbb{1}_d \end{bmatrix} \in \mathbb{C}^{d\times 3\times 3},
\end{aligned}$$

$$\begin{aligned}
\nabla_v S(x,z) = {}& \frac{1}{d}\nabla_v\psi(x)\Omega \begin{bmatrix} \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{d}\nabla_v\psi(x)\Omega \begin{bmatrix} \psi(x) & 0 & 0 \\ \beta^* & 0 & 0 \\ \mathbb{1}_d & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d}\operatorname{diag}(\psi(x)) R\big(z_2; \operatorname{diag}(v)\big)\Omega \begin{bmatrix} \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d}\operatorname{diag}(\beta^*) R\big(z_2; \operatorname{diag}(v)\big)\Omega \begin{bmatrix} 0 & 0 & 0 \\ \psi(x) & \beta^* & \mathbb{1}_d \\ 0 & 0 & 0 \end{bmatrix} \\
& + \frac{1}{d} R\big(z_2; \operatorname{diag}(v)\big)\Omega \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \psi(x) & \beta^* & \mathbb{1}_d \end{bmatrix} \in \mathbb{C}^{d\times 3\times 3}.
\end{aligned}$$

Moreover, the second derivatives of S are

$$\nabla_u^2 S(x,z) = \frac{2q_{11}}{d}\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{2q_{11}}{d}\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & 0 & 0 \\ \operatorname{diag}(\beta^*) & 0 & 0 \\ I_d & 0 & 0 \end{bmatrix}$$

$$
\begin{aligned}
&+ \frac{2}{d}\Omega \begin{bmatrix} (\nabla_u \psi(x))^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{2}{d}\nabla_u \psi(x) R\big(z_1; \operatorname{diag}(u)\big)\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{2}{d}\nabla_u \psi(x) R\big(z_1; \operatorname{diag}(u)\big)\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & 0 & 0 \\ \operatorname{diag}(\beta^*) & 0 & 0 \\ I_d & 0 & 0 \end{bmatrix} \\
&+ \frac{2}{d}\operatorname{diag}(\psi(x)) R\big(z_1; \operatorname{diag}(u)\big)^2 \Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{2}{d}\operatorname{diag}(\beta^*) R\big(z_1; \operatorname{diag}(u)\big)^2 \Omega \begin{bmatrix} 0 & 0 & 0 \\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{2}{d} R\big(z_1; \operatorname{diag}(u)\big)^2 \Omega \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \end{bmatrix} \in \mathbb{C}^{d\times d\times 3\times 3},
\end{aligned}
$$

$$
\begin{aligned}
\nabla^2_{uv} S(x,z) = {} & \frac{2q_{12}}{d}\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{2q_{12}}{d}\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & 0 & 0 \\ \operatorname{diag}(\beta^*) & 0 & 0 \\ I_d & 0 & 0 \end{bmatrix} \\
&+ \frac{2}{d}\Omega \begin{bmatrix} (\nabla_u \psi(x))(\nabla_v \psi(x)) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{1}{d}\nabla_v \psi(x) R\big(z_1; \operatorname{diag}(u)\big)\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{1}{d}\nabla_v \psi(x) R\big(z_1; \operatorname{diag}(u)\big)\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & 0 & 0 \\ \operatorname{diag}(\beta^*) & 0 & 0 \\ I_d & 0 & 0 \end{bmatrix} \\
&+ \frac{1}{d}\nabla_u \psi(x) R\big(z_2; \operatorname{diag}(v)\big)\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{1}{d}\nabla_u \psi(x) R\big(z_2; \operatorname{diag}(v)\big)\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & 0 & 0 \\ \operatorname{diag}(\beta^*) & 0 & 0 \\ I_d & 0 & 0 \end{bmatrix} \\
&+ \frac{1}{d}\operatorname{diag}(\psi(x)) R\big(z_1; \operatorname{diag}(u)\big) R\big(z_2; \operatorname{diag}(v)\big)\Omega \begin{bmatrix} \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&+ \frac{1}{d}\operatorname{diag}(\beta^*) R\big(z_1; \operatorname{diag}(u)\big) R\big(z_2; \operatorname{diag}(v)\big)\Omega \begin{bmatrix} 0 & 0 & 0 \\ \operatorname{diag}(\psi(x)) & \operatorname{diag}(\beta^*) & I_d \\ 0 & 0 & 0 \end{bmatrix}
\end{aligned}
$$

$$+\frac{1}{d}R\big(z_1;\mathrm{diag}(u)\big)R\big(z_2;\mathrm{diag}(v)\big)\Omega\begin{bmatrix}0&0&0\\0&0&0\\\mathrm{diag}(\psi(x))&\mathrm{diag}(\beta^*)&I_d\end{bmatrix}\in\mathbb{C}^{d\times d\times 3\times 3},$$

$$\nabla_v^2 S(x,z)=\frac{2q_{22}}{d}\Omega\begin{bmatrix}\mathrm{diag}(\psi(x))&\mathrm{diag}(\beta^*)&I_d\\0&0&0\\0&0&0\end{bmatrix}+\frac{2q_{22}}{d}\Omega\begin{bmatrix}\mathrm{diag}(\psi(x))&0&0\\\mathrm{diag}(\beta^*)&0&0\\I_d&0&0\end{bmatrix}$$

$$+\frac{2}{d}\Omega\begin{bmatrix}(\nabla_v\psi(x))^2&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

$$+\frac{2}{d}\nabla_v\psi(x)R\big(z_2;\mathrm{diag}(v)\big)\Omega\begin{bmatrix}\mathrm{diag}(\psi(x))&\mathrm{diag}(\beta^*)&I_d\\0&0&0\\0&0&0\end{bmatrix}$$

$$+\frac{2}{d}\nabla_v\psi(x)R\big(z_2;\mathrm{diag}(v)\big)\Omega\begin{bmatrix}\mathrm{diag}(\psi(x))&0&0\\\mathrm{diag}(\beta^*)&0&0\\I_d&0&0\end{bmatrix}$$

$$+\frac{2}{d}\,\mathrm{diag}(\psi(x))R\big(z_2;\mathrm{diag}(v)\big)^2\Omega\begin{bmatrix}\mathrm{diag}(\psi(x))&\mathrm{diag}(\beta^*)&I_d\\0&0&0\\0&0&0\end{bmatrix}$$

$$+\frac{2}{d}\,\mathrm{diag}(\beta^*)R\big(z_2;\mathrm{diag}(v)\big)^2\Omega\begin{bmatrix}0&0&0\\\mathrm{diag}(\psi(x))&\mathrm{diag}(\beta^*)&I_d\\0&0&0\end{bmatrix}$$

$$+\frac{2}{d}R\big(z_2;\mathrm{diag}(v)\big)^2\Omega\begin{bmatrix}0&0&0\\0&0&0\\\mathrm{diag}(\psi(x))&\mathrm{diag}(\beta^*)&I_d\end{bmatrix}\in\mathbb{C}^{d\times d\times 3\times 3}.$$

B.4 Concentration and Pseudo-Lipschitz Estimates

We use the subgaussian norm $\|\cdot\|_{\psi_2}$, which is equivalent up to universal constants to the optimal variance proxy in a Gaussian tail bound. Namely, for a real-valued random variable X,

$$\|X\|_{\psi_2}\asymp\inf\Big\{V>0:\mathbb{P}\big(|X|>t\big)\leqslant 2e^{-t^2/V^2}\text{ for all }t>0\Big\}. \tag{26}$$

Gaussian random variables are naturally subgaussian. Moreover, Gaussian measures satisfy the stronger property of dimension-free *Lipschitz concentration*, which gives cocnentration inequalities for nonlinear functions of Gaussian vectors. Specifically, let V_0 be a finite-dimensional Hilbert space, and let Z be a centered isotropic Gaussian vector in V_0. If $g:V_0\to\mathbb{R}$ is Lipschitz with constant $L(g)$, meaning that

$$|g(x)-g(y)|\leq L(g)\|x-y\|_{V_0}\qquad\text{for all }x,y\in V_0,$$

then we have

$$\|g(Z)-\mathbb{E}\,g(Z)\|_{\psi_2}\leq CL(f),$$

where $C>0$ is an absolute universal constant. In particular, the bound does not depend on the dimension of V_0.

Pseudo-Lipschitz functions. In our setting, we will also work with functions which are not quite Lipschitz, in that they are locally Lipschitz (Lipschitz on compact sets) and moreover have polynomial growth of their Lipschitz on norm-balls.

Definition B.4 (Constant Pseudo-Lipschitz functions)**.** A function $f : V_0 \to V_1$ is called pseudo-Lipschitz of order α if there exists a constant $L = L(\alpha, f)$ such that, for all $x, y \in V_0$, we have

$$\|f(x) - f(y)\|_{V_1} \leq L\|x - y\|_{V_0} \left(1 + \|x\|_{V_0}^{\alpha} + \|y\|_{V_0}^{\alpha}\right).$$

We call L an α-pseudo-Lipschitz constant for f.

We will often works with outer functions and statistics whose gradients are α-pseudo-Lipschitz. To reduce pseudo-Lipschitz estimates to Lipschitz estimates on bounded sets, we will use projection onto norm balls. For $\beta > 0$, we define the *projection operator onto the ball of radius* β, denoted $\mathrm{Proj}_\beta : V_0 \to V_0$, by

$$\mathrm{Proj}_\beta(x) := \operatorname*{arg\,min}_{y \in \beta\mathbb{B}} \|x - y\|_{V_0}^2 = \begin{cases} x, & \text{if } \|x\|_{V_0} \leqslant \beta, \\ \beta\left(\frac{x}{\|x\|_{V_0}}\right), & \text{otherwise.} \end{cases},$$

where $\mathbb{B}$ denotes the unit ball in V_0.

Lemma B.5. *Suppose $f : V_0 \to V_1$ is α-pseudo-Lipschitz with constant L. Then the composition $f \circ \mathrm{Proj}_\beta$ is Lipschitz with constant $L(1 + 2\beta^\alpha)$.*

Proof. Projection onto a closed convex set is 1-Lipschitz. Therefore, for all $x, y \in V_0$, it holds

$$\begin{aligned} \left\|(f \circ \mathrm{Proj}_\beta)(x) - (f \circ \mathrm{Proj}_\beta)(y)\right\|_{V_1} &\leq L \left\|\mathrm{Proj}_\beta(x) - \mathrm{Proj}_\beta(y)\right\|_{V_0} \\ &\quad \times \left(1 + \|\mathrm{Proj}_\beta(x)\|_{V_0}^{\alpha} + \|\mathrm{Proj}_\beta(y)\|_{V_0}^{\alpha}\right) \\ &\leq L(1 + 2\beta^\alpha)\|x - y\|_{V_0}, \end{aligned}$$

as we had to show. □

We will use the following growth estimate for moments of the first partial derivative of the α-pseudo-Lipschitz outer function f; see [Collins-Woodfin et al., 2024a].

Lemma B.6 (Growth of $\nabla_{r_1} f$)**.** *Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with pseudo-Lipschitz constant $L(f)$, as in Assumption 1.4. Then, for any $p > 0$ and $r \in \mathbb{R}^2$, we have*

$$\left|\nabla_{r_1} f(r)\right|^p \leq C(\alpha, p)\, L(f)^p \left(1 + \|r\|\right)^{\max\{1, \alpha p\}}, \tag{27}$$

where $\nabla_{r_1} f$ denotes the partial derivative of f with respect to its first coordinate.

Moreover, set $r = \frac{1}{\sqrt{d}} \begin{bmatrix} \psi(x) & \beta^* \end{bmatrix}^\top a \in \mathbb{R}^2$, and let

$$W(x) := [\, \psi(x) \mid \beta^* \mid \mathbb{1}_d \,] \in \mathbb{R}^{d \times 3}.$$

Then the following moment bound holds:

$$\begin{aligned} \mathbb{E}_a\left[\left|\nabla_{r_1} f(r)\right|^p\right] &\leq C(\alpha, p)\, L(f)^p \left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W(x)\|\right)^{\max\{1, \alpha p\}}, \\ \|1 + \|r\|\|_{\psi_2} &\leq C\left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W(x)\|\right). \end{aligned} \tag{28}$$

C Dynamics of the Resolvent Statistic

The purpose of this section is to identify the statistic that mediates between the stochastic dynamics and their deterministic limit. Recall that our goal is to prove that, for every statistic $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$ satisfying Assumption 3.6, the processes $\varphi(x_{\lfloor td \rfloor})$ and $\varphi(\mathcal{X}_t)$ are close in the sense that they converge to the same deterministic limit.

The central object in this comparison is the matrix-valued statistic

$$S(x, z) := \frac{1}{d} W(x)^\top \Omega(x, z) W(x) \in \mathbb{C}^{3 \times 3},$$

where

$$W(x) := [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3},$$

and $\Omega(x,z)$ is the resolvent product defined in (17). We study this statistic along both trajectories, namely $S(\mathcal{X}_t, z)$ (homogenized SGD updates) and $S(x_{\lfloor td\rfloor}, z)$ (SGD updates), for z on the fixed product contour of Remark 3.3.

The argument has two main components. First, we show that the discrete and homogenized dynamics are close when tested against S, in the sense that $S(x_{\lfloor td\rfloor}, z)$ and $S(\mathcal{X}_t, z)$ remain close uniformly over the relevant time interval. Second, we show that $S(\mathcal{X}_t, z)$ is itself close to a deterministic limit $\mathscr{S}(t,z)$, where $\mathscr{S}$ solves the partial integro-differential equation (30). Combining these two steps yields a deterministic description of $S(x_{\lfloor td\rfloor}, z)$.

This statistic is powerful because it encodes enough spectral information to recover the deterministic limits of the broader class of statistics φ considered in Assumption 3.6. We make this reduction explicit in Section C.2. Thus, the dynamics of $S(x,z)$ serve as the dynamical nexus of the proof: once S is controlled, the limiting behavior of the other admissible statistics follows by the contour representations developed below. Beyond this, the dynamics of the mapping $S(x,z)$ itself often provide useful insights into analyzing the optimization trajectories of particular optimization problems. Indeed, properties of the solutions to which the algorithms converge can be derived by looking at the mapping $S(x,z)$.

C.1 Approximate Solutions and Stability

We begin by recalling the quantities entering the partial integro-differential equation. By Assumptions 1.10 and 1.11, we have

$$\mathcal{R}(x) := h\left(B(x)\right) \quad \text{and} \quad \mathbb{E}_a[\nabla_{r_1} f(r)^2] := I\left(B(x)\right) \quad \text{with} \quad B(x) = \frac{1}{d} W(x)^\top K W(x),$$

where $h, I : \mathbb{R}^{3\times 3} \to \mathbb{R}$ are differentiable and α-pseudo-Lipschitz. It will be useful to isolate the first-column components of the gradient of h. We therefore define

$$H\left(B(x)\right) = \left[\begin{array}{c|c|c} \nabla_{11} h\left(B(x)\right) & 0 & 0 \\ \hline \nabla_{21} h\left(B(x)\right) & 0 & 0 \\ \hline \nabla_{31} h\left(B(x)\right) & 0 & 0 \end{array}\right].$$

With this notation in place, we now introduce the partial integro-differential equation. In what follows, we will work with a function $\mathcal{F}\big(z, \mathcal{S}(t,\cdot)\big)$. Its explicit formula can be found in the displayed lines after Remark D.3, but will not be important for what follows; the only feature that matters is that it can be written as a sum of simple terms described below.

Partial Integro-differential Equation for $\mathscr{S}(t,z)$. Fix a product contour $\Gamma \subset \mathbb{C}^4$ as in Remark 3.3. For a multi-index $\zeta = (\zeta_1, \zeta_2, \zeta_3, \zeta_4)$ and a complex vector $z = (z_1, z_2, z_3, z_4) \in \Gamma$, write

$$z^\zeta := \prod_{i=1}^{4} z_i^{\zeta_i}.$$

Let $\mathcal{F}\big(z, \mathcal{S}(t,\cdot)\big)$ be a finite sum of terms of the form

$$-2\gamma(t)\, H\big(\mathcal{B}(t)\big)^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\big(\mathcal{S}(t,\cdot)\big)$$

or (29)

$$C(Q)\,\gamma(t)^2\, I\big(\mathcal{B}(t)\big) z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\big(\mathcal{S}(t,\cdot)\big),$$

where $\bigcirc_{i=1}^4 \mathcal{G}_i$ denotes successive composition,

$$\left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\big(\mathcal{S}(t,\cdot)\big) := (\mathcal{G}_1 \circ \mathcal{G}_2 \circ \mathcal{G}_3 \circ \mathcal{G}_4)\big(\mathcal{S}(t,\cdot)\big).$$

Here each $\mathcal{G}_i$ acts only on the z_i-coordinate and is chosen from the following three operators:

$$\mathcal{G}_i\big(\mathcal{S}(t,\cdot)\big) \in \left\{ \frac{1}{2\pi i} \oint_{\Gamma_i} q(w)\, \mathcal{S}(t,z)\, \mathrm{d}w,\ -\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}(t,z),\ \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}(t,z) \right\}.$$

Finally,

$$\mathcal{B}(t) := \frac{1}{(2\pi)^4} \oint_{\Gamma} z_4\, \mathcal{S}(t,z)\, \mathrm{d}z.$$

We then define $\mathscr{S}$ as a solution to the evolution equation

$$\mathrm{d}\mathscr{S}(t,\cdot) = \mathcal{F}\big(z, \mathscr{S}(t,\cdot)\big)\, \mathrm{d}t, \tag{30}$$

with initial condition

$$\mathscr{S}(0,z) = S(x_0, z). \tag{31}$$

We next introduce a notion of approximate solution to (30). The point of this definition is that both $S(\mathcal{X}_t, z)$ and $S(x_{\lfloor td \rfloor}, z)$, which are functions of both homogenized SGD and SGD respectively, will be shown to satisfy the PDE up to a small error.

To measure errors uniformly on the fixed contour, we define a norm, $\|\cdot\|_\Gamma$, on a continuous function $S\colon \Gamma \subset \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ by

$$\|S\|_\Gamma = \max_{z\in\Gamma} \|S(z)\|.$$

The following lemma relates this contour norm to the parameter squared norm along both the homogenized and discrete trajectories.

Lemma C.1. *There exist constants $0 < c < C < \infty$, depending on $\|K\|_{\mathrm{op}}$, $\|\beta^*\|_\infty$, and Γ, such that for all $t \geqslant 0$,*

$$c \leqslant \frac{\|S(\mathcal{X}_t,\cdot)\|_\Gamma}{\frac{1}{d}\|\mathcal{W}_t\|^2}, \frac{\|S(x_{\lfloor td \rfloor},\cdot)\|_\Gamma}{\frac{1}{d}\|W_{\lfloor td \rfloor}\|^2} \leqslant C.$$

Proof. For homogenized SGD, we have

$$\frac{1}{d}\|\mathcal{W}_t\|^2 = \frac{1}{(2\pi)^4} \oint_\Gamma \operatorname{Tr} S(\mathcal{X}_t, z)\, \mathrm{d}z \leqslant C\, \|S(\mathcal{X}_t,\cdot)\|_\Gamma.$$

On the other hand, by Neumann series and since $|z_i| > \|D_i\|_{\mathrm{op}}$ on each Γ_i, we know that

$$R(z_i; D_i) = (z_i \cdot I_d - D_i)^{-1} = \frac{1}{z_i}\left(I_d - \frac{1}{z_i}D_i\right)^{-1} = \frac{1}{z_i}\sum_{j=0}^{\infty}\left(\frac{1}{z_i}D_i\right)^j,$$

therefore

$$\|R(z_i; D_i)\|_{\mathrm{op}} \leqslant \frac{1}{|z_i|} \sum_{j=0}^{\infty} \left(\frac{1}{|z_i|} \|D_i\|_{\mathrm{op}} \right)^j = \frac{1}{|z_i|} \cdot \frac{1}{1 - \frac{1}{|z_i|} \|D_i\|_{\mathrm{op}}} = \frac{1}{|z_i| - \|D_i\|_{\mathrm{op}}} \leqslant 2.$$

Thus we conclude

$$\|S(\mathcal{X}_t, \cdot)\|_\Gamma = \max_{z \in \Gamma} \left\| \frac{1}{d} \mathcal{W}_t^\top \Omega \mathcal{W}_t \right\| \leqslant \frac{1}{d} \|\mathcal{W}_t\|^2 \cdot \Pi_{i=1}^4 \max_{z_i \in \Gamma_i} \|R(z_i; D_i)\|_{\mathrm{op}} \leqslant \frac{2^4}{d} \|\mathcal{W}_t\|^2 .$$

The same bounds hold for SGD with obvious changes. □

We will be working with *approximate solutions to the partial integro-differential equation* defined as:

Definition C.2 ((ε, M, T)-approximate solution to the partial integro-differential equation)**.** For constants $M, T, \varepsilon > 0$, we say that a continuous function

$$\mathcal{S} \colon \{t \geqslant 0\} \times \Gamma \subset \{t \geqslant 0\} \times \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$$

is an (ε, M, T)-approximate solution to (30) if it satisfies the following properties.

Stopping time. Define the stopping time

$$\hat{\tau}_M(\mathcal{S}) := \inf \left\{ t \geqslant 0 : \|\mathcal{S}(t, \cdot)\|_\Gamma > M \text{ or } \mathcal{B}(t) \notin \mathcal{U} \right\}.$$

(i) **Lipschitz regularity of derivatives.** For any other (ε, M, T)–approximate solution $\tilde{\mathcal{S}}$ and for each coordinate z_i, the first and second partial derivatives $\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}$ and $\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}$ exist, and they depend Lipschitz–continuously on $\mathcal{S}$ in the sense that for any $s \geqslant 0$ and $z \in \Gamma \subset \mathbb{C}^4$, it holds:

$$\left\| \frac{\mathrm{d}}{\mathrm{d}z_i} \mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, z) - \frac{\mathrm{d}}{\mathrm{d}z_i} \tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, z) \right\|_\Gamma \leqslant C \left\| \mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, \cdot) - \tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, \cdot) \right\|_\Gamma ,$$

$$\left\| \frac{\mathrm{d}^2}{\mathrm{d}z_i^2} \mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, z) - \frac{\mathrm{d}^2}{\mathrm{d}z_i^2} \tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, z) \right\|_\Gamma \leqslant C \left\| \mathcal{S}(s \wedge \hat{\tau}_M^{\mathcal{S}}, \cdot) - \tilde{\mathcal{S}}(s \wedge \hat{\tau}_M^{\tilde{\mathcal{S}}}, \cdot) \right\|_\Gamma .$$

(ii) **Approximate satisfaction of the PDE.** The following integral-form error bound holds:

$$\sup_{0 \leqslant t \leqslant (\hat{\tau}_M(\mathcal{S}) \wedge T)} \left\| \mathcal{S}(t, \cdot) - \mathcal{S}(0, \cdot) - \int_0^t \mathcal{F}(\cdot, \mathcal{S}(s, \cdot)) \, \mathrm{d}s \right\|_\Gamma \leqslant \varepsilon,$$

where the initial condition is $\mathcal{S}(0, \cdot) = S(x_0, \cdot)$, with x_0 the initialization of SGD.

We suppress the $\mathcal{S}$ in the notation for $\hat{\tau}_M$, that is $\hat{\tau}_M = \hat{\tau}_M(\mathcal{S})$, when the function $\mathcal{S}$ is clear from context.

Remark C.3. Consider

$$S(x, z) = \frac{1}{d} W(x)^\top \Omega(x, z) W(x) \in \mathbb{C}^{3\times 3}$$

for

$$\Omega = R(z_1; \mathrm{diag}(u)) \cdot R(z_2; \mathrm{diag}(v)) \cdot R(z_3; \mathrm{diag}(\beta^*)) \cdot R(z_4; K) \in \mathbb{C}^{d\times d}.$$

Differentiating the resolvent representation gives

$$\frac{\mathrm{d}^2}{\mathrm{d}z_i^2} S(x, z) = \frac{2}{d} W^\top R(z_i; D_i)^2 \Omega(x, z) W, \qquad \frac{\mathrm{d}^3}{\mathrm{d}z_i^3} S(x, z) = -\frac{6}{d} W^\top R(z_i; D_i)^3 \Omega(x, z) W,$$

so by Lemma C.1 and (64), we get

$$\left\| \frac{\mathrm{d}^2}{\mathrm{d}z_i^2} S(x, z) \right\|_\Gamma \leqslant C \|S(x, \cdot)\|_\Gamma \quad \text{and} \quad \left\| \frac{\mathrm{d}^3}{\mathrm{d}z_i^3} S(x, z) \right\|_\Gamma \leqslant C \|S(x, \cdot)\|_\Gamma .$$

Consequently, the processes $S(x_{\lfloor td \rfloor}, z)$ and $S(\mathcal{X}_t, z)$ satisfy condition (i), since $\|\mathcal{S}(s \wedge \hat{\tau}_M, \cdot)\|_\Gamma \leqslant M$ for any $s \geqslant 0$. Furthermore, in Section D, we prove that they also satisfy condition (ii) and hence that they are (ε, M, T)-approximate solutions. Note that we must extend the discrete time of SGD to a continuous time (see Section D for details). Finally, it is clear by definition that the exact solution $\mathscr{S}$ of (30) is an $(0, M, T)$-approximate solution.

The first result is a *stability* statement: any two (ε, M, T)-approximate solutions remain uniformly close up to the stopping time.

Proposition C.4 (Stability). *For all (ε, M, T)-approximate solutions $\mathcal{S}_1$ and $\mathcal{S}_2$, there exists a positive constant $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha, T)$ such that*

$$\sup_{0 \leqslant t \leqslant T} \|\mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot)\|_\Gamma \leqslant C \cdot \varepsilon,$$

where $\tau_M = \min\{\hat{\tau}_M(\mathcal{S}_1), \hat{\tau}_M(\mathcal{S}_2)\}$.

Proof. First note that $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_1)$ and $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_2)$. Thus all estimates below are taken up to the common stopping time τ_M. Write $\mathcal{S}_1$ and $\mathcal{S}_2$ as

$$\begin{aligned} \mathcal{S}_1(t, \cdot) &= \mathcal{S}_1(0, \cdot) + \int_0^t \mathcal{F}(\cdot, \mathcal{S}_1(s, \cdot))\, \mathrm{d}s + \varepsilon(\mathcal{S}_1) \\ \mathcal{S}_2(t, \cdot) &= \mathcal{S}_2(0, \cdot) + \int_0^t \mathcal{F}(\cdot, \mathcal{S}_2(s, \cdot))\, \mathrm{d}s + \varepsilon(\mathcal{S}_2), \end{aligned} \tag{32}$$

where $\varepsilon(\mathcal{S}_1)$ and $\varepsilon(\mathcal{S}_2)$ are error terms from the (ε, M, T)-approximate solution inequality and we have for $j = 1, 2$ the estimate

$$\sup_{0 \leqslant t \leqslant (T \wedge \tau_M)} \|\varepsilon(\mathcal{S}_j)\|_\Gamma \leqslant \varepsilon.$$

We first prove the stability estimate under the following Lipschitz bound on $\mathcal{F}$. Suppose that there exists a constant $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha)$ such that, for all s,

$$\|\mathcal{F}(\cdot, \mathcal{S}_1(s \wedge \tau_M, \cdot)) - \mathcal{F}(\cdot, \mathcal{S}_2(s \wedge \tau_M, \cdot))\|_\Gamma \leqslant C \|\mathcal{S}_1(s \wedge \tau_M, \cdot) - \mathcal{S}_2(s \wedge \tau_M, \cdot)\|_\Gamma. \tag{33}$$

We verify (33) at the end of the proof. However, Equations (32) and (33) imply

$$\begin{aligned} \sup_{0 \leqslant t \leqslant T \wedge \tau_M} \|\mathcal{S}_1(t, \cdot) - \mathcal{S}_2(t, \cdot)\|_\Gamma &\leqslant 2\varepsilon + \sup_{0 \leqslant t \leqslant T \wedge \tau_M} \int_0^t \|\mathcal{F}(\cdot, \mathcal{S}_1(s, \cdot)) - \mathcal{F}(\cdot, \mathcal{S}_2(s, \cdot))\|_\Gamma \, \mathrm{d}s \\ &= 2\varepsilon + \sup_{0 \leqslant t \leqslant T} \int_0^t \|\mathcal{F}(\cdot, \mathcal{S}_1(s \wedge \tau_M, \cdot)) - \mathcal{F}(\cdot, \mathcal{S}_2(s \wedge \tau_M, \cdot))\|_\Gamma \, \mathrm{d}s \\ &\leqslant 2\varepsilon + C \int_0^T \|\mathcal{S}_1(s \wedge \tau_M, \cdot) - \mathcal{S}_2(s \wedge \tau_M, \cdot)\|_\Gamma \, \mathrm{d}s, \end{aligned}$$

where $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha)$ is a positive constant.

Define $Q_T := \sup_{0 \leqslant t \leqslant T} \|\mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot)\|_\Gamma$. Then one has that

$$Q_T = \sup_{0 \leqslant t \leqslant T \wedge \tau_M} \|\mathcal{S}_1(t, \cdot) - \mathcal{S}_2(t, \cdot)\|_\Gamma \leqslant 2\varepsilon + C \int_0^T Q_s \, \mathrm{d}s.$$

By an application of Gronwall's inequality,

$$\sup_{0 \leqslant t \leqslant T} \|\mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot)\|_\Gamma \leqslant 2\varepsilon e^{CT},$$

and the result is shown.

It remains to verify the Lipschitz estimate (33). We will do this in steps. First, define $\mathcal{B}_j(\cdot) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \, \mathcal{S}_j(\cdot, z) \, \mathrm{d}z$ for $j = 1, 2$. We will use the shorthand $\mathcal{B}_j^{\tau_M}(s) = \mathcal{B}_j(s \wedge \tau_M)$ and $\mathcal{S}_j^{\tau_M}(s, \cdot) =$

$\mathcal{S}_j(s \wedge \tau_M, \cdot)$. Now, by the α-pseudo-Lipschitzness of ∇h (Assumption 1.10), we have

$$\begin{aligned}
\left\|H(\mathcal{B}_1^{\tau_M}(s))^\top - H(\mathcal{B}_2^{\tau_M}(s))^\top\right\| &\leqslant L(h)\left(1 + \|\mathcal{B}_1^{\tau_M}(s)\|^\alpha + \|\mathcal{B}_2^{\tau_M}(s)\|^\alpha\right)\|\mathcal{B}_1^{\tau_M}(s) - \mathcal{B}_2^{\tau_M}(s)\| \\
&\leqslant C(L(h), M, \alpha)\,\|\mathcal{B}_1^{\tau_M}(s) - \mathcal{B}_2^{\tau_M}(s)\| \\
\left\|H(\mathcal{B}_j^{\tau_M}(s))^\top\right\| &\leqslant L(h)\left(1 + \left\|\mathcal{B}_j^{\tau_M}(s)\right\|^{\alpha+1}\right) \leqslant C(L(h), M, \alpha)
\end{aligned} \tag{34}$$

since we have the expression

$$\left\|\mathcal{B}_j^{\tau_M}(s)\right\| = \left\|\frac{1}{(2\pi)^4}\oint_\Gamma z_4\,\mathcal{S}_j^{\tau_M}(s,z)\,\mathrm{d}z\right\| \leqslant C(|\Gamma|, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty)\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C \cdot M.$$

Here we used the stopping time τ_M explicitly. Similarly,

$$\begin{aligned}
\|\mathcal{B}_1^{\tau_M}(s) - \mathcal{B}_2^{\tau_M}(s)\| &\leqslant C\oint_\Gamma |z_4|\,\|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma\,\mathrm{d}\,|z| \\
&\leqslant C(|\Gamma|, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty)\,\|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma.
\end{aligned}$$

Consequently, there exists a positive constant (independent of s) such that

$$\left\|H(\mathcal{B}_1^{\tau_M}(s))^\top - H(\mathcal{B}_2^{\tau_M}(s))^\top\right\| \leqslant C(L(h), M, \alpha, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty)\,\|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma. \tag{35}$$

Analogously,

$$\begin{aligned}
\left\|\oint_{\Gamma_i} q(z_i)\mathcal{S}_1^{\tau_M}(s,z)\,\mathrm{d}z_i - \oint_{\Gamma_i} q(z_i)\mathcal{S}_2^{\tau_M}(s,z)\,\mathrm{d}z_i\right\| &\leqslant C(|\Gamma|, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty)\cdot \\
&\qquad \|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma \\
\left\|\oint_{\Gamma_i} q(z_i)\mathcal{S}_j^{\tau_M}(s,z)\,\mathrm{d}z_i\right\| &\leqslant C(|\Gamma|, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty)\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \\
&\leqslant C \cdot M.
\end{aligned} \tag{36}$$

Lastly, by the definition of approximate solution (C.2), for every $1 \leqslant i \leqslant 4$, we have

$$\begin{aligned}
\left\|\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}_1^{\tau_M}(s,z) - \frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}_2^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\,\|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma \\
\left\|\frac{\mathrm{d}}{\mathrm{d}z_i}\mathcal{S}_j^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C \cdot M,
\end{aligned} \tag{37}$$

and

$$\begin{aligned}
\left\|\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}_1^{\tau_M}(s,z) - \frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}_2^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\,\|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma \\
\left\|\frac{\mathrm{d}^2}{\mathrm{d}z_i^2}\mathcal{S}_j^{\tau_M}(s,z)\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C \cdot M.
\end{aligned} \tag{38}$$

Therefore, since the composition of Lipschitz functions is Lipschitz, we have

$$\begin{aligned}
\left\|\left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)(\mathcal{S}_1^{\tau_M}(s,\cdot)) - \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)(\mathcal{S}_2^{\tau_M}(s,\cdot))\right\|_\Gamma &\leqslant C\,\|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma \\
\left\|\left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)(\mathcal{S}_j^{\tau_M}(s,\cdot))\right\|_\Gamma &\leqslant C\left\|\mathcal{S}_j^{\tau_M}(s,\cdot)\right\|_\Gamma \leqslant C \cdot M.
\end{aligned} \tag{39}$$

From equations (34), (35), and (39), it follows that there exists a positive constant $C = C(L(h), M, \alpha, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \bar{\gamma})$ such that

$$\left\| 2\gamma(s) H\left(\mathcal{B}_1^{\tau_M}(s)\right)^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\left(\mathcal{S}_1^{\tau_M}(s,\cdot)\right) - 2\gamma(s) H\left(\mathcal{B}_2^{\tau_M}(s)\right)^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\left(\mathcal{S}_1^{\tau_M}(s,\cdot)\right) \right\|_\Gamma \leqslant C \cdot \|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma . \tag{40}$$

Next, by α-pseudo-Lipschitzness of the squared gradients (Assumption 1.11),

$$\begin{aligned} \|I(\mathcal{B}_1^{\tau_M}(s)) - I(\mathcal{B}_2^{\tau_M}(s))\| &\leqslant C(L(I), M, \alpha)\, \|\mathcal{B}_1^{\tau_M}(s) - \mathcal{B}_2^{\tau_M}(s)\| \\ \left\|I(\mathcal{B}_j^{\tau_M}(s))\right\| &\leqslant L(I)\left(1 + \left\|\mathcal{B}_j^{\tau_M}(s)\right\|^{\alpha+1}\right) \leqslant C(L(I), M, \alpha). \end{aligned} \tag{41}$$

Therefore, we deduce that

$$\left\| C(Q)\gamma(s)^2 I(\mathcal{B}_1^{\tau_M}(s)) z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\left(\mathcal{S}_1^{\tau_M}(s,\cdot)\right) - C(Q)\gamma(s)^2 I(\mathcal{B}_2^{\tau_M}(s)) z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)\left(\mathcal{S}_2^{\tau_M}(s,\cdot)\right) \right\|_\Gamma \leqslant C \cdot \|\mathcal{S}_1^{\tau_M}(s,\cdot) - \mathcal{S}_2^{\tau_M}(s,\cdot)\|_\Gamma , \tag{42}$$

where $C = C(L(I), M, \alpha, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \bar{\gamma})$ is a positive constant. The Lipschitz condition for $\mathcal{F}$ (33) holds after applying expressions (40) and (42). □

We now transfer the stability estimate from the resolvent statistic S to any statistic $\varphi = g \circ Q$ satisfying Assumption 3.6. Here we set

$$Q(x) := \frac{1}{d} W^\top q_1(\mathrm{diag}(u)) q_2(\mathrm{diag}(v)) q_4(K) W,$$

where q_1, q_2, q_4 are polynomials. For an approximate solution $\mathcal{S}_i$, define

$$\mathcal{Q}_i(t) = \frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_i(t, z)\, \mathrm{d}z.$$

The following proposition shows that given two approximate solutions, $\mathcal{S}_1$ and $\mathcal{S}_2$, the function $g \circ \mathcal{Q}_1(t)$ is close to $g \circ \mathcal{Q}_2(t)$. The pseudo-Lipschitzness of g, together with the boundedness imposed by the stopping time, allows us to control the difference between $g(\mathcal{Q}_1)$ and $g(\mathcal{Q}_2)$ by the contour distance between $\mathcal{S}_1$ and $\mathcal{S}_2$:

$$\sup_{0 \leqslant t \leqslant T} |(g \circ \mathcal{Q}_1)(t \wedge \tau_M) - (g \circ \mathcal{Q}_2)(t \wedge \tau_M)| \leqslant C \sup_{0 \leqslant t \leqslant T} \|\mathcal{S}_1(t \wedge \tau_M, \cdot) - \mathcal{S}_2(t \wedge \tau_M, \cdot)\|_\Gamma,$$

and then Proposition C.4 finishes the result.

Proposition C.5. *Suppose $\varphi\colon \mathbb{R}^{2d} \to \mathbb{R}$ is a statistic satisfying Assumption 3.6 such that $\varphi(x) = g \circ Q(x)$. Suppose $\mathcal{S}_1$ and $\mathcal{S}_2$ are (ε, M, T)-approximate solutions. Then there exists a positive constant $C = C(L(h), L(I), \bar{\gamma}, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \alpha, T)$ such that*

$$\sup_{0 \leqslant t \leqslant T} \left| g\left(\frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_1^{\tau_M}(t, z)\, \mathrm{d}z\right) - g\left(\frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_2^{\tau_M}(t, z)\, \mathrm{d}z\right) \right| \leqslant C \cdot \varepsilon$$

where $\tau_M := \hat{\tau}_M(\mathcal{S}_1) \wedge \hat{\tau}_M(\mathcal{S}_2)$. Here $\mathcal{S}_i^{\tau_M}(t, \cdot) = \mathcal{S}_i(t \wedge \tau_M, \cdot)$.

Proof. Since $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_1)$ and $\tau_M \leqslant \hat{\tau}_M(\mathcal{S}_2)$, we can always work on the smaller time τ_M. We define $\mathcal{Q}_i(t) = \frac{1}{(2\pi)^4} \oint_\Gamma q_1(z_1) q_2(z_2) q_4(z_4) \mathcal{S}_i(t, z)\, \mathrm{d}z$ and the stopped process $\mathcal{Q}_i^{\tau_M}(t) = \mathcal{Q}_i(t \wedge \tau_M)$ for $i = 1, 2$. First, we observe

$$\begin{aligned} \|\mathcal{Q}_i^{\tau_M}(t)\| &\leqslant C \oint_\Gamma |q_1(z_1)|\, |q_2(z_2)|\, |q_4(z_4)|\, \|\mathcal{S}_i^{\tau_M}(t, z)\|\, \mathrm{d}z \\ &\leqslant C(\|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4})\, \|\mathcal{S}_i^{\tau_M}(t, \cdot)\|_\Gamma \leqslant C \cdot M. \end{aligned} \tag{43}$$

Moreover, the map $\mathcal{S} \mapsto \mathcal{Q}$ is Lipschitz in the contour norm:

$$\begin{aligned}\left\|\mathcal{Q}_1^{\tau_M}(t) - \mathcal{Q}_2^{\tau_M}(t)\right\| &\leqslant C(\|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4}) \oint_\Gamma \left\|\mathcal{S}_1^{\tau_M}(t,z) - \mathcal{S}_2^{\tau_M}(t,z)\right\| \mathrm{d}\,|z| \\ &\leqslant C(\|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4}) \left\|\mathcal{S}_1^{\tau_M}(t,\cdot) - \mathcal{S}_2^{\tau_M}(t,\cdot)\right\|_\Gamma . \end{aligned} \tag{44}$$

Using the α-pseudo-Lipschitzness of g, together with the boundedness estimate (43) and the Lipschitz estimate (44), we obtain

$$\begin{aligned}\left|(g \circ \mathcal{Q}_1^{\tau_M})(t) - (g \circ \mathcal{Q}_2^{\tau_M})(t)\right| &\leqslant L(g) \left\|\mathcal{Q}_1^{\tau_M}(t) - \mathcal{Q}_2^{\tau_M}(t)\right\| \left(1 + \left\|\mathcal{Q}_1^{\tau_M}(t)\right\|^\alpha + \left\|\mathcal{Q}_2^{\tau_M}(t)\right\|^\alpha\right) \\ &\leqslant C \cdot \left\|\mathcal{S}_1^{\tau_M}(t,\cdot) - \mathcal{S}_2^{\tau_M}(t,\cdot)\right\|_\Gamma ,\end{aligned}$$

where $C = C(\|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, M, \|q_1\|_{\Gamma_1}, \|q_2\|_{\Gamma_2}, \|q_4\|_{\Gamma_4}, L(g), \alpha)$ is a positive constant. Taking the supremum over all $0 \leqslant t \leqslant T$ and applying Proposition C.4 finishes the result. □

C.2 Main Argument of the Proof: Concentration of the Resolvent Statistic S

In this section, we prove concentration of both SGD and homogenized SGD under the resolvent statistic S around the deterministic solution $\mathscr{S}(t,z)$ of (30). We first prove a concentration result with a common stopping time, and then strengthen it to a one-sided stopping-time formulation in Theorem C.7. This resolvent-level result implies Theorem 3.4. The important statistic which will play a pivotal role is

$$S(x,z) = \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3}, \tag{45}$$

as well as the function

$$B(x) = \frac{1}{d} W(x)^\top K W(x) \in \mathbb{R}^{3\times 3}. \tag{46}$$

Here

$$W(x) := [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3}.$$

We will extend the iterates of SGD, $\{x_k\}$, defined on discrete time k to continuous time. This is so that we can compare SGD and homogenized SGD, $\{\mathcal{X}_t\}$. We relate the k-th iterate of SGD to the continuous time parameter t in homogenized SGD through the relationship $k = \lfloor td \rfloor$. Thus, when $t = 1$, SGD has done exactly d updates.

We are now ready to state and prove one of our main results.

Theorem C.6 (Concentration under a common stopping time)**.** *Suppose the risk function $\mathcal{R}(x)$ (7) satisfies Assumptions 1.4, 1.10, and 1.11. Suppose the stepsize schedule satisfies 1.13, the iterates x_k and $\mathcal{X}_t$ satisfy 3.1, and the hidden parameters β^* satisfy Assumption 1.9. Moreover the data $a \sim \mathcal{N}(0,K)$ satisfies Assumption 1.5. Write $W_{\lfloor td \rfloor} := W(x_{\lfloor td \rfloor})$ and $\mathcal{W}_t := W(\mathcal{X}_t)$ initialized with $\mathcal{X}_0 = x_0$. Then there is an $\varepsilon > 0$ so that for any $T, M > 0$ and d sufficiently large, with overwhelming probability,*

$$\begin{aligned}\sup_{0\leqslant t\leqslant T\wedge\tau_M(S(x_{\lfloor td\rfloor},\cdot),\mathscr{S}(t,\cdot))} \left\|S(x_{\lfloor td\rfloor},\cdot) - \mathscr{S}(t,\cdot)\right\|_\Gamma &\leqslant d^{-\varepsilon}, \\ \sup_{0\leqslant t\leqslant T\wedge\tau_M(S(\mathcal{X}_t,\cdot),\mathscr{S}(t,\cdot))} \left\|S(\mathcal{X}_t,\cdot) - \mathscr{S}(t,\cdot)\right\|_\Gamma &\leqslant d^{-\varepsilon}, \\ \sup_{0\leqslant t\leqslant T\wedge\tau_M(S(x_{\lfloor td\rfloor},\cdot),S(\mathcal{X}_t,\cdot))} \left\|S(x_{\lfloor td\rfloor},\cdot) - S(\mathcal{X}_t,\cdot)\right\|_\Gamma &\leqslant d^{-\varepsilon},\end{aligned} \tag{47}$$

where $\mathscr{S}$ solves the partial integro-differential equation (30), and

$$\tau_M(\mathcal{S}_1, \mathcal{S}_2) = \min\{\hat{\tau}_M(\mathcal{S}_1), \hat{\tau}_M(\mathcal{S}_2)\}.$$

Proof. We will consider $\mathcal{S}_1(t,z) = S(x_{\lfloor td\rfloor}, z)$ and $\mathcal{S}_2(t,z) = S(\mathcal{X}_t, z)$ and suppress the notation by setting $\tau_M(\mathcal{S}_1, \mathcal{S}_2) = \tau_M$. We also note that the cases when $\mathcal{S}_1(t,z) = S(x_{\lfloor td\rfloor}, z)$ and $\mathcal{S}_2(t,z) = \mathscr{S}(t,z)$ and $\mathcal{S}_1(t,z) = S(\mathcal{X}_t, z)$ and $\mathcal{S}_2(t,z) = \mathscr{S}(t,z)$ follow an analogous proof, so for brevity, we do not present them.

By Proposition D.2, for some $\tilde{\varepsilon} > 0$, we have that $S(\mathcal{X}_t, z)$ is an $(d^{-\tilde{\varepsilon}}, M, T)$-approximate solution with overwhelming probability. Moreover, by Proposition D.4, $S(x_{\lfloor td \rfloor}, z)$ is an $(d^{-\tilde{\varepsilon}}, M, T)$-approximate solution. (For the deterministic function $\mathscr{S}(t, z)$, it is an $(0, M+1, T)$-approximate solution by definition.) Applying the stability result, Proposition C.4, to the three pairs

$$(S(x_{\lfloor td \rfloor}, \cdot), S(\mathcal{X}_t, \cdot)), \qquad (S(x_{\lfloor td \rfloor}, \cdot), \mathscr{S}(t, \cdot)), \qquad (S(\mathcal{X}_t, \cdot), \mathscr{S}(t, \cdot))$$

gives the three stated bounds, after possibly decreasing ε. □

In the next theorem, we note that one can remove the condition that *both* processes must remain bounded in $\|\cdot\|_\Gamma$ and separated from $\mathcal{U}$, and reduce this to show that we only need *one* of the processes to satisfy these properties. In this way, we can show that SGD is well-behaved and then conclude that homogenized SGD must also be well-behaved.

For any (ε, M, T)-approximate solution $\mathcal{S}(t, \cdot)$, we define the stopping time

$$\hat{\tau}_{M,\eta}(\mathcal{S}) = \inf\{t \geqslant 0 : \|\mathcal{S}(t, \cdot)\|_\Gamma > M \text{ or } \sup_{\hat{\mathcal{B}} \notin \mathcal{U}} \left\| \mathcal{B}(t, \mathcal{S}) - \hat{\mathcal{B}} \right\| \leqslant \eta\}$$

where we set

$$\mathcal{B}(t, \mathcal{S}) = \frac{1}{(2\pi)^4} \oint_\Gamma z_4 \mathcal{S}(t, z) \, \mathrm{d}z.$$

Our main theorem requires that *only one* of the statistics stays bounded, and not, in particular, both. To define this, we introduce a stopping time

$$\Theta_{M,\eta}^{\mathcal{S}_1, \mathcal{S}_2} := \max_{i=1,2} \hat{\tau}_{M,\eta}(\mathcal{S}_i). \tag{48}$$

We note that $\hat{\tau}_{M,0} = \hat{\tau}_M$ with $\hat{\tau}_M$ defined in the (ε, M, T)-approximate solution definition.

Theorem C.7 (Concentration under a one-sided stopping time)**.** *Suppose the risk function $\mathcal{R}(x)$ (7) satisfies Assumptions 1.4, 1.10, and 1.11. Suppose the learning rate schedule satisfies 1.13, the iterates x_k and $\mathcal{X}_t$ satisfy 3.1, and the hidden parameters β^* satisfy 1.9. Moreover the data $a \sim \mathcal{N}(0, K)$ satisfies Assumption 1.5. Let $\Theta_{M,\eta}$ be defined by (48). Write $W_{\lfloor td \rfloor} := W(x_{\lfloor td \rfloor})$ and $\mathcal{W}_t := W(\mathcal{X}_t)$ initialized with $\mathcal{X}_0 = x_0$. Then there is an $\varepsilon > 0$ so that for any $T, M, \eta > 0$ and d sufficiently large, with overwhelming probability,*

$$\begin{aligned}
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(x_{\lfloor td \rfloor}, \cdot), \mathscr{S}(t,\cdot)}} \left\| S(x_{\lfloor td \rfloor}, \cdot) - \mathscr{S}(t, \cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon}, \\
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(\mathcal{X}_t, \cdot), \mathscr{S}(t,\cdot)}} \left\| S(\mathcal{X}_t, \cdot) - \mathscr{S}(t, \cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon}, \\
\sup_{0 \leqslant t \leqslant T \wedge \Theta_{M,\eta}^{S(x_{\lfloor td \rfloor}, \cdot), S(\mathcal{X}_t, \cdot)}} \left\| S(x_{\lfloor td \rfloor}, \cdot) - S(\mathcal{X}_t, \cdot) \right\|_\Gamma &\leqslant d^{-\varepsilon},
\end{aligned} \tag{49}$$

where $\mathscr{S}$ solves the partial integro-differential equation (30).

Proof. Fix an $\eta > 0$. For two mappings $\mathcal{S}_1$ and $\mathcal{S}_2$, we define the stopping time

$$\tau_{M+1,0}^{\mathcal{S}_1, \mathcal{S}_2} = \min\{\hat{\tau}_{M+1,0}(\mathcal{S}_1), \hat{\tau}_{M+1,0}(\mathcal{S}_2)\}. \tag{50}$$

As in the previous theorem, we will consider $\mathcal{S}_1(t, z) = S(x_{\lfloor td \rfloor}, z)$ and $\mathcal{S}_2(t, z) = S(\mathcal{X}_t, z)$ and suppress the notation by setting $\tau_{M,\eta}^{\mathcal{S}_1, \mathcal{S}_2} = \tau_{M,\eta}$. We also note that the cases when $\mathcal{S}_1(t, z) = S(x_{\lfloor td \rfloor}, z)$ and $\mathcal{S}_2(t, z) = \mathscr{S}(t, z)$ and $\mathcal{S}_1(t, z) = S(\mathcal{X}_t, z)$ and $\mathcal{S}_2(t, z) = \mathscr{S}(t, z)$ follow an analogous proof so for brevity we do not present them.

By Theorem C.6, we have that

$$\sup_{0 \leqslant t \leqslant T \wedge \tau_{M+1,0}} \left\| S(x_{\lfloor td \rfloor}, \cdot) - S(\mathcal{X}_t, \cdot) \right\|_\Gamma \leqslant d^{-\varepsilon} \quad \text{w.o.p.} \tag{51}$$

The remaining component is to replace the stopping time $\tau_{M+1,0}$ which requires *both* statistics to have Γ-norm less than $M+1$ with $\Theta_{M,\eta}$ which only requires *one* of the statistics to remain in the

good set. Denote the event that (51) occurs by A_ε and its complement by $A_\varepsilon^{\mathrm{C}}$. Then for sufficiently large d, we have

$$\mathbb{P}\left[\Theta_{M,\eta} > \tau_{M+1,0}\right] \leqslant \mathbb{P}\left[A_\varepsilon^{\mathrm{C}}\right]. \tag{52}$$

To see this, suppose $\Theta_{M,\eta} > \tau_{M+1,0}$. At time $t = \tau_{M+1,0}$, one of the following exits must occur: $\left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_\Gamma \geqslant M+1$ or $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\| \leqslant 0$ or $\|S(\mathcal{X}_t,\cdot)\|_\Gamma \geqslant M+1$ or $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(\mathcal{X}_t)-\hat{B}\right\| \leqslant 0$. On the other hand, since $\tau_{M+1,0} = t < \Theta_{M,\eta}$, then either $\left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_\Gamma \geqslant M$ or $\|S(\mathcal{X}_t,\cdot)\|_\Gamma \geqslant M$ and then $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\| > \eta$ or $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(\mathcal{X}_t)-\hat{B}\right\| > \eta$.

Now we consider cases. Suppose $\|S(\mathcal{X}_t,\cdot)\|_\Gamma \geqslant M+1$. Then $\|S(\mathcal{X}_t,\cdot)\|_\Gamma$ cannot be less than or equal to M so it must have been that $\left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_\Gamma \leqslant M$. Since $t=\tau_{M+1,0}$, working on the event that (51) occurs, we have that

$$\|S(\mathcal{X}_t,\cdot)\|_\Gamma \leqslant \left\|S(x_{\lfloor td\rfloor},\cdot)-S(\mathcal{X}_t,\cdot)\right\|_\Gamma + \left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_\Gamma \leqslant d^{-\varepsilon}+M.$$

For sufficiently large d, then $\|S(\mathcal{X}_t,\cdot)\|_\Gamma < M+1$ which is a contradiction.

Suppose $\left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_\Gamma \geqslant M+1$. Then by reversing the roles of $x_{\lfloor td\rfloor}$ and $\mathcal{X}_t$ in the previous case, we see that this cannot occur.

Next suppose that $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(\mathcal{X}_t)-\hat{B}\right\| \leqslant 0$. Then $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(\mathcal{X}_t)-\hat{B}\right\|$ cannot be greater than η. Thus it had to be the case that $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\| > \eta$. Now working on the event that (51) occurs, we have that

$$\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\| \leqslant \left\|B(x_{\lfloor td\rfloor})-B(\mathcal{X}_t)\right\| \leqslant C\cdot\sup_{z_4\in\Gamma_4}|z_4|\cdot\left\|S(x_{\lfloor td\rfloor},\cdot)-S(\mathcal{X}_t,\cdot)\right\|_\Gamma \leqslant \tilde{C}\cdot d^{-\varepsilon}.$$

where C and $\tilde{C}$ are positive constants. Hence for sufficiently large d, we have $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\| < \eta$, a contradiction.

Lastly suppose $\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\| \leqslant 0$. By reversing the roles of $x_{\lfloor td\rfloor}$ and $\mathcal{X}_t$, we reach the same conclusion as the previous case.

Hence the inequality (52) holds and thus, $\tau_{M+1,0} \geqslant \Theta_{M,\eta}$ with overwhelming probability. The result follows. □

We immediately get a corollary which shows that SGD and homogenized SGD concentrate around the deterministic function $\mathscr{S}(t,z)$ which is a solution to the partial integro-differential equation (30) provided Assumption 3.1 holds.

Corollary C.8 (Non-explosiveness and concentration)**.** *Suppose the assumptions of Theorem C.7 hold. Suppose, in addition, for a fixed $T>0$ and $\eta>0$ that*

$$\sup_{0\leqslant t\leqslant T}\sup_{\hat{B}\notin\mathcal{U}}\left\|B(x_{\lfloor td\rfloor})-\hat{B}\right\| > \eta \quad \textit{w.o.p.} \tag{53}$$

Then there is an $\varepsilon>0$ so that for d sufficiently large, with overwhelming probability,

$$\sup_{0\leqslant t\leqslant T}\left\|S(x_{\lfloor td\rfloor},\cdot)-\mathscr{S}(t,\cdot)\right\|_\Gamma \leqslant d^{-\varepsilon} \quad \textit{and} \quad \sup_{0\leqslant t\leqslant T}\left\|S(x_{\lfloor td\rfloor},\cdot)-S(\mathcal{X}_t,\cdot)\right\|_\Gamma \leqslant d^{-\varepsilon}, \tag{54}$$

and therefore

$$\sup_{0\leqslant t\leqslant T}\|\mathscr{S}(t,\cdot)-S(\mathcal{X}_t,\cdot)\|_\Gamma \leqslant 2d^{-\varepsilon}. \tag{55}$$

Proof. Define the following stopping time similar to $\Theta_{M,\eta}$ in (48) by

$$\tilde{\Theta}_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2} = \max_{i=1,2}\inf\{t\geqslant 0 : \sup_{\hat{\mathcal{B}}\notin\mathcal{U}}\left\|\mathcal{B}(t,\mathcal{S}_i)-\hat{\mathcal{B}}\right\| \leqslant \eta\}. \tag{56}$$

Here we think of $\mathcal{S}_1$ as either homogenized SGD or $\mathscr{S}$ and $\mathcal{S}_2$ as SGD. The stopping time $\Theta_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2}$ from (48) is controlled by the non-explosiveness and separation assumptions on the SGD trajectory. By Lemma C.1 and Assumption 1.9, there exists some $C > 0$ independent of d such that

$$\left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_\Gamma \leqslant \frac{2^4}{d}\left\|W_{\lfloor td\rfloor}\right\|^2 \leqslant 2^4\left(\left\|x_{\lfloor td\rfloor}\right\|_\infty^4 + \|\beta^*\|_\infty^2 + 1\right) \leqslant C\left(\left\|x_{\lfloor td\rfloor}\right\|_\infty^4 + 1\right).$$

Consequently, this translates into

$$\{t \geqslant 0 : \left\|S(x_{\lfloor td\rfloor},\cdot)\right\|_\Gamma > C\left(M^4+1\right)\} \subset \{t \geqslant 0 : \left\|x_{\lfloor td\rfloor}\right\|_\infty > M\},$$

and so the infimum of the right-hand-side is smaller than the infimum of the left-hand-side. Moreover, we have by Assumption 3.1 that

$$T \leqslant \inf\{t \geqslant 0 : \left\|x_{\lfloor td\rfloor}\right\|_\infty > M\} \quad \text{w.o.p.}$$

Similarly we have that

$$T \leqslant \inf\{t \geqslant 0 : \sup_{\hat{B}\notin\mathcal{U}} \left\|B(x_{\lfloor td\rfloor}) - \hat{B}\right\| \leqslant \eta\} \quad \text{w.o.p.}$$

Thus, we have that

$$T \leqslant \tilde{\Theta}_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2} \leqslant \Theta_{C(M^4+1),\eta}^{\mathcal{S}_1,\mathcal{S}_2} \quad \text{w.o.p.},$$

where $\mathcal{S}_1$ is either $S(\mathcal{X}_t,\cdot)$ or $\mathscr{S}(t,\cdot)$. By Theorem C.7, we immediately get the result in (54). A simple triangle inequality gives the result in (55). □

Lastly, we make one final connection to Theorem 3.4, proving the result below.

Proof of Theorem 3.4. The result immediately follows from Theorem C.7 and Corollary C.8 after noting that

$$B(x) = \frac{1}{(2\pi)^4}\oint_\Gamma z_4 S(x,z)\,\mathrm{d}z \quad \text{and} \quad \mathscr{B}(t) = \frac{1}{(2\pi)^4}\oint_\Gamma z_4 \mathscr{S}(t,z)\,\mathrm{d}z$$

and Lipschitzness of the integral, that is,

$$\left\|\oint_\Gamma z_4\mathcal{S}_1(t,\cdot)\,\mathrm{d}z - \oint_\Gamma z_4\mathcal{S}_2(t,\cdot)\,\mathrm{d}z\right\| \leqslant C\cdot\left\|\mathcal{S}_1(t,\cdot) - \mathcal{S}_2(t,\cdot)\right\|_\Gamma \quad \text{for some positive} \quad C > 0.$$

□

C.3 Concentration of General Statistics

We now transfer the resolvent-level concentration result of Theorem C.7 to any statistic $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$ satisfying Assumption 3.6. This gives Theorem C.9, which is a reformulation of Theorem 3.7. The result applies, in particular, to the risk and curvature curves, $\mathcal{R}(x)$ and $\mathrm{tr}(\nabla^2\mathcal{R}(x))$, as well as to other generalization metrics covered by Assumption 3.6.

In this section, the statistics $\varphi\colon\ \mathbb{R}^{2d} \to \mathbb{R}$ of interest satisfy a composite structure

$$\varphi(x) = g\left(\frac{1}{d}W(x)^\top q_1(\mathrm{diag}(u))q_2(\mathrm{diag}(v))q_4(K)W(x)\right)$$

where $g\colon\ \mathbb{R}^{3\times 3} \to \mathbb{R}$ is α-pseudo-Lipschitz on $\mathcal{U}$ and q_1, q_2, q_4 are polynomials (see Assumption 3.6). Note that we assume that g is evaluated on the real-valued matrix recovered by the contour integral, which lies in $\mathcal{U}$ up to the stopping time.

Define

$$\mathscr{Q}(t) := \frac{1}{(2\pi)^4}\oint_\Gamma q_1(z_1)q_2(z_2)q_4(z_4)\mathscr{S}(t,z)\,\mathrm{d}z,$$

where $\mathscr{S}$ solves (30). The deterministic equivalent of $\varphi(x_{\lfloor td\rfloor})$ and $\varphi(\mathcal{X}_t)$ is then

$$\phi(t) := g(\mathscr{Q}(t)). \tag{57}$$

Thus we state our concentration theorem for $\varphi(x_{\lfloor td\rfloor})$ and $\varphi(\mathcal{X}_t)$.

Theorem C.9 (Concentration of general statistics)**.** *Suppose the Assumptions of Theorem C.7 hold. Suppose, in addition, the statistic satisfies a composite structure,*

$$\varphi(x) = g\left(\frac{1}{d}W(x)^\top q_1(\mathrm{diag}(u))q_2(\mathrm{diag}(v))q_4(K)W(x)\right)$$

where $g\colon \mathbb{R}^{3\times 3} \to \mathbb{R}$ *is* α*-pseudo-Lipschitz on* $\mathcal{U}$ *and* q_1, q_2, q_4 *are polynomials (see Assumption 3.6). Then there is an* $\varepsilon > 0$ *so that for any* $T, M, \eta > 0$ *and* d *sufficiently large, with overwhelming probability,*

$$\begin{aligned}
\sup_{0\leqslant t\leqslant T\wedge\Theta_{M,\eta}^{S(x_{\lfloor td\rfloor},\cdot),\mathscr{S}(t,\cdot)}} \left|\varphi(x_{\lfloor td\rfloor}) - \phi(t)\right| &\leqslant d^{-\varepsilon},\\
\sup_{0\leqslant t\leqslant T\wedge\Theta_{M,\eta}^{S(\mathcal{X}_t,\cdot),\mathscr{S}(t,\cdot)}} \left|\varphi(\mathcal{X}_t) - \phi(t)\right| &\leqslant d^{-\varepsilon},\\
\sup_{0\leqslant t\leqslant T\wedge\Theta_{M,\eta}^{S(x_{\lfloor td\rfloor},\cdot),S(\mathcal{X}_t,\cdot)}} \left|\varphi(x_{\lfloor td\rfloor}) - \varphi(\mathcal{X}_t)\right| &\leqslant d^{-\varepsilon},
\end{aligned} \tag{58}$$

where ϕ *is defined in* (57) *and the stopping time* $\Theta_{M,\eta}^{\mathcal{S}_1,\mathcal{S}_2}$ *is defined in* (48)*.*

Proof. As in the proof of Theorem C.7, we define the stopping time $\tau_{M+1,\eta}^{\mathcal{S}_1,\mathcal{S}_2}$ as in (50) and suppress the notation by setting $\tau_{M+1,\eta}^{\mathcal{S}_1,\mathcal{S}_2} = \tau_{M,\eta}$. We will consider the case when $\mathcal{S}_1(t,\cdot) = S(x_{\lfloor td\rfloor},\cdot)$ and $\mathcal{S}_2(t,\cdot) = S(\mathcal{X}_t,\cdot)$. The other cases will follow by analogous proof.

By Proposition D.2, we have that $S(\mathcal{X}_t, z)$ is an $(d^{-\tilde{\varepsilon}}, M+1, T)$-approximate solution with overwhelming probability. Moreover, by Proposition D.4, the function $S(x_{\lfloor td\rfloor}, z)$ is an $(d^{-\tilde{\varepsilon}}, M+1, T)$-approximate solution. (For the deterministic function $\mathscr{S}$, it is a $(0, M+1, T)$-approximate solution by definition.) We observe that

$$\frac{1}{(2\pi)^4}\oint_\Gamma q_1(z_1)q_2(z_2)q_4(z_4)S(x_{\lfloor td\rfloor}, z)\,\mathrm{d}z = \frac{1}{d}W_{\lfloor td\rfloor}^\top q_1(\mathrm{diag}(u_{\lfloor td\rfloor}))q_2(\mathrm{diag}(v_{\lfloor td\rfloor}))q_4(K)W_{\lfloor td\rfloor},$$

$$\frac{1}{(2\pi)^4}\oint_\Gamma q_1(z_1)q_2(z_2)q_4(z_4)S(\mathcal{X}_t, z)\,\mathrm{d}z = \frac{1}{d}\mathcal{W}_t^\top q_1(\mathrm{diag}(_t))q_2(\mathrm{diag}(_t))q_4(K)\mathcal{W}_t.$$

We apply Proposition C.5 to conclude that there exists a $\varepsilon > 0$ such that

$$\sup_{0\leqslant t\leqslant T\wedge\tau_{M+1,0}} \left|\varphi(x_{\lfloor td\rfloor}) - \varphi(\mathcal{X}_t)\right| \leqslant d^{-\varepsilon} \quad \text{w.o.p.} \tag{59}$$

Using the same argument as in Theorem C.7, we can remove the stopping time $\tau_{M+1,0}$ and replace it with $\Theta_{M,0}$ for sufficiently large d. □

Lastly we formulate an immediate corollary which follows directly from the proofs of Theorem C.9 and Corollary C.8.

Corollary C.10. *Suppose the Assumptions of Theorem C.9 and Corollary C.8 hold. For any fixed* $T > 0$*, there exists* $\varepsilon > 0$ *such that, for* d *sufficiently large, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T}\left|\varphi(x_{\lfloor td\rfloor}) - \phi(t)\right| \leqslant d^{-\varepsilon} \quad \textit{and} \quad \sup_{0\leqslant t\leqslant T}\left|\varphi(x_{\lfloor td\rfloor}) - \varphi(\mathcal{X}_t)\right| \leqslant d^{-\varepsilon}, \tag{60}$$

and therefore

$$\sup_{0\leqslant t\leqslant T}\left|\varphi(\mathcal{X}_t) - \phi(t)\right| \leqslant 2d^{-\varepsilon}. \tag{61}$$

Theorem 3.7 follows by applying Corollary C.10 to the statistic φ appearing in the theorem statement.

D SGD and Homogenized SGD are Approximate Solutions

In this section, we show that the resolvent statistic $t \mapsto S(x_{\lfloor td\rfloor}, z)$ associated with SGD and the corresponding statistic $t \mapsto S(\mathcal{X}_t, z)$ associated with homogenized SGD satisfy the partial integro-differential equation (30) up to a small error. The argument uses a martingale method in the spirit

of diffusion approximation [Ethier and Kurtz, 1986]: after applying a Doob decomposition to the statistic S, we show that the martingale terms are negligible and that the drift terms match the operator $\mathcal{F}$.

Recall the matrix-valued statistics

$$S(x,z) = \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3}, \qquad B(x) = \frac{1}{d} W(x)^\top K W(x) \in \mathbb{R}^{3\times 3},$$

where $W(x) = [\,\psi(x) \mid \beta^* \mid \mathbb{1}_d\,]$ and $\Omega(x,z)$ is defined in (17).

We first show that both homogenized SGD and SGD on $S(\cdot, z)$ are (ε, M, T)-approximate solutions as defined in Definition C.2. Then by Proposition C.4, it is immediately implied that both homogenized SGD and SGD on $S(\cdot, z)$ are uniformly close. Finally, Proposition C.5 establishes that the same hold for any statistic $\varphi(x)$ satisfying Assumption 3.6. In order to show that both homogenized SGD and SGD on $S(\cdot, z)$ are (ε, M, T)-approximate solutions, we perform a Doob decomposition for both homogenized SGD and SGD and then show that both martingale terms are small.

To compare homogenized SGD with SGD, we rescale time by setting $k = \lfloor td \rfloor$. Thus, one unit of continuous time corresponds to d SGD updates, and we write

$$x_{td} = x_{\lfloor td \rfloor} \quad \text{(SGD)} \qquad \text{and} \qquad \mathfrak{X}_t \quad \text{(HSGD)}.$$

Throughout this section, expressions such as $S(x_{\lfloor td \rfloor}, z)$ are understood under this identification, and scalar statistics are real-valued $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$.

When applying the scalar Itô, Taylor, Doob decomposition, or martingale estimates to the complex-valued matrix statistic $S(\cdot, z) \in \mathbb{C}^{3\times 3}$, we do so coordinatewise. Namely, for $a, b \in \{1, 2, 3\}$ and fixed $z \in \Gamma$, define the real-valued functions

$$S_{ab}^{\mathrm{Re},z}(x) := \operatorname{Re} S_{ab}(x,z), \qquad S_{ab}^{\mathrm{Im},z}(x) := \operatorname{Im} S_{ab}(x,z).$$

All scalar identities and estimates below are applied to these real-valued functions. The corresponding complex matrix-valued identities are then obtained by recombining real and imaginary parts. For example,

$$\big(\mathcal{M}_t(S)(z)\big)_{ab} := \mathcal{M}_t\big(S_{ab}^{\mathrm{Re},z}\big) + i\,\mathcal{M}_t\big(S_{ab}^{\mathrm{Im},z}\big).$$

Since there are only finitely many entries, passing from scalar estimates for the real and imaginary parts to matrix estimates for S only changes constants.

Our first argument is a net argument showing that we do not need to work with every $z \in \Gamma \subset \mathbb{C}^4$, but only polynomially many in d. For this, recall Γ defined in Remark 3.3. For a fixed $\delta > 0$, we say that Γ_i^δ is a $d^{-\delta}$-mesh of Γ_i if $\Gamma_i^\delta \subset \Gamma_i$ and for every $z_i \in \Gamma_i$ there exists a $\bar{z}_i \in \Gamma_i^\delta$ such that $|z_i - \bar{z}_i| < d^{-\delta}$.

Lemma D.1 (Net argument). *Fix $T, M > 0$ and let $\delta > 0$. For each $i = 1, \dots, 4$, let Γ_i^δ be a $d^{-\delta}$-mesh of Γ_i, chosen so that*

$$|\Gamma_i^\delta| \leqslant C_i d^\delta.$$

Set

$$\Gamma^\delta := \Gamma_1^\delta \times \Gamma_2^\delta \times \Gamma_3^\delta \times \Gamma_4^\delta.$$

Then

$$|\Gamma^\delta| = \prod_{i=1}^4 |\Gamma_i^\delta| \leqslant C_\Gamma d^{4\delta},$$

where $C_\Gamma := \prod_{i=1}^4 C_i$ depends only on the fixed contour Γ.

Let $S(t,z)$ denote either $S(x_{\lfloor td \rfloor}, z)$ or $S(\mathfrak{X}_t, z)$, and suppose that

$$\sup_{0 \leqslant t \leqslant \hat{\tau}_M(S) \wedge T} \left\| S(t,\cdot) - S(0,\cdot) - \int_0^t \mathcal{F}\big(\cdot, S(s,\cdot)\big)\,\mathrm{d}s \right\|_{\Gamma^\delta} \leqslant \varepsilon, \tag{62}$$

where

$$\hat{\tau}_M(S) := \inf\left\{ t \geq 0 : \|S(t,\cdot)\|_\Gamma > M \text{ or } B(t,S) \notin \mathcal{U} \right\}, \qquad B(t,S) := \frac{1}{(2\pi)^4} \oint_\Gamma z_4 S(t,z)\,\mathrm{d}z.$$

Then the approximate PDE residual bound (i) in Definition C.2 holds on the full contour Γ with error $\varepsilon + Cd^{-\delta}$, that is,

$$\sup_{0\leqslant t\leqslant \hat{\tau}_M(S)\wedge T} \left\| S(t,\cdot) - S(0,\cdot) - \int_0^t \mathcal{F}\big(\cdot, S(s,\cdot)\big)\,\mathrm{d}s \right\|_\Gamma \leq \varepsilon + Cd^{-\delta}.$$

Here $C = C(M, T, \Gamma, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty, \bar{\gamma}, L(I), L(h))$ is positive and independent of d. Consequently, if S also satisfies the derivative regularity condition (i), then S is an $(\varepsilon + Cd^{-\delta}, M, T)$-approximate solution.

Proof. We consider only $S(t,z) = S(\mathcal{X}_t, z)$ as the same argument will also hold for SGD. We also will always work with the stopped process, that is, $S(t\wedge\hat{\tau}_M, z)$, where $\hat{\tau}_M = \inf\{t \geqslant 0 : \|S(t,\cdot)\|_\Gamma \geqslant M\}$. To simplify the notation, we suppress the $\hat{\tau}_M$ and use $S(t,z)$.

First, we state some resolvent identities. One such resolvent identity gives

$$\|R(z_i; D_i) - R(\bar{z}_i; D_i)\|_{\mathrm{op}} \leqslant |z_i - \bar{z}_i|\,\|R(z_i; D_i)R(\bar{z}_i; D_i)\|_{\mathrm{op}} \quad \text{for any } z_i, \bar{z}_i \in \Gamma_i. \tag{63}$$

Furthermore, by Neumann series and since $|z_i| > \|D_i\|_{\mathrm{op}}$ on each Γ_i, we know that

$$R\left(z_i; D_i\right) = \left(z_i \cdot I_d - D_i\right)^{-1} = \frac{1}{z_i}\left(I_d - \frac{1}{z_i}D_i\right)^{-1} = \frac{1}{z_i}\sum_{j=0}^{\infty}\left(\frac{1}{z_i}D_i\right)^j,$$

so then

$$\left\|R\left(z_i; D_i\right)\right\|_{\mathrm{op}} \leqslant \frac{1}{|z_i|}\sum_{j=0}^{\infty}\left(\frac{1}{|z_i|}\|D_i\|_{\mathrm{op}}\right)^j = \frac{1}{|z_i|}\cdot\frac{1}{1-\frac{1}{|z_i|}\|D_i\|_{\mathrm{op}}} = \frac{1}{|z_i| - \|D_i\|_{\mathrm{op}}} \leqslant C$$

and we immediately get

$$\sup_{z_i\in\Gamma_i}\|R(z_i; D_i)\|_{\mathrm{op}} \leqslant C(M, \|K\|_{\mathrm{op}}, \|\beta^*\|_\infty). \tag{64}$$

These bounds will be useful later in the proof.

Next, with these bounds, we can get estimates on quantities involving $S(t,\cdot)$ where t is fixed and z varies. Fix $z \in \Gamma$ and let $\bar{z} \in \Gamma_\delta$ be such that $|z_i - \bar{z}_i| < d^{-\delta}$ for each $i = 1, \cdots, 4$. Then, using Lemma C.1 (and the stopping time $\hat{\tau}_M$), we obtain

$$\begin{aligned}\|S(t,z) - S(t,\bar{z})\| &\leqslant \frac{C}{d}\|\mathcal{W}_t\|^2\sum_{i=1}^4 |z_i - \bar{z}_i|\,\|R(z_i; D_i)\|_{\mathrm{op}}\,\|R(\bar{z}_i; D_i)\|_{\mathrm{op}} \\ &\leqslant C\,\|S(\mathcal{X}_t,\cdot)\|_\Gamma\, d^{-\delta} \\ &\leqslant C\cdot M\cdot d^{-\delta},\end{aligned} \tag{65}$$

where we used the boundedness of the product contour Γ in the last inequality.

Similarly, for any $z \in \Gamma$, we have

$$\|S(t,z)\| \leqslant \frac{1}{d}\|\mathcal{W}_t\|^2\,\Pi_{i=1}^4\|R(z_i; D_i)\|_{\mathrm{op}} \leqslant C\,\|S(\mathcal{X}_t,\cdot)\|_\Gamma \leqslant C\cdot M.$$

Since $S(t,z)$ is a product of resolvents and $t \leqslant \hat{\tau}_M$, all finitely many z-derivatives appearing in the terms defining $\mathcal{F}$ are uniformly bounded on Γ. Hence each term

$$z^\zeta\left(\bigcirc_i \mathcal{G}_i\right)S(t,z)$$

is Lipschitz in z on Γ, with Lipschitz constant depending only on the stopped bounds and the fixed contour. Thus, since $z, \bar{z} \in \Gamma$ and the contour Γ is bounded,

$$\left\|z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)(S(s,z)) - \bar{z}^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)(S(s,\bar{z}))\right\| \leqslant C\cdot M\cdot d^{-\delta}, \tag{66}$$

where we used equations (39) and (65).

Now we are ready to prove the main result of the proposition. For a fixed $t \leqslant \hat{\tau}_M$ and $z \in \Gamma$ with $\bar{z} \in \Gamma_\delta$ such that $|z_i - \bar{z}_i| \leqslant d^{-\delta}$ for each $i = 1, \cdots, 4$,

$$
\begin{aligned}
&\left\| S(t,z) - S(0,z) - \int_0^t \mathcal{F}(z, S(s,\cdot))\,\mathrm{d}s \right\| \\
&\quad \leqslant \|S(t,z) - S(t,\bar{z})\| + \|S(0,z) - S(0,\bar{z})\| + \int_0^t \|\mathcal{F}(z,S(s,\cdot)) - \mathcal{F}(\bar{z},S(s,\cdot))\|\,\mathrm{d}s \\
&\quad + \left\| S(t,\bar{z}) - S(0,\bar{z}) - \int_0^t \mathcal{F}(\bar{z}, S(s,\cdot))\,\mathrm{d}s \right\| \\
&\quad \leqslant CMd^{-\delta} \\
&\quad + 2\bar{\gamma} \int_0^t \|H(B(s))\| \cdot \left\| z^\zeta \left(\textstyle\bigcirc_{i=1}^4 \mathcal{G}_i\right)(S(s,z)) - \bar{z}^\zeta \left(\textstyle\bigcirc_{i=1}^4 \mathcal{G}_i\right)(S(s,\bar{z})) \right\| \mathrm{d}s \\
&\quad + C(Q)\bar{\gamma}^2 \int_0^t |I(B(s))| \cdot \left\| z^\zeta \left(\textstyle\bigcirc_{i=1}^4 \mathcal{G}_i\right)(S(s,z)) - \bar{z}^\zeta \left(\textstyle\bigcirc_{i=1}^4 \mathcal{G}_i\right)(S(s,\bar{z})) \right\| \mathrm{d}s \\
&\quad + \varepsilon,
\end{aligned}
\tag{67}
$$

where the omitted terms are bounded in the same way, since $\mathcal{F}$ is a finite sum of terms of the two forms in (29).

Here we used (65) to bound the first two terms in the first inequality and ε for the last term by the assumption (62) in the statement. In the difference $\mathcal{F}(z, S(s,\cdot)) - \mathcal{F}(\bar{z}, S(s,\cdot))$, the coefficients depending only on s and on $\mathcal{B}(s)$ are identical. Thus the difference is controlled by the variation in the monomial and coordinate-operator factor as z is replaced by $\bar{z}$.

As we have already shown that $z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right)(S(s,z))$ is Lipschitz in z, we only need to bound $|I(B(s))|$ and $\|H(B(s))\|$. We have already shown a uniform bound on $|I(B(s))|$ and $\|H(B(s))\|$ in the proof of Proposition C.4. Notably, we showed that for $s \leqslant \hat{\tau}_M$, we have that $|I(B(s))| \leqslant C(L(I), M, \alpha)$ and $\|H(B(s))\| \leqslant C(L(h), M, \alpha)$.

Last, by taking the supremum over $z \in \Gamma$ and then the supremum over $0 \leqslant t \leqslant (\hat{\tau}_M \wedge T)$ on the left-hand-side of (67) and then using the bounds (65) and (66), yields the result. □

It remains to verify the approximate-solution property for the two processes

$$
t \mapsto S(\mathcal{X}_t, z) \qquad \text{and} \qquad t \mapsto S(x_{\lfloor td \rfloor}, z).
$$

We do this by applying a Doob decomposition to each process and showing that the resulting martingale terms are negligible.

To do so, it will be convenient to work directly with the stopped process $x_{t \wedge \hat{\tau}_M}$ on the iterates. Since $\hat{\tau}_M$ is a time based on S-values, it is often difficult to apply to iterates of SGD and homogenized SGD, so we introduce equivalent stopping times

$$
\begin{aligned}
\vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d}\left\|W_{\lfloor td \rfloor}\right\|^2 > M \text{ or } B\left(x_{\lfloor td \rfloor}\right) \notin \mathcal{U}\} \quad \text{or} \\
\vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d}\left\|\mathcal{W}_t\right\|^2 > M \text{ or } B\left(\mathcal{X}_t\right) \notin \mathcal{U}\}.
\end{aligned}
\tag{68}
$$

We overload the notation ϑ_M to be either applied to SGD iterates, $x_{\lfloor td \rfloor}$ or homogenized SGD iterates, $\mathcal{X}_t$, for which it will be made clear in the context which criterion is used. These stopping times are equivalent to $\hat{\tau}_M$ (see Lemma C.1). Moreover, we often drop the M so that $\vartheta = \vartheta_M$. It will be convenient to work with the stopped processes, $x^\vartheta_{\lfloor td \rfloor} = x_{\lfloor d(t \wedge \vartheta) \rfloor}$ and $\mathcal{X}^\vartheta_t = \mathcal{X}_{t \wedge \vartheta}$.

D.1 Homogenized SGD Under the Resolvent Statistic S

We first verify the approximate-solution property for the homogenized process. Specifically, we show that the resolvent statistic $t \mapsto S(\mathcal{X}_t, z)$ satisfies the partial integro-differential equation (30) up to a martingale error, and that this martingale error is negligible uniformly on the fixed contour.

With this, we recall *homogenized SGD*

$$\mathrm{d}\mathcal{X}_t = -\gamma(t) d \nabla \mathcal{R}(\mathcal{X}_t)\mathrm{d}t + \gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)}\,\left(\nabla\psi(\mathcal{X}_t)\right)^\top \sqrt{K}\,\mathrm{d}\mathfrak{B}_t,$$

where $\mathcal{X}_t$ is a stochastic process taking values in $\mathbb{R}^{2d}$ with initial conditions $\mathcal{X}_0 = x_0$, and $\mathrm{d}\mathfrak{B}_t$ is the differential of a standard Brownian motion in $\mathbb{R}^d$.

Along the homogenized trajectory, write

$$\mathcal{W}_t := [\,\psi\left(\mathcal{X}_t\right) \mid \beta^* \mid \mathbb{1}_d\,] \in \mathbb{R}^{d\times 3}, \quad \text{and} \quad \rho_t := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(\mathcal{X}_t) \\ \beta^* \end{pmatrix}^\top a \in \mathbb{R}^2\,.$$

We study the homogenized process through the resolvent statistic

$$x \in \mathbb{R}^{2d} \mapsto S(x,z) = \frac{1}{d} W(x)^\top \Omega(x,z) W(x) \in \mathbb{C}^{3\times 3} \quad \text{for} \quad z \in \Gamma \subset \mathbb{C}^4.$$

We will show that $S\left(\mathcal{X}_t, z\right)$ is an approximate solution (C.2) to the partial integro-differential equation (30) which we state below.

Proposition D.2 (Homogenized SGD is an approximate solution)**.** *Fix $T, M > 0$ and $0 < \delta < 1/2$. Then $S\left(\mathcal{X}_t, z\right)$ is a $(d^{-\delta}, M, T)$-approximate solution w.o.p., that is,*

$$\sup_{0\leqslant t\leqslant(\hat{\tau}_M \wedge T)} \left\| S(\mathcal{X}_t,\cdot) - S(x_0,\cdot) - \int_0^t \mathcal{F}\left(\cdot, S(\mathcal{X}_s,\cdot)\right)\mathrm{d}s \right\|_\Gamma \leqslant d^{-\delta} \quad \textit{w.o.p.}$$

The proof of this Proposition is deferred to Section D.1.2.

D.1.1 Doob Decomposition for Homogenized SGD

We begin by applying Itô calculus to homogenized SGD under smooth statistics $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$. Later, when applying the resulting identity to $S(\cdot, z)$, we apply it separately to $S_{ab}^{\mathrm{Re},z}$ and $S_{ab}^{\mathrm{Im},z}$ for each matrix entry.

Applying Itô's lemma, we deduce that

$$\begin{aligned}
\mathrm{d}\varphi(\mathcal{X}_t) &= \langle \nabla\varphi(\mathcal{X}_t), \mathrm{d}\mathcal{X}_t\rangle + \frac{1}{2}\langle \nabla^2\varphi(\mathcal{X}_t), (\mathrm{d}\mathcal{X}_t)^{\otimes 2}\rangle \\
&= -\gamma(t) d \langle \nabla\varphi(\mathcal{X}_t), \nabla\mathcal{R}(\mathcal{X}_t)\rangle\,\mathrm{d}t + \gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)}\langle \nabla\varphi(\mathcal{X}_t), \left(\nabla\psi(\mathcal{X}_t)\right)^\top\sqrt{K}\,\mathrm{d}\mathfrak{B}_t\rangle \\
&\quad + \frac{\gamma(t)^2}{2} I\left(B(\mathcal{X}_t)\right)\langle \nabla^2\varphi(\mathcal{X}_t), \left(\left(\nabla\psi(\mathcal{X}_t)\right)^\top\sqrt{K}\,\mathrm{d}\mathfrak{B}_t\right)^{\otimes 2}\rangle.
\end{aligned} \tag{69}$$

We seek to simplify some of the terms in (69). For this, we flatten the second term in sum:

$$\langle \nabla\varphi(\mathcal{X}_t), \left(\nabla\psi(\mathcal{X}_t)\right)^\top\sqrt{K}\,\mathrm{d}\mathfrak{B}_t\rangle = \langle \left(\nabla\psi(\mathcal{X}_t)\right)^\top\sqrt{K}, \nabla\varphi(\mathcal{X}_t)\left(\mathrm{d}\mathfrak{B}_t\right)^\top\rangle_{\mathbb{R}^{2d\times d}}. \tag{70}$$

Next, we look at the second derivative term of φ,

$$\langle \nabla^2\varphi(\mathcal{X}_t), \left(\left(\nabla\psi(\mathcal{X}_t)\right)^\top\sqrt{K}\,\mathrm{d}\mathfrak{B}_t\right)^{\otimes 2}\rangle = \langle \nabla^2\varphi(\mathcal{X}_t), \left(\nabla\psi(\mathcal{X}_t)\right)^\top K \nabla\psi(\mathcal{X}_t)\rangle\,\mathrm{d}t, \tag{71}$$

where we used the symmetry of $\sqrt{K}$. With this, we can now identify the martingale increment for homogenized SGD,

$$\begin{aligned}
\mathrm{d}\varphi(\mathcal{X}_t) = &-\gamma(t) d \langle \nabla\varphi(\mathcal{X}_t), \nabla\mathcal{R}(\mathcal{X}_t)\rangle\,\mathrm{d}t \\
&+ \frac{\gamma(t)^2}{2} I\left(B(\mathcal{X}_t)\right)\langle \nabla^2\varphi(\mathcal{X}_t), \left(\nabla\psi(\mathcal{X}_t)\right)^\top K\nabla\psi(\mathcal{X}_t)\rangle\,\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(\varphi),
\end{aligned}$$

where $\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(\varphi) := \gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)}\langle(\nabla\psi(\mathcal{X}_t))^\top\sqrt{K}, \nabla\varphi(\mathcal{X}_t)\left(\mathrm{d}\mathfrak{B}_t\right)^\top\rangle_{\mathbb{R}^{2d\times d}}$. By integrating, we derive the Doob decomposition for $\varphi(\mathcal{X}_t)$

$$\begin{aligned}\varphi(\mathcal{X}_t) = \varphi(x_0) &- \int_0^t \gamma(s) d\langle\nabla\varphi(\mathcal{X}_s), \nabla\mathcal{R}(\mathcal{X}_s)\rangle\,\mathrm{d}s \\ &+ \frac{1}{2}\int_0^t \gamma(s)^2 I\left(B(\mathcal{X}_s)\right)\langle\nabla^2\varphi(\mathcal{X}_s), (\nabla\psi(\mathcal{X}_s))^\top K\nabla\psi(\mathcal{X}_s)\rangle\,\mathrm{d}s + \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(\varphi).\end{aligned} \tag{72}$$

D.1.2 $S(\mathcal{X}_t, z)$ is an Approximate Solution, Proof of Proposition D.2

The goal in this section is to prove Proposition D.2, that is, show that

$$S(\mathcal{X}_t, z) = \frac{1}{d}\mathcal{W}_t^\top\Omega(x, z)\mathcal{W}_t \in \mathbb{C}^{3\times 3}$$

is an approximate solution (C.2) to the partial integro-differential equation in (30).

The first step is to derive a closed equation for $S(\mathcal{X}_t, z)$ using Itô calculus.

Itô calculus applied to $S(\mathcal{X}_t, z)$. Recall the expected risk $\mathcal{R}$ can be expressed as a composition, $\mathcal{R}(\mathcal{X}_t) = h\left(B(\mathcal{X}_t)\right)$, for some function $h\colon\ \mathbb{R}^{3\times 3} \to \mathbb{R}$ and

$$B(\mathcal{X}_t) = \frac{1}{d}\mathcal{W}_t^\top K\mathcal{W}_t \in \mathbb{R}^{3\times 3}.$$

A straightforward application of the chain rule shows that

$$\begin{aligned}\mathbb{R}^{2d} \ni \nabla\mathcal{R}(\mathcal{X}_t) &= \langle\nabla B(\mathcal{X}_t), \nabla h\left(B(\mathcal{X}_t)\right)\rangle_{\mathbb{R}^{3\times 3}} \\ &= \frac{2}{d}\left(\nabla h_{11}\cdot(\nabla\psi(\mathcal{X}_t))^\top K\psi(\mathcal{X}_t) + \nabla h_{21}\cdot(\nabla\psi(\mathcal{X}_t))^\top K\beta^* + \nabla h_{31}\cdot(\nabla\psi(\mathcal{X}_t))^\top K\mathbb{1}_d\right).\end{aligned}$$

Using the product rule for Itô derivatives, we obtain

$$\begin{aligned}\mathrm{d}S =& -\gamma(t)d\langle\nabla S(\mathcal{X}_t, z), \nabla\mathcal{R}(\mathcal{X}_t)\rangle\,\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S) \\ &+ \frac{\gamma(t)^2}{2} I\left(B(\mathcal{X}_t)\right)\langle\nabla^2 S(\mathcal{X}_t, z), (\nabla\psi(\mathcal{X}_t))^\top K\nabla\psi(\mathcal{X}_t)\rangle\,\mathrm{d}t \\ =& -\gamma(t)d\left(\langle\nabla_u S(\mathcal{X}_t, z), \nabla_u\mathcal{R}(\mathcal{X}_t)\rangle + \langle\nabla_v S(\mathcal{X}_t, z), \nabla_v\mathcal{R}(\mathcal{X}_t)\rangle\right)\mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S) \\ &+ \frac{\gamma(t)^2}{2} I\left(B(\mathcal{X}_t)\right)\Big[\langle\nabla_u^2 S(\mathcal{X}_t, z), K\left(\nabla_u\psi(\mathcal{X}_t)\right)^2\rangle \\ &+ 2\langle\nabla_{uv}^2 S(\mathcal{X}_t, z), (\nabla_u\psi(\mathcal{X}_t))^\top K\nabla_v\psi(\mathcal{X}_t)\rangle + \langle\nabla_v^2 S(\mathcal{X}_t, z), K\left(\nabla_v\psi(\mathcal{X}_t)\right)^2\rangle\Big]\mathrm{d}t.\end{aligned} \tag{73}$$

Remark D.3 (Applying scalar formulas to S)**.** The statistic $S(x, z)$ is complex-valued:

$$S(x, z) \in \mathbb{C}^{3\times 3}.$$

However, all scalar statistic formulas above were stated for real-valued functions $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$. Therefore, whenever we apply these formulas to S, we apply them to the real-valued coordinate functions

$$S_{ab}^{\mathrm{Re},z}(x) := \mathrm{Re}\, S_{ab}(x, z), \qquad S_{ab}^{\mathrm{Im},z}(x) := \mathrm{Im}\, S_{ab}(x, z),$$

for each $a, b \in \{1, 2, 3\}$ and fixed $z \in \Gamma$.

For example, we define the complex-valued HSGD martingale associated with S entrywise by

$$\left(\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S)(z)\right)_{ab} := \mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}\left(S_{ab}^{\mathrm{Re},z}\right) + i\,\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}\left(S_{ab}^{\mathrm{Im},z}\right)$$

for

$$\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S_{ab}^{\mathrm{Re},z}) := \gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)}\langle(\nabla\psi(\mathcal{X}_t))^\top\sqrt{K}, \nabla S_{ab}^{\mathrm{Re},z}(\mathcal{X}_t, z)\left(\mathrm{d}\mathfrak{B}_t\right)^\top\rangle_{\mathbb{R}^{2d\times d}},$$

and analogously for $S_{ab}^{\mathrm{Im},z}$.

Thus matrix-valued stochastic identities involving S are shorthand for the collection of real-valued identities for the real and imaginary parts of its entries.

Similarly, we define

$$\mathcal{M}_t^{\mathrm{HSGD}}(S) = \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(S).$$

We consider the first term in the summation above, and plugging in $\nabla\mathcal{R}$ and using Lemma B.3, we have

$$\begin{aligned}
\mathbb{C}^{3\times 3} \ni \langle \nabla_u S(\mathcal{X}_t, z), \nabla_u \mathcal{R}(\mathcal{X}_t)\rangle_{\mathbb{R}^d} &= \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad (\nabla_u\psi(\mathcal{X}_t))^2 K\Omega \begin{bmatrix}\psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \\
&+ \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad (\nabla_u\psi(\mathcal{X}_t))^2 K\Omega \begin{bmatrix}\psi(\mathcal{X}_t) & 0 & 0\\ \beta^* & 0 & 0\\ \mathbb{1}_d & 0 & 0\end{bmatrix} \\
&+ \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\quad \nabla_u\psi(\mathcal{X}_t)\,\mathrm{diag}(\psi(\mathcal{X}_t))KR\left(z_1;\mathrm{diag}(\mathcal{U}_t)\right)\Omega \begin{bmatrix}\psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \\
&+ \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\quad \nabla_u\psi(\mathcal{X}_t)\,\mathrm{diag}(\beta^*)KR\left(z_1;\mathrm{diag}(\mathcal{U}_t)\right)\Omega \begin{bmatrix}0 & 0 & 0\\ \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d\\ 0 & 0 & 0\end{bmatrix} \\
&+ \frac{2}{d^2}(\nabla h_{11}\cdot\psi(\mathcal{X}_t) + \nabla h_{21}\cdot\beta^* + \nabla h_{31}\cdot \mathbb{1}_d)^\top\cdot \\
&\qquad \nabla_u\psi(\mathcal{X}_t)KR\left(z_1;\mathrm{diag}(\mathcal{U}_t)\right)\Omega \begin{bmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ \psi(\mathcal{X}_t) & \beta^* & \mathbb{1}_d\end{bmatrix} \\
&= \frac{2}{d}H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^2 K\Omega\mathcal{W}_t \\
&\quad + \frac{2}{d^2}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^2 K\Omega\mathcal{W}_t\cdot H \\
&+ \frac{2}{d}H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\nabla_u\psi(\mathcal{X}_t)\,\mathrm{diag}(\psi(\mathcal{X}_t))KR\left(z_1;\mathrm{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t \\
&+ \frac{2}{d}E_{21}\cdot H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\nabla_u\psi(\mathcal{X}_t)\,\mathrm{diag}(\beta^*)KR\left(z_1;\mathrm{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t \\
&+ \frac{2}{d}E_{31}\cdot H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\nabla_u\psi(\mathcal{X}_t)KR\left(z_1;\mathrm{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t,
\end{aligned}$$

$$\begin{aligned}
\mathbb{C}^{3\times 3} \ni \langle \nabla_v S(\mathcal{X}_t, z), \nabla_v \mathcal{R}(\mathcal{X}_t)\rangle_{\mathbb{R}^d} &= \frac{2}{d}H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_v\psi(\mathcal{X}_t)\right)^2 K\Omega\mathcal{W}_t \\
&\quad + \frac{2}{d^2}\mathcal{W}_t^\top\left(\nabla_v\psi(\mathcal{X}_t)\right)^2 K\Omega\mathcal{W}_t\cdot H \\
&+ \frac{2}{d}H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\nabla_v\psi(\mathcal{X}_t)\,\mathrm{diag}(\psi(\mathcal{X}_t))KR\left(z_2;\mathrm{diag}(\mathcal{V}_t)\right)\Omega\mathcal{W}_t \\
&+ \frac{2}{d}E_{21}\cdot H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\nabla_v\psi(\mathcal{X}_t)\,\mathrm{diag}(\beta^*)KR\left(z_2;\mathrm{diag}(\mathcal{V}_t)\right)\Omega\mathcal{W}_t
\end{aligned}$$

$$+\frac{2}{d}E_{31}\cdot H^\top\cdot\frac{1}{d}\mathcal{W}_t^\top\nabla_v\psi(\mathcal{X}_t)KR\left(z_2;\operatorname{diag}(\mathcal{V}_t)\right)\Omega\mathcal{W}_t,$$

for

$$H\left(B(\mathcal{X}_t)\right)=\left[\begin{array}{c|c|c}\nabla_{11}h & 0 & 0\\\hline \nabla_{21}h & 0 & 0\\\hline \nabla_{31}h & 0 & 0\end{array}\right].$$

Similarly for the second term we have

$$\begin{aligned}\mathbb{C}^{3\times3}\ni\langle\nabla_u^2S(\mathcal{X}_t,z),K\left(\nabla_u\psi(x)\right)^2\rangle_{\mathbb{R}^{d\times d}}&=\frac{2q_{11}}{d}\mathbb{1}_d^\top\cdot\left(\nabla_u\psi(x)\right)^2K\Omega\begin{bmatrix}\psi(\mathcal{X}_t)&\beta^*&\mathbb{1}_d\\0&0&0\\0&0&0\end{bmatrix}\\&+\frac{2q_{11}}{d}\mathbb{1}_d^\top\cdot\left(\nabla_u\psi(x)\right)^2K\Omega\begin{bmatrix}\psi(\mathcal{X}_t)&0&0\\\beta^*&0&0\\\mathbb{1}_d&0&0\end{bmatrix}\\&+\frac{2}{d}\mathbb{1}_d^\top\cdot\left(\nabla_u\psi(x)\right)^3KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\begin{bmatrix}\psi(\mathcal{X}_t)&\beta^*&\mathbb{1}_d\\0&0&0\\0&0&0\end{bmatrix}\\&+\frac{2}{d}\mathbb{1}_d^\top\cdot\left(\nabla_u\psi(x)\right)^3KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\begin{bmatrix}\psi(\mathcal{X}_t)&0&0\\\beta^*&0&0\\\mathbb{1}_d&0&0\end{bmatrix}\\&+\frac{2}{d}\begin{bmatrix}\operatorname{Tr}\left(\left(\nabla_u\psi(x)\right)^4K\Omega\right)&0&0\\0&0&0\\0&0&0\end{bmatrix}\\&+\frac{2}{d}(\psi(\mathcal{X}_t))^\top\cdot\left(\nabla_u\psi(x)\right)^2KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\begin{bmatrix}\psi(\mathcal{X}_t)&\beta^*&\mathbb{1}_d\\0&0&0\\0&0&0\end{bmatrix}\\&+\frac{2}{d}(\beta^*)^\top\cdot\left(\nabla_u\psi(x)\right)^2KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\begin{bmatrix}0&0&0\\\psi(\mathcal{X}_t)&\beta^*&\mathbb{1}_d\\0&0&0\end{bmatrix}\\&+\frac{2}{d}\mathbb{1}_d^\top\cdot\left(\nabla_u\psi(x)\right)^2KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\begin{bmatrix}0&0&0\\0&0&0\\\psi(\mathcal{X}_t)&\beta^*&\mathbb{1}_d\end{bmatrix}\end{aligned}$$

$$\begin{aligned}&=2q_{11}E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^2K\Omega\mathcal{W}_t\\&+2q_{11}\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^2K\Omega\mathcal{W}_t\cdot E_{31}\\&+2E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^3KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t\\&+2\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^3KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t\cdot E_{31}\\&+2E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)^4K\Omega\mathcal{W}_t\cdot E_{31}\\&+2\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(x)\right)^2KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)^2\Omega\mathcal{W}_t,\end{aligned}$$

$$\begin{aligned}\mathbb{C}^{3\times3}\ni\langle\nabla_{uv}^2S(\mathcal{X}_t,z),\left(\nabla_u\psi(x)\right)^\top K\nabla_v\psi(x)\rangle_{\mathbb{R}^{d\times d}}=&\\2q_{12}E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)\left(\nabla_v\psi(\mathcal{X}_t)\right)K\Omega\mathcal{W}_t&\\+2q_{12}\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)\left(\nabla_v\psi(\mathcal{X}_t)\right)K\Omega\mathcal{W}_t\cdot E_{31}&\\+E_{13}\cdot\frac{1}{d}\mathcal{W}_t^\top\left(\nabla_u\psi(\mathcal{X}_t)\right)\left(\nabla_v\psi(\mathcal{X}_t)\right)^2KR\left(z_1;\operatorname{diag}(\mathcal{U}_t)\right)\Omega\mathcal{W}_t&\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right) \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 KR\left(z_1; \operatorname{diag}(\mathcal{U}_t)\right) \Omega \mathcal{W}_t \cdot E_{31} \\
&+ E_{13} \cdot \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right)^2 \left(\nabla_v \psi(\mathcal{X}_t)\right) KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \\
&+ \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right)^2 \left(\nabla_v \psi(\mathcal{X}_t)\right) KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2E_{13} \cdot \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right)^2 \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \cdot E_{31} \\
&+ \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_u \psi(\mathcal{X}_t)\right) \left(\nabla_v \psi(\mathcal{X}_t)\right) KR\left(z_1; \operatorname{diag}(\mathcal{U}_t)\right) R\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{C}^{3\times 3} \ni \langle \nabla_v^2 S(\mathcal{X}_t, z), K\left(\nabla_v \psi(x)\right)^2 \rangle_{\mathbb{R}^{d\times d}} &= 2q_{22} E_{13} \cdot \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \\
&+ 2q_{22} \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^2 K \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2E_{13} \cdot \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^3 KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \\
&+ 2\frac{1}{d}\mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^3 KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right) \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2E_{13} \cdot \frac{1}{d}\mathcal{W}_t^\top \left(\nabla_v \psi(\mathcal{X}_t)\right)^4 K \Omega \mathcal{W}_t \cdot E_{31} \\
&+ 2\frac{1}{d}\mathcal{W}_t^\top \left(\nabla_v \psi(x)\right)^2 KR\left(z_2; \operatorname{diag}(\mathcal{V}_t)\right)^2 \Omega \mathcal{W}_t.
\end{aligned}
$$

Now recall that

$$
\nabla \psi(\mathcal{X}_t) = \left[2q_{11} \operatorname{diag}(\mathcal{U}_t) + 2q_{12} \operatorname{diag}(\mathcal{V}_t) + l_1 I_d \; 2q_{12} \operatorname{diag}(\mathcal{U}_t) + 2q_{22} \operatorname{diag}(\mathcal{V}_t) + l_2 I_d\right] \in \mathbb{R}^{d\times 2d}.
$$

Accordingly, each term in (73) takes a form such as

$$
\begin{aligned}
&- 2\gamma(t) H^\top \cdot \frac{1}{d}\mathcal{W}_t^\top \operatorname{diag}(\mathcal{U}_t)^{m_1} R\big(z_1; \operatorname{diag}(\mathcal{U}_t)\big)^{\rho_1} \\
&\quad \cdot \operatorname{diag}(\mathcal{V}_t)^{m_2} R\big(z_2; \operatorname{diag}(\mathcal{V}_t)\big)^{\rho_2} \operatorname{diag}(\beta^*)^{m_3} R\big(z_3; \operatorname{diag}(\beta^*)\big) KR\big(z_4; K\big) \mathcal{W}_t, \\
\text{or} \quad &C(Q)\gamma(t)^2 I\big(B(\mathcal{X}_t)\big) \cdot \frac{1}{d}\mathcal{W}_t^\top \operatorname{diag}(\mathcal{U}_t)^{m_1} R\big(z_1; \operatorname{diag}(\mathcal{U}_t)\big)^{\rho_1} \\
&\quad \cdot \operatorname{diag}(\mathcal{V}_t)^{m_2} R\big(z_2; \operatorname{diag}(\mathcal{V}_t)\big)^{\rho_2} R\big(z_3; \operatorname{diag}(\beta^*)\big) KR\big(z_4; K\big) \mathcal{W}_t.
\end{aligned}
$$

Expanding these terms and applying the Cauchy integral formula along each contour Γ_i, we write

$$
q(D_i) = \frac{1}{2\pi i} \oint_{\Gamma_i} q(z_i) R\left(z_i; D_i\right) \, \mathrm{d}z_i,
$$

together with the standard resolvent identities,

$$
R\left(z_i; D_i\right)^2 = -\frac{\mathrm{d}}{\mathrm{d}z_i} R\left(z_i; D_i\right) \quad \text{and} \quad R\left(z_i; D_i\right)^3 = \frac{1}{2} \cdot \frac{\mathrm{d}^2}{\mathrm{d}z_i^2} R\left(z_i; D_i\right).
$$

It then follows that

$$
\mathrm{d}S(\mathcal{X}_t, z) = \mathcal{F}\left(z, S(\mathcal{X}_t, z)\right) \mathrm{d}t + \mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(S), \tag{74}
$$

with $\mathcal{X}_0 = x_0$, where the summands of $\mathcal{F}\left(z, S(\mathcal{X}_t, z)\right)$ are of the form

$$
\begin{aligned}
&-2\gamma(t) H^\top \cdot \frac{1}{d}\mathcal{W}_t^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right) \left(R\left(z_i; D_i\right)\right) \mathcal{W}_t \quad \text{or} \\
&C(Q)\gamma(t)^2 I(B(\mathcal{X}_t)) \cdot \frac{1}{d}\mathcal{W}_t^\top z^\zeta \left(\bigcirc_{i=1}^4 \mathcal{G}_i\right) \left(R\left(z_i; D_i\right)\right) \mathcal{W}_t,
\end{aligned} \tag{75}
$$

for

$$
\mathcal{G}_i(R\left(z_i; D_i\right)) = \frac{1}{2\pi i} \oint_{\Gamma_i} q(z_i) R\left(z_i; D_i\right) \, \mathrm{d}z_i, \quad -\frac{\mathrm{d}}{\mathrm{d}z_i} R\left(z_i; D_i\right), \quad \text{or} \quad \frac{1}{2} \cdot \frac{\mathrm{d}^2}{\mathrm{d}z_i^2} R\left(z_i; D_i\right). \tag{76}
$$

We now prove Proposition D.2.

Proof of Proposition D.2. By Itô's Lemma, we have seen that

$$S(\mathcal{X}_t, \cdot) = S(x_0, \cdot) + \int_0^t \mathcal{F}\left(\cdot, S(\mathcal{X}_s, \cdot)\right) \mathrm{d}s + \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(S(\mathcal{X}_s, \cdot)).$$

Thus to show that $S(\mathcal{X}_t, \cdot)$ is an approximate solution of the partial integro-differential equation (30) it amounts to bounding the martingale term where C is a positive constant independent of d. For all $z \in \Gamma$, we note that for some constants $C, c > 0$ we have $\vartheta_{c \cdot M} \leqslant \hat{\tau}_M \leqslant \vartheta_{C \cdot M}$ (see Lemma C.1). Consequently, we can work with the stopped process $\mathcal{X}_t^{\vartheta} = \mathcal{X}_{t \wedge \vartheta}$ instead of using $\hat{\tau}_M$. We thus have that for all $z \in \Gamma$

$$\sup_{0 \leqslant t \leqslant (\hat{\tau}_M \wedge T)} \left\| S(\mathcal{X}_t, z) - S(x_0, z) - \int_0^t \mathcal{F}\left(z, S(\mathcal{X}_s, z)\right) \, \mathrm{d}s \right\| \leqslant \sup_{0 \leqslant t \leqslant (\vartheta_{C \cdot M} \wedge T)} \left\| \mathcal{M}_t^{\mathrm{HSGD}}(S(\cdot, z)) \right\|.$$

Fix a constant $\delta > 0$. Let $\Gamma^\delta = \Gamma_1^\delta \times \Gamma_2^\delta \times \Gamma_3^\delta \times \Gamma_4^\delta$ where each Γ_i^δ is a $d^{-\delta}$-mesh of Γ_i with $|\Gamma^\delta| \leqslant C_\Gamma d^{4\delta}$ for positive $C_\Gamma > 0$ depending on $\|K\|_{\mathrm{op}}$ and $\|\beta^*\|_\infty$.

By the martingale error proposition, Proposition D.8, which we have deferred the proof to Section D.3.1, we have that for any $\hat{\delta} > 0$

$$\sup_{0 \leqslant t \leqslant T} \left\| \mathcal{M}_{t \wedge \vartheta_{C \cdot M}}^{\mathrm{HSGD}}(S(\cdot, z)) \right\| \leqslant CL(f) d^{\hat{\delta}/2 - 1/2} \quad \text{w.o.p.}$$

By a union bound over $z \in \Gamma^\delta$, using $|\Gamma^\delta| \leqslant C_\Gamma d^{4\delta}$, the martingale bound holds uniformly on Γ^δ with overwhelming probability:

$$\sup_{z \in \Gamma^\delta} \sup_{0 \leqslant t \leqslant T} \left\| \mathcal{M}_{t \wedge \vartheta_{C \cdot M}}^{\mathrm{HSGD}}(S(\cdot, z)) \right\| \leqslant CL(f) d^{\hat{\delta}/2 - 1/2} \quad \text{w.o.p.}$$

Consequently, we deduce that

$$\begin{aligned}
\sup_{0 \leqslant t \leqslant (\hat{\tau}_M \wedge T)} \left\| S(\mathcal{X}_t, z) - S(x_0, z) - \int_0^t \mathcal{F}\left(z, S(\mathcal{X}_s, z)\right) \, \mathrm{d}s \right\|_{\Gamma^\delta} & \\
\leqslant \sup_{0 \leqslant t \leqslant (\vartheta_{C \cdot M} \wedge T)} \left\| \mathcal{M}_t^{\mathrm{HSGD}}(S(\cdot, z)) \right\|_{\Gamma^\delta} & \\
\leqslant CL(f) d^{\hat{\delta}/2 - 1/2} \quad \text{w.o.p.} &
\end{aligned}$$

An application of the net argument, Lemma D.1, finishes the proof after setting $\hat{\delta} = 1 - 2\delta$.

The derivative regularity condition (i) is automatic for $t \mapsto S(\mathcal{X}_t, \cdot)$: for each fixed trajectory, the dependence on z is through a finite product of resolvents on the fixed contour Γ, and the resolvent derivative bounds from Remark C.3 give the required Lipschitz control up to $\hat{\tau}_M$. Thus the estimate above verifies condition (ii), and hence $S(\mathcal{X}_t, \cdot)$ is an $(d^{-\varepsilon}, M, T)$-approximate solution. □

D.2 SGD Under the Resolvent Statistic S

In this section, we show that $S\left(x_{\lfloor td \rfloor}, z\right)$ is an approximate solution (C.2) to the partial integro-differential equation (30) which we state below.

Proposition D.4 (SGD is an approximate solution)**.** *Fix a $T, M > 0$ and $0 < \delta < 1/2$. Then $S\left(x_{\lfloor td \rfloor}, z\right)$ is a $(d^{-\delta}, M, T)$-approximate solution w.o.p., that is,*

$$\sup_{0 \leqslant t \leqslant (\hat{\tau}_M \wedge T)} \left\| S(x_{\lfloor td \rfloor}, z) - S(x_0, z) - \int_0^t \mathcal{F}\left(z, S(x_{\lfloor sd \rfloor}, z)\right) \, \mathrm{d}s \right\| \leqslant d^{-\delta} \quad \textit{w.o.p.}$$

The proof of this Proposition is deferred to Section D.2.2.

Recall our iterates satisfy the recurrence

$$x_{k+1} = x_k - \gamma_k \nabla_x \Psi(x_k; a_{k+1}) = x_k - \frac{\gamma_k}{\sqrt{d}} \nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla \psi(x_k)\right)^\top \cdot a_{k+1} \tag{77}$$

with $r_k := \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix}^\top a_{k+1}$.

The derivation follows the martingale-decomposition strategy used in [Collins-Woodfin et al., 2024a]. We first derive a one-step martingale decomposition for a smooth scalar statistic $\varphi : \mathbb{R}^{2d} \to \mathbb{R}$. We later apply this decomposition to the real-valued coordinate functions $S_{ab}^{\mathrm{Re},z}$ and $S_{ab}^{\mathrm{Im},z}$ and then recombine the resulting identities to obtain the complex matrix-valued decomposition for $S(x_{\lfloor td \rfloor}, z)$.

Applying Taylor's expansion, we have

$$\begin{aligned}
\varphi(x_{k+1}) &= \varphi(x_k) - \gamma_k \langle \nabla\varphi(x_k), \nabla_x \Psi(x_k; a_{k+1}) \rangle + \frac{\gamma_k^2}{2} \langle \nabla^2 \varphi(x_k), (\nabla_x \Psi(x_k; a_{k+1}))^{\otimes 2} \rangle + \dots \\
&= \varphi(x_k) - \frac{\gamma_k}{\sqrt{d}} \langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1} \rangle \\
&+ \frac{\gamma_k^2}{2d} \langle \nabla^2 \varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right)^{\otimes 2} \rangle + \mathcal{E}_k^{\mathrm{High}}(\varphi).
\end{aligned} \tag{78}$$

We will show in Section D.3.3 that the third and higher-order terms, $\mathcal{E}_k^{\mathrm{High}}(\varphi)$, vanish when $d \to \infty$.

D.2.1 Doob Decomposition for SGD

Recall that, for a Hessian A and vector v,

$$\langle A, v^{\otimes 2} \rangle = v^\top A v.$$

To write the Doob decomposition, the idea is to condition on $r_k = \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix}^\top a_{k+1}$ and $W_k = [\, \psi(x_k) \,|\, \beta^* \,|\, \mathbb{1}_d \,] \in \mathbb{R}^{d \times 3}$. For this, we will introduce some notation. Define the σ-algebras

$$\mathcal{F}_k := \sigma\left(\{W_i\}_{i=0}^k\right) \subset \mathcal{G}_k := \sigma\left(\{r_i\}_{i=0}^k, \{W_i\}_{i=0}^k\right).$$

Gradient term in Taylor expansion. First, consider the conditional expectation with respect to $\mathcal{F}_k$ of the gradient term in equation (78). Note we can safely assume the interchange of the gradient and the expectation in the risk formula $\mathcal{R}$ since a follows a Gaussian distribution, and generally, this assumption holds true. Therefore, we have

$$\frac{1}{\sqrt{d}} \mathbb{E}_{a_{k+1}} \left[\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1} \rangle \middle| \mathcal{F}_k \right] = \langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k) \rangle.$$

By applying Doob's decomposition in this context, the gradient term takes the form

$$\frac{\gamma_k}{\sqrt{d}} \langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1} \rangle = \gamma_k \langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k) \rangle - \Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi) \tag{79}$$

where $\Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi)$ represents a martingale increment that tends to zero as d tends to infinity.

Hessian term in the Taylor expansion. Proceeding to the Hessian term in (78), we initiate the analysis by conditioning first on $\mathcal{G}_k$ and subsequently on $\mathcal{F}_k$, using the tower property. A simple computation leads to

$$\langle \nabla^2 \varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right) \cdot \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right)^{\otimes 2} \rangle = \langle \mathcal{B}, a_{k+1}^{\otimes 2} \rangle,$$

where $\mathcal{B} := \nabla_{r_1} f\left(r_k\right)^2 \nabla\psi(x_k) \nabla^2\varphi(x_k) \left(\nabla\psi(x_k)\right)^\top$. To proceed, we must evaluate

$$\begin{aligned}
\mathbb{E}_{a_{k+1}} \left[\langle \mathcal{B}, a_{k+1}^{\otimes 2} \rangle \middle| \mathcal{G}_k \right] &= \langle \mathcal{B}, \mathbb{E}_{a_{k+1}} \left[a_{k+1}^{\otimes 2} \middle| \mathcal{G}_k \right] \rangle \\
&= \langle \mathcal{B}, \mathbb{E}_{a_{k+1}} \left[\left(a_{k+1} - \mathbb{E}_{a_{k+1}}\left[a_{k+1} | \mathcal{G}_k\right]\right)^{\otimes 2} \middle| \mathcal{G}_k \right] \rangle \\
&+ \langle \mathcal{B}, \left(\mathbb{E}_{a_{k+1}}\left[a_{k+1} | \mathcal{G}_k\right]\right)^{\otimes 2} \rangle.
\end{aligned} \tag{80}$$

To establish the conditional distribution of a_{k+1} given r_k, we invoke a standard conditioning lemma, see e.g. [Collins-Woodfin et al., 2024a].

Lemma D.5 (Gaussian conditioning). *Let $v \sim \mathcal{N}(0, I_d)$ and let $Q \in \mathbb{R}^{d\times 2}$ have orthonormal columns. Then, conditional on $Q^\top v$, we have*

$$v \overset{\mathcal{D}}{=} (I_d - QQ^\top)g + QQ^\top v$$

where $g \sim \mathcal{N}(0, I_d)$ is independent of $Q^\top v$. In particular,

$$\mathbb{E}[v \mid Q^\top v] = QQ^\top v, \qquad \mathrm{Cov}(v \mid Q^\top v) = I_d - QQ^\top.$$

By Assumption 1.5, we write $a_{k+1} = \sqrt{K}v_k$ with $v_k \sim \mathcal{N}(0, I_d)$, and express $\sqrt{K}\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}$ as $Q_k R_k$, where $Q_k \in \mathbb{R}^{d\times 2}$ is orthogonal and $R_k \in \mathbb{R}^{2\times 2}$ is upper triangular and invertible (obtained through QR decomposition, assuming $2 \ll d$). Observe here $\Pi_k := Q_k Q_k^\top$ has rank 2.

Consequently, we can simplify $\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1}$ as follows

$$\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1} = \begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top \sqrt{K}v_k = R_k^\top Q_k^\top v_k.$$

Applying the conditioning lemma, we derive

$$\begin{aligned} a_{k+1}\Big|\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}^\top a_{k+1} &\overset{\mathcal{D}}{=} \sqrt{K}v_k | Q_k^\top v_k \\ &\overset{\mathcal{D}}{=} \sqrt{K}\left((I_d - Q_kQ_k^\top)g_k + Q_kQ_k^\top v_k\right) \\ &= \sqrt{K}\left(g_k - \Pi_k g_k\right) + \sqrt{K}\Pi_k v_k, \end{aligned}$$

where $g_k \sim \mathcal{N}(0, I_d)$ is independent of $Q_k^\top v_k$. From this, we have that

$$\mathbb{E}_{a_{k+1}}\left[a_{k+1}|\mathcal{G}_k\right] = \sqrt{K}\Pi_k v_k \quad \text{where } v_k \sim \mathcal{N}(0, I_d). \tag{81}$$

Moreover, the conditional covariance of a_{k+1} is precisely

$$\mathbb{E}_{a_{k+1}}\left[\left(a_{k+1} - \mathbb{E}_{a_{k+1}}\left[a_{k+1}|\mathcal{G}_k\right]\right)^{\otimes 2}\Big|\mathcal{G}_k\right] = \sqrt{K}\left(I_d - \Pi_k\right)\sqrt{K}. \tag{82}$$

Thus, from (80) we have

$$\mathbb{E}_{a_{k+1}}\left[\langle \mathcal{B}, a_{k+1}^{\otimes 2}\rangle\big|\mathcal{G}_k\right] = \langle \mathcal{B}, K\rangle - \langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle + \langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle. \tag{83}$$

We will later see, in Section D.3.4, that the term

$$\mathcal{E}_k^{\mathrm{Hess}}(\varphi) := -\frac{\gamma_k^2}{2d}\langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle + \frac{\gamma_k^2}{2d}\langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle$$

is of lower order in expectation and will disappear as $d \to \infty$. So we may write

$$\frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle \mathcal{B}, a_{k+1}^{\otimes 2}\rangle\big|\mathcal{F}_k\right] = \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle \mathcal{B}, K\rangle|\mathcal{F}_k\right] + \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}(\varphi)\big|\mathcal{F}_k\right].$$

Moreover, observe that

$$\begin{aligned} \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle \mathcal{B}, K\rangle|\mathcal{F}_k\right] &= \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\nabla_{r_1} f\left(r_k\right)^2\Big|\mathcal{F}_k\right]\langle \nabla\psi(x_k)\nabla^2\varphi(x_k)\left(\nabla\psi(x_k)\right)^\top, K\rangle \\ &= \frac{\gamma_k^2}{2d}\,I(B(x_k))\,\langle \nabla\psi(x_k)\nabla^2\varphi(x_k)\left(\nabla\psi(x_k)\right)^\top, K\rangle \\ &= \frac{\gamma_k^2}{2d}\,I(B(x_k))\,\langle \nabla^2\varphi(x_k), \langle K, \left(\nabla\psi(x_k)\right)^{\otimes 2}\rangle\rangle. \end{aligned} \tag{84}$$

By applying Doob's decomposition in this context, the Hessian term takes the form

$$\begin{aligned} &\frac{\gamma_k^2}{2d}\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle = \\ &\frac{\gamma_k^2}{2d}\,I(B(x_k))\,\langle \nabla^2\varphi(x_k), \langle K, \left(\nabla\psi(x_k)\right)^{\otimes 2}\rangle\rangle + \Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) + \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}(\varphi)\big|\mathcal{F}_k\right] \end{aligned} \tag{85}$$

where $\Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi)$ represents a martingale increment that tends to zero as d tends to infinity.

We have successfully identified the martingale increments of a *single* update of SGD, that is, by (79) and (85) in the Taylor expansion (78)

$$\begin{aligned}\varphi(x_{k+1}) = \varphi(x_k) &- \gamma_k \langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\rangle + \Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi) \\ &+ \frac{\gamma_k^2}{2d} I(B(x_k)) \langle \nabla^2\varphi(x_k), \langle K, (\nabla\psi(x_k))^{\otimes 2}\rangle\rangle + \Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) \\ &+ \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}(\varphi)\middle|\mathcal{F}_k\right] + \mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{High}}(\varphi)\middle|\mathcal{F}_k\right],\end{aligned} \tag{86}$$

where the error terms look like

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi) &= \gamma_k \langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\rangle - \frac{\gamma_k}{\sqrt{d}} \langle \nabla\varphi(x_k), \nabla_{r_1} f(r_k) \cdot (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \\ \Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) &= \frac{\gamma_k^2}{2d}\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f(r_k) \cdot (\nabla\psi(x_k))^\top \cdot a_{k+1}\right)^{\otimes 2}\rangle \\ &\quad - \frac{\gamma_k^2}{2d}\mathbb{E}_{a_{k+1}}\left[\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f(r_k) \cdot (\nabla\psi(x_k))^\top \cdot a_{k+1}\right)^{\otimes 2}\rangle\middle|\mathcal{F}_k\right] \\ \mathcal{E}_k^{\mathrm{Hess}}(\varphi) &= -\frac{\gamma_k^2}{2d}\langle \nabla_{r_1} f(r_k)^2 \nabla\psi(x_k)\nabla^2\varphi(x_k) (\nabla\psi(x_k))^\top, \sqrt{K}\Pi_k\sqrt{K}\rangle \\ &\quad + \frac{\gamma_k^2}{2d}\langle \nabla_{r_1} f(r_k)^2 \nabla\psi(x_k)\nabla^2\varphi(x_k) (\nabla\psi(x_k))^\top, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle.\end{aligned}$$

We now pass from the one-step decomposition to the continuous-time embedding. For $t \geqslant 0$, write

$$n_t := \lfloor td\rfloor, \qquad t_d := \frac{n_t}{d}.$$

Thus $0 \leqslant t - t_d < 1/d$ and $x_{\lfloor td\rfloor} = x_{n_t}$.

Define the predictable drift integrand

$$A_j(\varphi) := -d\,\gamma_j \langle \nabla\varphi(x_j), \nabla\mathcal{R}(x_j)\rangle + \frac{\gamma_j^2}{2} I(B(x_j)) \left\langle \nabla^2\varphi(x_j), \left\langle K, (\nabla\psi(x_j))^{\otimes 2}\right\rangle\right\rangle.$$

Equivalently, the predictable part of one SGD step is $(1/d)A_j(\varphi)$. Let

$$A_\varphi^{(d)}(s) := A_{\lfloor sd\rfloor}(\varphi)$$

be its piecewise-constant interpolation. Then

$$\sum_{j=0}^{n_t-1} \frac{1}{d} A_j(\varphi) = \int_0^{t_d} A_\varphi^{(d)}(s)\,\mathrm{d}s.$$

Therefore,

$$\sum_{j=0}^{n_t-1} \frac{1}{d} A_j(\varphi) = \int_0^{t} A_\varphi^{(d)}(s)\,\mathrm{d}s + \xi_t^{\mathrm{mesh}}(\varphi),$$

where the mesh error is

$$\xi_t^{\mathrm{mesh}}(\varphi) := -\int_{t_d}^{t} A_\varphi^{(d)}(s)\,\mathrm{d}s.$$

Since $t - t_d \leqslant 1/d$, we have

$$\left|\xi_t^{\mathrm{mesh}}(\varphi)\right| \leqslant \frac{1}{d} \sup_{0\leqslant s\leqslant T} \left|A_\varphi^{(d)}(s)\right|.$$

Using the one-step decomposition above, we obtain

$$\varphi(x_{\lfloor td\rfloor}) = \varphi(x_0) + \int_0^t A_\varphi^{(d)}(s)\,\mathrm{d}s + \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(\varphi) + \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(\varphi)$$

$$+\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(\varphi)\,\middle|\,\mathcal{F}_j\right]+\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{High}}(\varphi)\,\middle|\,\mathcal{F}_j\right]+\xi_t^{\mathrm{mesh}}(\varphi). \tag{87}$$

where

$$\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(\varphi):=\sum_{j=0}^{\lfloor td\rfloor-1}\Delta\mathcal{M}_j^{\mathrm{Grad}}(\varphi),\qquad \mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(\varphi):=\sum_{j=0}^{\lfloor td\rfloor-1}\Delta\mathcal{M}_j^{\mathrm{Hess}}(\varphi).$$

In Section D.3, we prove that the error terms in (87) are negligible as $d\to\infty$. The other two terms in $\int_0^t A_\varphi^{(d)}(s)\,\mathrm{d}s$ survive the limit. Next, we show that SGD on S is an (ε,M,T)-approximate solution.

D.2.2 $S\left(x_{\lfloor td\rfloor},z\right)$ is an Approximate Solution, Proof of Proposition D.4

The goal in this section is to prove Proposition D.4, that is, show that

$$S\left(x_{\lfloor td\rfloor},z\right)=\frac{1}{d}W_{\lfloor td\rfloor}^{\top}\Omega\left(x_{\lfloor td\rfloor},z\right)W_{\lfloor td\rfloor}\in\mathbb{C}^{3\times3}$$

is an approximate solution to the partial integro-differential equation (30).

Proof of Proposition D.4. Applying the preceding real-valued identity to the real and imaginary parts of each matrix entry of $S(\cdot,z)$, and then recombining, yields

$$\begin{aligned}S\left(x_{\lfloor td\rfloor},z\right)&=S\left(x_0,z\right)+\int_0^t\mathcal{F}\left(z,S(x_{\lfloor sd\rfloor},z)\right)\mathrm{d}s+\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(S)+\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(S)\\&\quad+\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(S)\,\middle|\,\mathcal{F}_j\right]+\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{High}}(S)\,\middle|\,\mathcal{F}_j\right]+\xi_t^{\mathrm{mesh}}(S).\end{aligned}$$

Thus to show that $S\left(x_{\lfloor td\rfloor},z\right)$ is an approximate solution of the partial integro-differential equation (30) it suffices to bound the martingales and error terms where C is a positive constant independent of d. We thus have for all $z\in\Gamma$, the estimate

$$\begin{aligned}&\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|S\left(x_{\lfloor td\rfloor},z\right)-S\left(x_0,z\right)-\int_0^t\mathcal{F}\left(z,S(x_{\lfloor sd\rfloor},z)\right)\mathrm{d}s\right\|\\&\qquad\leqslant\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Grad}}(S)\right\|+\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|\mathcal{M}_{\lfloor td\rfloor}^{\mathrm{Hess}}(S)\right\|\\&+\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(S)\middle|\mathcal{F}_j\right]\right\|+\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|\sum_{j=0}^{\lfloor td\rfloor-1}\mathbb{E}_{a_{j+1}}\left[\mathcal{E}_j^{\mathrm{High}}(S)\middle|\mathcal{F}_j\right]\right\|\\&+\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|\xi_t^{\mathrm{mesh}}(S)\right\|.\end{aligned}$$

Next, fix a constant $\delta>0$. Let $\Gamma^\delta=\Gamma_1^\delta\times\Gamma_2^\delta\times\Gamma_3^\delta\times\Gamma_4^\delta$ where each Γ_i^δ is a $d^{-\delta}$-mesh of Γ_i with $|\Gamma^\delta|\leqslant C_\Gamma d^{4\delta}$ for positive $C_\Gamma>0$ depending on $\|K\|_{\mathrm{op}}$ and $\|\beta^*\|_\infty$. For all $z\in\Gamma$, we note that for some constants $C,c>0$ we have $\vartheta_{c\cdot M}\leqslant\hat\tau_M\leqslant\vartheta_{C\cdot M}$ (see Lemma C.1). Consequently, we evaluate the error with the stopped process $x_{td}^{\vartheta}=x_{td\wedge\vartheta}$ instead of using $\hat\tau_M$. By the martingale error propositions, Proposition D.9 and D.10, the proofs of which are deferred to Section D.3.2, we have for any $\hat\delta>0$ the estimates

$$\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}^{\mathrm{Grad}}_{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor}(S(\cdot,z))\right\|<d^{-\frac{2+\alpha}{2}+\hat\delta}\quad\text{w.o.p.},$$

$$\text{and}\quad\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}^{\mathrm{Hess}}_{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor}(S(\cdot,z))\right\|<d^{-(2+\alpha)+\hat\delta}\quad\text{w.o.p.}$$

In addition, for the Hessian and higher order terms errors, by Propositions D.11 and D.12, the proofs of which are deferred to Sections D.3.3 and D.3.4, we have

$$\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\sum_{j=0}^{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor-1}\left\|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_j^{\mathrm{Hess}}(S)\middle|\mathcal{F}_j\right]\right\|\leqslant Cd^{-1}\quad\text{w.o.p.},$$

$$\text{and}\quad\sup_{z\in\Gamma^\delta}\sup_{0\leqslant t\leqslant T}\sum_{j=0}^{\lfloor(t\wedge\vartheta_{C\cdot M})d\rfloor-1}\left\|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_j^{\mathrm{High}}(S)\middle|\mathcal{F}_j\right]\right\|\leqslant Cd^{-\frac{1}{2}}\quad\text{w.o.p.}$$

The mesh error is the contribution of the last fractional time interval:

$$\xi_t^{\mathrm{mesh}}(S)(z)=-\int_{t_d}^t\mathcal{F}\big(z,S(x_{\lfloor sd\rfloor},\cdot)\big)\,\mathrm{d}s,\qquad t_d=\frac{\lfloor td\rfloor}{d}.$$

Hence, since $0\leqslant t-t_d\leqslant 1/d$,

$$\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|\xi_t^{\mathrm{mesh}}(S)(\cdot)\right\|_\Gamma\leqslant\frac{1}{d}\sup_{0\leqslant s\leqslant(\hat\tau_M\wedge T)}\left\|\mathcal{F}\big(\cdot,S(x_{\lfloor sd\rfloor},\cdot)\big)\right\|_\Gamma.$$

Last we show that in the stopped interval, the right-hand side is bounded by Cd^{-1}. For any $0\leqslant s\leqslant(\hat\tau_M\wedge T)$, we have

$$\begin{aligned}\left\|\mathcal{F}(z,S(x_{\lfloor sd\rfloor},z))\right\|&\leqslant\bar\gamma C\left\|H\left(B(x_{\lfloor sd\rfloor})\right)\right\|\cdot\left\|z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)(S(x_{\lfloor sd\rfloor},z))\right\|\\&\quad+\bar\gamma^2C\left\|I\left(B(x_{\lfloor sd\rfloor})\right)\right\|\cdot\left\|z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)(S(x_{\lfloor sd\rfloor},z))\right\|+\cdots\end{aligned}$$

such that $B(x_{\lfloor sd\rfloor})=\frac{1}{(2\pi)^4}\oint_\Gamma z_4\,S(x_{\lfloor sd\rfloor},z)\,\mathrm{d}z$. Next, plugging equations (34), (39), (41), and

$$\left\|z^\zeta\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)(S(x_{\lfloor sd\rfloor},z))\right\|\leqslant C(|\Gamma|)\cdot\left\|\left(\bigcirc_{i=1}^4\mathcal{G}_i\right)(S(x_{\lfloor sd\rfloor},z))\right\|,$$

we know there is a positive constant $C=C(L(h),L(I),\bar\gamma,\|K\|_{\mathrm{op}},\|\beta^*\|_\infty,M,\alpha)$, such that $\left\|\mathcal{F}\big(\cdot,S(x_{\lfloor sd\rfloor},\cdot)\big)\right\|_\Gamma\leqslant C$. Therefore,

$$\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|\xi_t^{\mathrm{mesh}}(S)(\cdot)\right\|_\Gamma\leqslant Cd^{-1}.\tag{88}$$

Consequently, combining all the errors, we deduce that for some $C>0$, which does not depend on d,

$$\sup_{0\leqslant t\leqslant(\hat\tau_M\wedge T)}\left\|S\left(x_{\lfloor td\rfloor},z\right)-S\left(x_0,z\right)-\int_0^t\mathcal{F}\left(z,S(x_{\lfloor sd\rfloor},z)\right)\mathrm{d}s\right\|_{\Gamma^\delta}\leqslant Cd^{\hat\delta/2-1/2}\quad\text{w.o.p.}$$

An application of the net argument, Lemma D.1, finishes the proof after setting $\hat\delta=1-2\delta$ for $\delta\in(0,1/2)$.

The derivative regularity condition (i) is automatic for the continuous-time embedding $t\mapsto S(x_{\lfloor td\rfloor},\cdot)$: for each fixed trajectory, the dependence on z is through a finite product of resolvents on the fixed contour Γ, and the resolvent derivative bounds from Remark C.3 give the required Lipschitz control up to $\hat\tau_M$. Thus the estimate above verifies condition (ii), and hence $S(x_{\lfloor td\rfloor},\cdot)$ is an $(d^{-\varepsilon},M,T)$-approximate solution. □

D.3 Error Bounds

Recall that we are interested in the statistic

$$S(x,z):=\frac{1}{d}W(x)^\top\Omega(x,z)W(x)\in\mathbb{C}^{3\times3},$$

where

$$W(x):=[\,\psi(x)\mid\beta^*\mid\mathbb{1}_d\,]\in\mathbb{R}^{d\times3}$$

and $\Omega(x,z)$ is defined in (17). Throughout this section, the contour Γ is as in Remark 3.3. We control the error terms arising in the comparison between SGD and homogenized SGD under S, with $\mathcal{F}$ given by the partial integro-differential equation (30). All estimates are taken on the stopped event where $\|x_{\lfloor td\rfloor}\|_\infty$ and $\|\mathfrak{X}_t\|_\infty$ remain uniformly bounded for $0\leqslant t\leqslant T$ (see Assumption 3.1). Constants may depend on this bound, but not on d.

Before proceeding, we present some bounds on the derivatives of S.

Lemma D.6. *There exist constants $c, C > 0$ such that, for large enough d, the following hold:*

$$\frac{c}{d}\|W\|^2 \leqslant \|S(x,\cdot)\|_\Gamma \leqslant \frac{C}{d}\|W\|^2,$$

$$\|\nabla_x S(x,\cdot)\|_\Gamma \leqslant \frac{C(\|x\|_\infty)}{d}\|W\|,$$

and $$\left\|\nabla_x^2 S(x,\cdot)\right\|_\Gamma \leqslant \frac{C(\|x\|_\infty)}{d}.$$

Proof. The first result follows from the proof of Lemma C.1.

For the derivative, observe that

$$\|\mathrm{diag}(\psi(x))\|_{\mathrm{op}} = \|\psi(x)\|_\infty = \max_i |\psi_i(x)| \leqslant 2\|Q\|\|x\|_\infty^2 + \|l\|_1\|x\|_\infty + |c|. \tag{89}$$

Moreover, by Lemma B.1, we have

$$\begin{aligned}\|\nabla_u \psi(x)\|_{\mathrm{op}} &= \|2q_{11}\,\mathrm{diag}(u) + 2q_{12}\,\mathrm{diag}(v) + l_1 I_d\|_{\mathrm{op}} \leqslant 2\sqrt{2}\|x\|_\infty + |l_1|,\\ \|\nabla_v \psi(x)\|_{\mathrm{op}} &= \|2q_{12}\,\mathrm{diag}(u) + 2q_{22}\,\mathrm{diag}(v) + l_2 I_d\|_{\mathrm{op}} \leqslant 2\sqrt{2}\|x\|_\infty + |l_2|.\end{aligned} \tag{90}$$

Taking norms on Lemma B.3 and using that $\sup_{z_i\in\Gamma_i}\|R(z_i;D_i)\|_{\mathrm{op}} \leqslant 2$, the second result follows.

Analogously, for the Hessian, the bound immediately follows. □

Consequently, the same bounds hold for the real-valued coordinate functions $S_{ab}^{\mathrm{Re},z}$ and $S_{ab}^{\mathrm{Im},z}$, uniformly in $a, b \in \{1,2,3\}$ and $z \in \Gamma$.

To control the errors, we will need to make an *a priori* estimate that effectively shows that the iterates of homogenized SGD and SGD remain bounded. Thus, recall our definition, for fixed $M > 0$, of the stopping times

$$\begin{aligned}\vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d}\left\|W_{\lfloor td\rfloor}\right\|^2 > M \text{ or } B\left(x_{\lfloor td\rfloor}\right) \notin \mathcal{U}\} \quad \text{or}\\ \vartheta_M &= \inf\{t \geqslant 0 : \frac{1}{d}\|\mathcal{W}_t\|^2 > M \text{ or } B\left(\mathcal{X}_t\right) \notin \mathcal{U}\},\end{aligned} \tag{91}$$

depending on whether we are working with SGD iterates or homogenized SGD iterates. We often drop the M so that $\vartheta := \vartheta_M$. It will be convenient to work with the stopped processes, $W^\vartheta_{\lfloor td\rfloor} := W_{\lfloor (t\wedge\vartheta)d\rfloor}$ and $\mathcal{W}_t^\vartheta := \mathcal{W}_{t\wedge\vartheta}$.

Remark D.7. The stopping time $\hat{\tau}_M = \inf\{t \geqslant 0 : \left\|S(x_{\lfloor td\rfloor}, z)\right\|_\Gamma > M \text{ or } B(x_{\lfloor td\rfloor}) \notin \mathcal{U}\}$ and $\hat{\tau}_M = \inf\{t \geqslant 0 : \|S(\mathcal{X}_t, z)\|_\Gamma > M \text{ or } B(\mathcal{X}_t) \notin \mathcal{U}\}$ are related to ϑ_M by positive constants $c, C > 0$, $\vartheta_{c\cdot M} \leqslant \hat{\tau}_M \leqslant \vartheta_{C\cdot M}$ (see Lemma D.6).

In the remainder of this section, we establish a series of propositions that provide bounds on the martingale terms arising from both homogenized SGD and SGD. Throughout the proofs below, C denotes a positive constant independent of d. It may depend on $L(h)$, $L(I)$, $\bar{\gamma}$, T, $\|K\|_{\mathrm{op}}$, $\|\beta^*\|_\infty$, M, and α, and may change from line to line and not necessarily be the same as the constant C in Lemma D.6.

D.3.1 Homogenized SGD Martingale Error

In this section we control the martingale that arises in homogenized SGD, that is, for a test function $\varphi\colon \mathbb{R}^{2d} \to \mathbb{R}$, define

$$\mathcal{M}_t^{\mathrm{HSGD}}(\varphi) := \int_0^t \mathrm{d}\mathcal{M}_s^{\mathrm{HSGD}}(\varphi) = \int_0^t \gamma(s)\sqrt{I\left(B(\mathcal{X}_s)\right)}\langle(\nabla\psi(\mathcal{X}_s))^\top\sqrt{K}, \nabla\varphi(\mathcal{X}_s)\left(\mathrm{d}\mathfrak{B}_s\right)^\top\rangle. \tag{92}$$

As introduced in Remark D.3, we are interested in controlling $\mathcal{M}_t^{\mathrm{HSGD}}(S(\cdot,z))$.

To control the fluctuations of this martingale, we need to control its quadratic variation, defined as follows. Consider a partition of time for $[0,t]$, that is, $0=t_0<t_1<\cdots<t_n=t$ such that the size of the partition $\Delta t=\max_i\{t_i-t_{i-1}\}\to 0$. We define for the continuous process Y,

$$[Y_t(n)]=\sum_{k=1}^{n}\left(Y_{t_k}-Y_{t_{k-1}}\right)^2.$$

If, for every partition of time $[0,t]$ such that $\Delta t\to 0$, the process $[Y_t(n)]$ converges in probability to a process $[Y_t]$ as $n\to\infty$, we call $[Y_t]$ the *quadratic variation* of Y. Using the quadratic variation of $\mathcal{M}_t^{\mathrm{HSGD}}$, we will show that the martingale arising from homogenized SGD is small.

Proposition D.8 (Homogenized SGD martingale small)**.** *Suppose $f\colon \mathbb{R}^2\to\mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 1.4). Let the statistic $S\colon \mathbb{R}^{2d}\times\Gamma\subset\mathbb{R}^{2d}\times\mathbb{C}^4\to\mathbb{C}^{3\times 3}$ be defined as in (18). Then for each fixed $z\in\Gamma$ and each $T,\zeta>0$, there is some constant C such that, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(S(\cdot,z))\right\|\leqslant CL(f)d^{\zeta/2-1/2}. \tag{93}$$

Proof. Fix indices $i,j\in\{1,2,3\}$. Let φ denote either

$$\varphi(x)=S_{ij}^{\mathrm{Re},z}(x)=\operatorname{Re}S_{ij}(x,z)$$

or

$$\varphi(x)=S_{ij}^{\mathrm{Im},z}(x)=\operatorname{Im}S_{ij}(x,z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot,z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. First, we rewrite the martingale increment, $\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}$, as

$$\mathrm{d}\mathcal{M}_t^{\mathrm{HSGD}}(\varphi)=\gamma(t)\sqrt{I\left(B(\mathcal{X}_t)\right)}\langle\left(\nabla\psi(\mathcal{X}_t)\right)^\top\sqrt{K},\nabla_x\varphi(\mathcal{X}_t,z)\left(\mathrm{d}\mathfrak{B}_t\right)^\top\rangle_{\mathbb{R}^{2d\times d}}. \tag{94}$$

The quadratic variation of $\mathcal{M}_t^{\mathrm{HSGD}}$ is

$$\left[\mathcal{M}_t^{\mathrm{HSGD}}(\varphi)\right]=\int_0^t\gamma(s)^2\left|I\left(B(\mathcal{X}_s)\right)\right|\left\|\langle\nabla_x\varphi(\mathcal{X}_s,z),\nabla\psi(\mathcal{X}_s)^\top\sqrt{K}\rangle_{\mathbb{R}^{2d}}\right\|^2\mathrm{d}s. \tag{95}$$

We need to compute $\sup_{0\leqslant t\leqslant T}\left[\mathcal{M}_t^{\mathrm{HSGD}}(\varphi)\right]$ and show that this quantity is small. In particular, we only need to show that the integrand is small. For this, we see that

$$\left\|\langle\nabla_x\varphi(\mathcal{X}_s,z),\nabla\psi(\mathcal{X}_s)^\top\sqrt{K}\rangle_{\mathbb{R}^{2d}}\right\|^2\leqslant\|K\|_{\mathrm{op}}\|\nabla\psi(\mathcal{X}_s)\|_{\mathrm{op}}^2\|\nabla_x\varphi(\mathcal{X}_s,z)\|^2.$$

Using (90), we have that $\|\nabla\psi(\mathcal{X}_s)\|_{\mathrm{op}}\leqslant C(\|\mathcal{X}_s\|_\infty)$. By Lemma D.6, we have a bound on $\|\nabla_x\varphi(\mathcal{X}_s,z)\|\leqslant\|\nabla_x S(\mathcal{X}_s,\cdot)\|_\Gamma\leqslant\frac{C(\|\mathcal{X}_s\|_\infty)}{d}\|\mathcal{W}_s\|$. From Lemma B.6, the growth condition on $|I\left(B(\mathcal{X}_s)\right)|=\mathbb{E}_a[|\nabla_{r_1}f(\rho_s)|^2]$ yields

$$\begin{aligned}
&\left|I\left(B(\mathcal{X}_s)\right)\right|\left\|\langle\nabla_x\varphi(\mathcal{X}_s,z),\nabla\psi(\mathcal{X}_s)^\top\sqrt{K}\rangle_{\mathbb{R}^{2d}}\right\|^2\\
&\quad\leqslant\frac{C(\|\mathcal{X}_s\|_\infty)}{d^2}(L(f))^2\|\mathcal{W}_s\|^2\left(1+\frac{1}{\sqrt{d}}\|K\|_{\mathrm{op}}^{1/2}\|\mathcal{W}_s\|\right)^{\max\{1,2\alpha\}}\\
&\quad\leqslant\frac{C(\|\mathcal{X}_s\|_\infty)}{d}(L(f))^2M\left(1+\sqrt{M}\right)^{\max\{1,2\alpha\}}.
\end{aligned}\tag{96}$$

Thus, (95) and (96), together

$$\sup_{0\leqslant t\leqslant T}\left[\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right]\leqslant C(\|\mathcal{X}_s\|_\infty)(L(f))^2\cdot\bar{\gamma}^2\cdot d^{-1}. \tag{97}$$

Using the fact that if $\sup_{0\leqslant t\leqslant T}\left[\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right]\leqslant b$ a.s. then $\mathbb{P}\left[\sup_{0\leqslant t\leqslant T}\left|\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right|>p\right]\leqslant\exp(-p^2/2b)$, for any $\zeta>0$ and $p=\sqrt{C}L(f)d^{\zeta/2-1/2}$,

$$\mathbb{P}\left[\sup_{0\leqslant t\leqslant T}\left|\mathcal{M}_{t\wedge\vartheta}^{\mathrm{HSGD}}(\varphi)\right|>p\right]\leqslant C\exp(-d^\zeta).$$

Applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot,z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

D.3.2 Bounds on the Martingales $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Hess}}$

Now we move on to the martingale increments coming from SGD applied to test functions φ. Recall, the expressions for the martingale increments for any quartic statistics φ

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Grad}}(\varphi) &= \gamma_k \left\langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\right\rangle - \frac{\gamma_k}{\sqrt{d}}\left\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \\ \Delta\mathcal{M}_k^{\mathrm{Hess}}(\varphi) &= \frac{\gamma_k^2}{2d}\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right)^{\otimes 2}\rangle \\ &\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle \nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right)^{\otimes 2}\rangle\middle|\mathcal{F}_k\right]\end{aligned}$$

with

$$\mathcal{M}_k(\varphi) = \sum_{j=0}^{k-1}\Delta\mathcal{M}_j(\varphi).$$

Proposition D.9 (Gradient martingale)**.** *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 1.4). Let the statistic $S\colon \mathbb{R}^{2d}\times\Gamma \subset \mathbb{R}^{2d}\times\mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ be defined as in (18). Then for each fixed $z\in\Gamma$ and each $T,\zeta>0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T}\left\|\mathcal{M}^{\mathrm{Grad}}_{\lfloor d(t\wedge\vartheta)\rfloor}(S(\cdot,z))\right\| \leqslant d^{-\frac{2+\alpha}{2}+\zeta}. \tag{98}$$

Proof. Fix indices $i,j\in\{1,2,3\}$. Let φ denote either

$$\varphi(x) = S_{ij}^{\mathrm{Re},z}(x) = \operatorname{Re} S_{ij}(x,z)$$

or

$$\varphi(x) = S_{ij}^{\mathrm{Im},z}(x) = \operatorname{Im} S_{ij}(x,z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot,z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. Throughout the proof of this proposition, we will be working on the stopped version of the martingale, $\mathcal{M}^{\mathrm{Grad}}_{\lfloor d(t\wedge\vartheta)\rfloor}$. However, to lighten the notation, we will suppress the ϑ dependence in the subscript as well as the φ and simply write $\mathcal{M}^{\mathrm{Grad}}_{\lfloor td\rfloor} := \mathcal{M}^{\mathrm{Grad}}_{\lfloor d(t\wedge\vartheta)\rfloor}(\varphi)$. We have the martingale increments

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Grad}} &= \gamma_k \left\langle \nabla\varphi(x_k), \nabla\mathcal{R}(x_k)\right\rangle - \frac{\gamma_k}{\sqrt{d}}\left\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \\ &= \frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\left[\left\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle\middle|\mathcal{F}_k\right] \\ &\quad - \frac{\gamma_k}{\sqrt{d}}\left\langle \nabla\varphi(x_k), \nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \\ &= \frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\left[\left\langle \nabla\varphi(x_k), \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \nabla_{r_1} f\left(r_k\right)\middle|\mathcal{F}_k\right] \\ &\quad - \frac{\gamma_k}{\sqrt{d}}\left\langle \nabla\varphi(x_k), \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \nabla_{r_1} f\left(r_k\right).\end{aligned}$$

We define $\mathcal{M}_k^{\mathrm{Grad},\beta}$ to be a new martingale with increments

$$\begin{aligned}\Delta\mathcal{M}_k^{\mathrm{Grad},\beta} &= \frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\left[\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left\langle \nabla\varphi(x_k), \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \nabla_{r_1} f\circ \mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left(r_k\right)\middle|\mathcal{F}_k\right] \\ &\quad - \frac{\gamma_k}{\sqrt{d}}\,\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left\langle \nabla\varphi(x_k), \left(\nabla\psi(x_k)\right)^\top \cdot a_{k+1}\right\rangle \nabla_{r_1} f\circ \mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left(r_k\right),\end{aligned}$$

where we note that there are two projections and the projection of r_k is in both coordinates of $\mathbb{R}^2$, even though the gradient $\nabla_{r_1} f$ is only with respect to the first coordinate. We take $\beta = d^{\zeta}$ in the projection radius for some $\zeta>0$ to be determined later. We will bound $\mathcal{M}_k^{\mathrm{Grad},\beta}$ first, and then bound the difference between $\mathcal{M}_k^{\mathrm{Grad}}$ and $\mathcal{M}_k^{\mathrm{Grad},\beta}$.

We begin by computing subgaussian bounds on the quantities that are going to be projected, namely r_k and $\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle$. For the purposes of this section, when we refer to a vector as *subgaussian*, we mean that its entries individually satisfy the stated subgaussian concentration bound. We can rewrite the quantities r_k and $\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle$ as

$$\begin{aligned} r_k = \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix}^\top a_{k+1} = \frac{1}{\sqrt{d}} \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix}^\top \sqrt{K} v_k \quad \text{and} \\ \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle = \langle \sqrt{K} \cdot \nabla\psi(x_k) \cdot \nabla\varphi(x_k), v_k\rangle, \end{aligned} \tag{99}$$

so r_k is $\frac{1}{\sqrt{d}} \left\|\sqrt{K}\right\|_{\text{op}} \cdot \left\| \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix} \right\|_{\text{op}}$ - subgaussian and $\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle$ is $\left\|\sqrt{K}\right\|_{\text{op}} \cdot \|\nabla\psi(x_k)\|_{\text{op}} \cdot \|\nabla\varphi(x_k)\|_{\text{op}}$ - subgaussian where

$$\|\nabla\varphi(x_k)\|_{\text{op}} = \|\nabla_x S_{ij}(x_k, z)\|_{\text{op}} \leqslant \|\nabla_x S(x_k, \cdot)\|_\Gamma \leqslant \frac{C(\|x_k\|_\infty)}{d} \|W_k\|.$$

by Lemma D.6, and $\left\| \begin{pmatrix} \psi(x_k) \\ \beta^* \end{pmatrix} \right\|_{\text{op}}, \|\nabla\psi(x_k)\|_{\text{op}} \leqslant C(\|x_k\|_\infty)$ by (89) and (90). Thus, since we are working on the stopped processes,

$$\|r_k\|_{\psi_2} = d^{-\frac{1}{2}} C(\|x_k\|_\infty) \quad \text{and} \quad \left\|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right\|_{\psi_2} = d^{-\frac{1}{2}} C(\|x_k\|_\infty). \tag{100}$$

These subgaussian bounds will be used to bound the difference between $\mathcal{M}_k^{\text{Grad}}$ and $\mathcal{M}_k^{\text{Grad},\beta}$.

Furthermore, leveraging the projections and the fact that f is α-pseudo-Lipschitz, which implies the growth bound on $\nabla_{r_1} f(r)$ in Lemma B.6, we can derive the norm bounds

$$\left|\nabla_{r_1} f \circ \text{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right| \leqslant L(f) C (1 + d^{-\frac{1}{2}}\beta)^{\max\{1,\alpha\}}, \tag{101}$$

$$\left|\text{Proj}_{d^{-\frac{1}{2}}\beta} \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| \leqslant d^{-\frac{1}{2}}\beta. \tag{102}$$

This gives us the bound

$$\left|\text{Proj}_{d^{-\frac{1}{2}}\beta} \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \nabla_{r_1} f \circ \text{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right| \leqslant L(f) C d^{-\frac{2+\alpha}{2}} \beta^{2+\alpha} \tag{103}$$

and, since this is an almost sure bound, it holds for the expectation as well, and we get

$$\left|\Delta\mathcal{M}_k^{\text{Grad},\beta}\right| \leqslant 2\gamma_k L(f) C d^{-\frac{3+\alpha}{2}} \beta^{2+\alpha}.$$

Applying Azuma's inequality with the assumption $n = O(d)$, we obtain

$$\sup_{1\leqslant k\leqslant n} \mathbb{P}\left[\left|\mathcal{M}_k^{\text{Grad},\beta}\right| > t\right] < 2\exp\left(\frac{-t^2}{2n \cdot (C d^{-\frac{3+\alpha}{2}} \beta^{2+\alpha})^2}\right) \leqslant 2\exp\left(\frac{-t^2}{C' d^{-(2+\alpha)} \beta^{2(2+\alpha)}}\right).$$

Thus, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n} \left|\mathcal{M}_k^{\text{Grad},\beta}\right| < d^{-\frac{2+\alpha}{2}} \beta^{3+\alpha}. \tag{104}$$

Finally, we bound the difference between $\{\mathcal{M}_k^{\text{Grad}}\}_{k=1}^n$ and $\{\mathcal{M}_k^{\text{Grad},\beta}\}_{k=1}^n$. For ease of notation, we write

$$\begin{aligned} G_k &:= \frac{\gamma_k}{\sqrt{d}} \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \nabla_{r_1} f(r_k) \\ G_{k,\beta} &:= \frac{\gamma_k}{\sqrt{d}} \text{Proj}_{d^{-\frac{1}{2}}\beta} \langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \nabla_{r_1} f \circ \text{Proj}_{d^{-\frac{1}{2}}\beta}(r_k). \end{aligned}$$

The quantity we are trying to bound is

$$\left|\Delta\mathcal{M}_k^{\text{Grad}} - \Delta\mathcal{M}_k^{\text{Grad},\beta}\right| = \left|\left(G_k - \mathbb{E}_{a_{k+1}} G_k\right) - \left(G_{k,\beta} - \mathbb{E}_{a_{k+1}} G_{k,\beta}\right)\right|$$

$$\leqslant |G_k - G_{k,\beta}| + \left|\mathbb{E}_{a_{k+1}}\left(G_k - G_{k,\beta}\right)\right|.$$

First, we will show that $G_k - G_{k,\beta} = 0$ with overwhelming probability. Using the subgaussian bounds on r_k and $\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle$, we have

$$\begin{aligned}\mathbb{P}\left[G_k \neq G_{k,\beta}\right] &\leqslant \mathbb{P}\left[\|r_k\| > d^{-\frac{1}{2}}\beta\right] + \mathbb{P}\left[\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right] \\ &< 4\exp\left(-\frac{\beta^2}{2C}\right).\end{aligned}$$

Since $\beta = d^\zeta$ for some $\zeta > 0$, the probability bounds above imply that $G_k - G_{k,\beta} = 0$ with overwhelming probability, and it remains to bound the difference in their expectations. For this, we have

$$\begin{aligned}\left|\mathbb{E}_{a_{k+1}}\left[G_k - G_{k,\beta}\right]\right| &= \left|\mathbb{E}_{a_{k+1}}\left[\left(G_k - G_{k,\beta}\right)\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| \\ &\leqslant \left|\mathbb{E}_{a_{k+1}}\left[G_k \cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| + \left|\mathbb{E}_{a_{k+1}}\left[G_{k,\beta}\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right|.\end{aligned}$$

For $\left|\mathbb{E}_{a_{k+1}}\left[G_{k,\beta}\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right|$, we have

$$\begin{aligned}\left|\mathbb{E}_{a_{k+1}}\left[G_{k,\beta}\cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| &\leqslant \max|G_{k,\beta}|\cdot \mathbb{P}\left[G_k \neq G_{k,\beta}\right] \\ &\leqslant L(f)Cd^{-\frac{2+\alpha}{2}}\beta^{3+\alpha}\cdot 4\exp\left(-\frac{\beta^2}{2C}\right).\end{aligned}$$

For $\left|\mathbb{E}_{a_{k+1}}\left[G_k \cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right|$, we have

$$\begin{aligned}\left|\mathbb{E}_{a_{k+1}}\left[G_k \cdot \mathbb{1}\{G_k \neq G_{k,\beta}\}\right]\right| &\leqslant \mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_1\}| \\ &\quad + \mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_2\}| \\ &\quad + \mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_3\}|,\end{aligned}$$

$$\begin{aligned}\text{where } E_1 &:= \{\|r_k\| \leqslant d^{-\frac{1}{2}}\beta\} \cap \left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}, \\ E_2 &:= \{\|r_k\| > d^{-\frac{1}{2}}\beta\} \cap \left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| \leqslant d^{-\frac{1}{2}}\beta\right\}, \\ E_3 &:= \{\|r_k\| > d^{-\frac{1}{2}}\beta\} \cap \left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}.\end{aligned}$$

The term $\mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_1\}|$ can be bounded as

$$\mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_1\}| \leqslant L(f)C(d^{-\frac{1}{2}}\beta)^{\max\{1,\alpha\}}.$$

$$\mathbb{E}_{a_{k+1}}\left[\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| \cdot \mathbb{1}\left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}\right],$$

where the expectation on the right-hand side is exponentially small due to being a tail of a subgaussian first moment (where β^2 is larger than the sub-gaussian variance and grows with d). By similar reasoning, $\mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_2\}|$ is also exponentially small (using the growth bound on $\nabla_{r_1} f$). For $\mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_3\}|$, we have

$$\begin{aligned}\mathbb{E}_{a_{k+1}}|G_k\cdot \mathbb{1}\{E_3\}| &\leqslant \frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\Big[\left|\nabla_{r_1} f\left(r_k\right)\cdot \mathbb{1}\{\|r_k\| > d^{-\frac{1}{2}}\beta\}\right| \cdot \\ &\qquad \left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \cdot \mathbb{1}\left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}\right|\Big] \\ &\leqslant \frac{\gamma_k}{\sqrt{d}}\,\mathbb{E}_{a_{k+1}}\left|\nabla_{r_1} f\left(r_k\right)\cdot \mathbb{1}\{\|r_k\| > d^{-\frac{1}{2}}\beta\}\right|^2 \cdot \\ &\qquad \mathbb{E}_{a_{k+1}}\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle \cdot \mathbb{1}\left\{\left|\langle \nabla\varphi(x_k), (\nabla\psi(x_k))^\top \cdot a_{k+1}\rangle\right| > d^{-\frac{1}{2}}\beta\right\}\right|^2.\end{aligned}$$

This is a product of tails of Gaussian moments, which is again exponentially small. Thus, we conclude that, with overwhelming probability, $\sup_{1\leqslant k\leqslant n}\left|\Delta\mathcal{M}_k^{\mathrm{Grad}} - \Delta\mathcal{M}_k^{\mathrm{Grad},\beta}\right|$ is exponentially small and thus, taking $\beta = d^\zeta$, we conclude that, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n}\left|\mathcal{M}_k^{\mathrm{Grad}}\right| < d^{-\frac{2+\alpha}{2}+\zeta(3+\alpha)}.$$

Adjusting the value of ζ, recalling that all of this has been proved on the stopped process, and applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot, z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

Proposition D.10 (Hessian martingale)**.** *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 1.4). Let the statistic $S\colon \mathbb{R}^{2d} \times \Gamma \subset \mathbb{R}^{2d} \times \mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ be defined as in (18). Then for each fixed $z \in \Gamma$ and each $T, \zeta > 0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T} \left|\mathcal{M}^{\mathrm{Hess}}_{\lfloor d(t\wedge\vartheta)\rfloor}(S(\cdot, z))\right| \leqslant d^{-(2+\alpha)+\zeta}. \tag{105}$$

Proof. As in the proof of the previous proposition, we will work on the stopped version of the martingale but will suppress the ϑ dependence in the subscript in order to lighten the notation. We also, as before, let φ denote either

$$\varphi(x) = S^{\mathrm{Re},z}_{ij}(x) = \operatorname{Re} S_{ij}(x, z)$$

or

$$\varphi(x) = S^{\mathrm{Im},z}_{ij}(x) = \operatorname{Im} S_{ij}(x, z).$$

We have the martingale increment

$$\begin{aligned}
\Delta\mathcal{M}^{\mathrm{Hess}}_k &= \frac{\gamma_k^2}{2d}\langle\nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle \\
&\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle\nabla^2\varphi(x_k), \left(\nabla_{r_1} f\left(r_k\right)\cdot\left(\nabla\psi(x_k)\right)^\top\cdot a_{k+1}\right)^{\otimes 2}\rangle\middle|\mathcal{F}_k\right] \\
&= \frac{\gamma_k^2}{2d}\,\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\left(r_k\right)\right)^2 \\
&\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\left(r_k\right)\right)^2\middle|\mathcal{F}_k\right].
\end{aligned}$$

We define $\mathcal{M}^{\mathrm{Hess},\beta}_k$ to be a new martingale with increments

$$\begin{aligned}
\Delta\mathcal{M}^{\mathrm{Hess},\beta}_k &= \frac{\gamma_k^2}{2d}\,\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left(r_k\right)\right)^2 \\
&\quad - \frac{\gamma_k^2}{2d}\,\mathbb{E}_{a_{k+1}}\left[\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1} f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\left(r_k\right)\right)^2\middle|\mathcal{F}_k\right].
\end{aligned}$$

The approach here is similar to the procedure for bounding $\mathcal{M}^{\mathrm{Grad}}_k$. As we saw in the proof of the previous Proposition, r_k is $\frac{1}{\sqrt{d}}\left\|\sqrt{K}\right\|_{\mathrm{op}}\cdot\left\|\begin{pmatrix}\psi(x_k)\\ \beta^*\end{pmatrix}\right\|_{\mathrm{op}}$–subgaussian in each entry. To obtain a concentration bound for $\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle$, we rewrite it as

$$\begin{aligned}
\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle &= \langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\right]^{\otimes 2}\rangle, \left(\sqrt{K}v_{k+1}\right)^{\otimes 2}\rangle \\
&= \langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle, v_{k+1}^{\otimes 2}\rangle,
\end{aligned}$$

where the vector v_k has independent standard gaussian entries. Since $\nabla^2\varphi(x_k) \in \mathbb{R}^{2d\times 2d}$ and $\left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2} \in (\mathbb{R}^{2d\times d})^{\otimes 2}$, we get $\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle \in \mathbb{R}^{d\times d}$.

Moreover, by Lemma D.6 and (90), we have

$$\left\|\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|_{\mathrm{op}} \leqslant \|K\|_{\mathrm{op}}\cdot\|\nabla\psi(x_k)\|^2_{\mathrm{op}}\cdot\left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}} \leqslant \frac{C(\|x_k\|_\infty)}{d},$$

so by Hanson-Wright inequality we obtain that, for large enough t,

$$\mathbb{P}\left[\langle\langle\nabla^2\varphi(x_k), \left[\left(\nabla\psi(x_k)\right)^\top\sqrt{K}\right]^{\otimes 2}\rangle, v_{k+1}^{\otimes 2}\rangle > t\right]$$

$$< 2\exp\left(-C\min\left\{\frac{t^2}{\left\|\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|^2},\frac{t}{\left\|\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|_{\mathrm{op}}}\right\}\right).$$

Using the fact that

$$\left\|\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|^2 \leqslant d\left\|\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle\right\|_{\mathrm{op}}^2 \leqslant \frac{C(\|x_k\|_\infty)}{d},$$

we conclude that

$$\mathbb{P}\left[\left|\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\sqrt{K}\right]^{\otimes 2}\rangle,v_{k+1}^{\otimes 2}\rangle\right|>t\right] < 2\exp\left(-\frac{d\cdot\min\{t^2,t\}}{C}\right).$$

In particular, this tells us that, for any $\zeta>0$,

$$\left|\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle,a_{k+1}^{\otimes 2}\rangle\right| \leqslant d^{-\frac{1}{2}+\zeta}$$

with overwhelming probability.

Having obtained concentration bounds for r_k and $\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle,a_{k+1}^{\otimes 2}\rangle$, we proceed to bound $\mathcal{M}_k^{\mathrm{Hess},\beta}$ and show that it is close to $\mathcal{M}_k^{\mathrm{Hess}}$. From the projections and the growth bound on $\nabla_{r_1}f$ in Lemma B.6, we can derive the norm bounds

$$\left|\nabla_{r_1}f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right|^2 \leqslant \left(L(f)C(1+d^{-\frac{1}{2}}\beta)^{\max\{1,\alpha\}}\right)^2 \tag{106}$$

$$\left|\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle,a_{k+1}^{\otimes 2}\rangle\right| \leqslant d^{-\frac{1}{2}}\beta, \tag{107}$$

and thus

$$\left|\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle,a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1}f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right)^2\right| \leqslant (L(f)C)^2 d^{-\frac{3+2\alpha}{2}}\beta^{3+2\alpha}.$$

Since this is an almost sure bound, it holds for the expectation as well and we get

$$\left|\Delta\mathcal{M}_k^{\mathrm{Hess},\beta}\right| \leqslant \gamma_k^2 (L(f)C)^2 d^{-\frac{5+2\alpha}{2}}\beta^{3+2\alpha}.$$

Applying Azuma's inequality with $n=O(d)$, we obtain

$$\sup_{1\leqslant k\leqslant n}\mathbb{P}\left[\left|\mathcal{M}_k^{\mathrm{Hess},\beta}\right|>t\right] < 2\exp\left(\frac{-t^2}{2n\cdot\left(Cd^{-\frac{5+2\alpha}{2}}\beta^{3+2\alpha}\right)^2}\right) \leqslant 2\exp\left(\frac{-t^2}{C'd^{-2(2+\alpha)}\beta^{2(3+2\alpha)}}\right)$$

so, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n}\left|\mathcal{M}_k^{\mathrm{Hess},\beta}\right| < d^{-(2+\alpha)}\beta^{4+2\alpha}. \tag{108}$$

It remains only to bound the difference between $\{\mathcal{M}_k^{\mathrm{Hess}}\}_{k=1}^n$ and $\{\mathcal{M}_k^{\mathrm{Hess},\beta}\}_{k=1}^n$. This follows a very similar argument to what was in the proof of Proposition D.9, we write

$$G_k := \frac{\gamma_k^2}{2d}\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle,a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1}f(r_k)\right)^2$$
$$G_{k,\beta} := \frac{\gamma_k^2}{2d}\,\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}\langle\langle\nabla^2\varphi(x_k),\left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle,a_{k+1}^{\otimes 2}\rangle\left(\nabla_{r_1}f\circ\mathrm{Proj}_{d^{-\frac{1}{2}}\beta}(r_k)\right)^2.$$

The quantity we are trying to bound is

$$\left|\Delta\mathcal{M}_k^{\mathrm{Hess}}-\Delta\mathcal{M}_k^{\mathrm{Hess},\beta}\right| = \left|\left(G_k-\mathbb{E}_{a_{k+1}}G_k\right)-\left(G_{k,\beta}-\mathbb{E}_{a_{k+1}}G_{k,\beta}\right)\right|$$

$$\leqslant |G_k - G_{k,\beta}| + \left|\mathbb{E}_{a_{k+1}}\left(G_k - G_{k,\beta}\right)\right|.$$

As in the proof of Proposition D.9, the first of the terms on the right-hand side is 0 with overwhelming probability, while the second is exponentially small. Computing the bound for $\left|\mathbb{E}_{a_{k+1}}\left(G_k - G_{k,\beta}\right)\right|$ is similar to what was done in the previous proof and is not repeated here. To see that $|G_k - G_{k,\beta}| = 0$ with overwhelming probability, we write

$$\mathbb{P}\left[G_k \neq G_{k,\beta}\right] \leqslant \mathbb{P}[\|r_k\| > d^{-\frac{1}{2}}\beta] + \mathbb{P}[\left|\langle\langle \nabla^2\varphi(x_k), \left[(\nabla\psi(x_k))^\top\right]^{\otimes 2}\rangle, a_{k+1}^{\otimes 2}\rangle\right| > d^{-\frac{1}{2}}\beta]$$

$$< 2\exp\left(-\frac{\beta^2}{2C}\right) + 2\exp\left(-\frac{\min\{\beta^2, d^{\frac{1}{2}}\beta\}}{2C}\right).$$

Thus, $\left|\mathcal{M}_k^{\mathrm{Hess}} - \mathcal{M}_k^{\mathrm{Hess},\beta}\right|$ is exponentially small with overwhelming probability. Using (108) and setting β to be an arbitrarily small power of d, we obtain that, with overwhelming probability,

$$\sup_{1\leqslant k\leqslant n}\left|\mathcal{M}_k^{\mathrm{Hess}}\right| < d^{-(2+\alpha)+\zeta(4+2\alpha)}.$$

Adjusting the value of ζ, recalling that all of this has been proved on the stopped process, and applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot, z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

D.3.3 Bounds on the Higher Order Error Term in the Taylor Expansion, $\mathcal{E}_t^{\mathrm{High}}$

This section is devoted to showing that the high-order terms (third and higher-order terms derived in the Taylor expansion of the statistic φ) are negligible. Recall the third-order derivative in the Taylor expansion (78) is

$$-\frac{\gamma_k^3}{2}\int_0^1 (1-s)^2\langle\nabla^3\varphi(x_k - \gamma_k\, s\,\nabla_x\Psi(x_k; a_{k+1})), (\nabla_x\Psi(x_k; a_{k+1}))^{\otimes 3}\rangle\,\mathrm{d}s. \tag{109}$$

At first glance, these error terms look, in terms of d, large due to the tensor $(\nabla_x\Psi(x_k; a_{k+1}))^{\otimes 3}$, but the scaling will be sufficient to control this term.

Proposition D.11 (Higher-order error term)**.** *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 1.4). Let the statistic $S\colon \mathbb{R}^{2d}\times\Gamma \subset \mathbb{R}^{2d}\times\mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ be defined as in (18). Then for each fixed $z \in \Gamma$ and each $T > 0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T}\sum_{k=0}^{\lfloor(t\wedge\vartheta)d\rfloor-1}\left|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{High}}(S(\cdot, z))|\mathcal{F}_k\right]\right| \leqslant Cd^{-1/2}. \tag{110}$$

Proof. Fix indices $i, j \in \{1, 2, 3\}$. Let φ denote either

$$\varphi(x) = S_{ij}^{\mathrm{Re},z}(x) = \mathrm{Re}\, S_{ij}(x, z)$$

or

$$\varphi(x) = S_{ij}^{\mathrm{Im},z}(x) = \mathrm{Im}\, S_{ij}(x, z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot, z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. First, because ψ_i and the diagonal resolvents depend only on the coordinate pair (u_i, v_i), the third derivative of each entry of S is supported only on triples of derivative indices belonging to the same coordinate pair, so the inner product in (109) can be simplified as

$$\langle\nabla^3\varphi(x_k - \gamma_k\, s\,\nabla_x\Psi(x_k; a_{k+1})), (\nabla_x\Psi(x_k; a_{k+1}))^{\otimes 3}\rangle$$

$$= \sum_{h,i,j=1}^{2d}\left(\nabla^3\varphi(x_k - \gamma_k\, s\,\nabla_x\Psi(x_k; a_{k+1}))\right)_{hij}\left(\nabla_x\Psi(x_k)\right)_h\left(\nabla_x\Psi(x_k)\right)_i\left(\nabla_x\Psi(x_k)\right)_j$$

$$= \sum_{\substack{h=1 \\ i,j\in\{h,h+d\}}}^{d} \left(\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1}))\right)_{hij} \left(\nabla_x\Psi(x_k)\right)_h \left(\nabla_x\Psi(x_k)\right)_i \left(\nabla_x\Psi(x_k)\right)_j$$

$$+ \sum_{\substack{h=d+1 \\ i,j\in\{h,h-d\}}}^{2d} \left(\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1}))\right)_{hij} \left(\nabla_x\Psi(x_k)\right)_h \left(\nabla_x\Psi(x_k)\right)_i \left(\nabla_x\Psi(x_k)\right)_j .$$

Moreover, a simple computation shows that, for any $0 < s < 1$, we have

$$\begin{aligned}\left\|x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})\right\|_\infty &= \left\|x_k - s\frac{\gamma_k}{\sqrt{d}}\nabla_{r_1} f\left(r_k\right)\left(\nabla\psi(x_k)\right)^\top a_{k+1}\right\|_\infty \\ &\leqslant \left\|x_k\right\|_\infty + \frac{\bar\gamma}{\sqrt{d}}\left|\nabla_{r_1} f(r_k)\right| \left\|\left(\nabla\psi(x_k)\right)^\top a_{k+1}\right\|_\infty .\end{aligned}$$

If we write $a_{k+1} = \sqrt{K}v_k$ with $v_k \sim \mathcal{N}(0, I_d)$, then by (90) we have that, with overwhelming probability,

$$\left\|\left(\nabla\psi(x_k)\right)^\top a_{k+1}\right\|_\infty = \left\|\left(\nabla\psi(x_k)\right)^\top \sqrt{K}v_k\right\|_\infty \leqslant \left\|\nabla\psi(x_k)\right\|_{\mathrm{op}} \left\|\sqrt{K}\right\|_{\mathrm{op}} \left\|v_k\right\| \leqslant C(\left\|x_k\right\|_\infty)\sqrt{d}.$$

Additionally, by Lemma B.6,

$$\left|\nabla_{r_1} f(r_k)\right| \leqslant C(\alpha)L(f)\left(1 + \frac{1}{\sqrt{d}}\left\|K\right\|_{\mathrm{op}}^{1/2}\left\|W_k\right\|\right)^{\max\{1,\alpha\}} \qquad \text{w.o.p.}$$

Thus, for any $k \leqslant (t \wedge \vartheta)d$, we get

$$\left\|x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})\right\|_\infty \leqslant C(\left\|x_k\right\|_\infty),$$

where C is a constant depending on $\bar\gamma$, $\left\|K\right\|_{\mathrm{op}}$, $L(f)$, and α, but independent of d.

Next, we know

$$\left|\left(\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1}))\right)_{hij}\right| \leqslant \frac{1}{d}C(\left\|x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})\right\|_\infty) \leqslant \frac{C(\left\|x_k\right\|_\infty)}{d},$$

and so

$$\begin{aligned}&\left|\langle\nabla^3\varphi(x_k - \gamma_k\, s\, \nabla_x\Psi(x_k;a_{k+1})), \left(\nabla_x\Psi(x_k;a_{k+1})\right)^{\otimes 3}\rangle\right| \\ &\qquad\leqslant \frac{C(\left\|x_k\right\|_\infty)}{d} \sum_{\substack{h=1 \\ i,j\in\{h,h+d\}}}^{d} \left|\left(\nabla_x\Psi(x_k;a_{k+1})\right)_h \left(\nabla_x\Psi(x_k;a_{k+1})\right)_i \left(\nabla_x\Psi(x_k;a_{k+1})\right)_j\right| \\ &\qquad+ \frac{C(\left\|x_k\right\|_\infty)}{d} \sum_{\substack{h=d+1 \\ i,j\in\{h,h-d\}}}^{2d} \left|\left(\nabla_x\Psi(x_k;a_{k+1})\right)_h \left(\nabla_x\Psi(x_k;a_{k+1})\right)_i \left(\nabla_x\Psi(x_k;a_{k+1})\right)_j\right|\end{aligned}$$

with overwhelming probability.

Furthermore, we can bound

$$\begin{aligned}&\mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_x\Psi(x_k;a_{k+1})\right)_h \left(\nabla_x\Psi(x_k;a_{k+1})\right)_i \left(\nabla_x\Psi(x_k;a_{k+1})\right)_j\right| \middle| \mathcal{F}_k\right] \\ &\qquad\leqslant \frac{1}{d\sqrt{d}}\left\|\nabla\psi(x_k)\right\|_{\mathrm{op}}^3 \mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_{r_1} f\left(r_k\right)\right)^3 \left(a_{k+1}\right)_h \left(a_{k+1}\right)_i \left(a_{k+1}\right)_j\right| \middle| \mathcal{F}_k\right],\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_{r_1} f\left(r_k\right)\right)^3 \left(a_{k+1}\right)_h \left(a_{k+1}\right)_i \left(a_{k+1}\right)_j\right| \middle| \mathcal{F}_k\right] \\ &\qquad\leqslant \sqrt{\mathbb{E}_{a_{k+1}}\left[\left(\nabla_{r_1} f\left(r_k\right)\right)^6 \middle| \mathcal{F}_k\right]} \cdot \sqrt{\mathbb{E}_{a_{k+1}}\left[\left(a_{k+1}\right)_h^2 \left(a_{k+1}\right)_i^2 \left(a_{k+1}\right)_j^2\right]}\end{aligned}$$

$$\leqslant \sqrt{C(\alpha)(L(f))^6 \left(1+\frac{1}{\sqrt{d}}\left\|K\right\|_{\mathrm{op}}^{1/2}\left\|W_k\right\|\right)^{\max\{1,6\alpha\}}} \cdot \left\|K\right\|_{\mathrm{op}}^{3/2}.$$

Therefore, noting that $k \leqslant (t \wedge \vartheta)d$ and using (90), yields

$$\mathbb{E}_{a_{k+1}}\left[\left|\left(\nabla_x \Psi(x_k;a_{k+1})\right)_h \left(\nabla_x \Psi(x_k;a_{k+1})\right)_i \left(\nabla_x \Psi(x_k;a_{k+1})\right)_j\right| \middle| \mathcal{F}_k\right] \leqslant \frac{C(\|x_k\|_\infty)}{d\sqrt{d}}.$$

Consequently, for any $k \leqslant (t \wedge \vartheta)d$, the quantity

$$\mathbb{E}_{a_{k+1}}\left[\left|-\frac{\gamma_k^3}{2}\int_0^1 (1-s)^2 \langle \nabla^3 \varphi(x_k - \gamma_k\, s\, \nabla_x \Psi(x_k;a_{k+1})), \left(\nabla_x \Psi(x_k;a_{k+1})\right)^{\otimes 3}\rangle \,\mathrm{d}s\right| \middle| \mathcal{F}_k\right]$$

is bounded by $C(\|x_k\|_\infty)/(d\sqrt{d})$. Summing over at most Td indices gives

$$\sup_{0\leqslant t\leqslant T} \sum_{k=0}^{\lfloor (t\wedge\vartheta)d\rfloor -1} \left|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{High}}(\varphi)\middle|\mathcal{F}_k\right]\right| \leqslant Cd^{-1/2}.$$

Applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot, z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. Thus the third-order Taylor remainder in (78) vanish when d grows large after summing up over k. □

D.3.4 Bounds on the Lower Order Terms in the Hessian, $\mathcal{E}_t^{\mathrm{Hess}}$

We now bound the error term, $\sup_{0\leqslant t\leqslant T}\sum_{k=0}^{\lfloor (t\wedge\vartheta)d\rfloor -1}\left|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}\middle|\mathcal{F}_k\right]\right|$, in (87). For this, we utilize the operator norm and its dual norm, the nuclear norm.

Proposition D.12 (Hessian error term)**.** *Suppose $f\colon \mathbb{R}^2 \to \mathbb{R}$ is α-pseudo-Lipschitz with constant $L(f)$ (see Assumption 1.4). Let the statistic $S\colon \mathbb{R}^{2d}\times\Gamma \subset \mathbb{R}^{2d}\times\mathbb{C}^4 \to \mathbb{C}^{3\times 3}$ be defined as in (18). Then for each fixed $z\in\Gamma$ and each $T>0$, with overwhelming probability,*

$$\sup_{0\leqslant t\leqslant T} \sum_{k=0}^{\lfloor (t\wedge\vartheta)d\rfloor -1} \left|\mathbb{E}_{a_{k+1}}\left[\mathcal{E}_k^{\mathrm{Hess}}(S(\cdot,z))\middle|\mathcal{F}_k\right]\right| \leqslant C(L(f))^4 d^{-1}. \tag{111}$$

Proof. Fix indices $i,j\in\{1,2,3\}$. Let φ denote either

$$\varphi(x) = S_{ij}^{\mathrm{Re},z}(x) = \mathrm{Re}\, S_{ij}(x,z)$$

or

$$\varphi(x) = S_{ij}^{\mathrm{Im},z}(x) = \mathrm{Im}\, S_{ij}(x,z).$$

We prove the scalar estimate for this real-valued φ. The corresponding matrix-valued estimate for $S(\cdot,z)$ follows by applying the same bound to all real and imaginary parts and taking a finite union bound over the 3×3 entries. Define $\Pi_k := Q_k Q_k^\top$ and note that $\|\Pi_k\|^2 = \mathrm{rank}(\Pi_k) = 2$. Recall

$$\mathcal{E}_k^{\mathrm{Hess}}(\varphi) := -\frac{\gamma_k^2}{2d}\langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle + \frac{\gamma_k^2}{2d}\langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle$$

where $\mathcal{B} := \nabla_{r_1} f\,(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top$. First, we consider the following term

$$\begin{aligned}
\left|\langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle\right| &= \left|\langle \nabla_{r_1} f\,(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top, \sqrt{K}\Pi_k\sqrt{K}\rangle\right| \\
&\leqslant \left\|\sqrt{K}\Pi_k\sqrt{K}\right\|_* \left\|\nabla_{r_1} f\,(r_k)^2\, \nabla\psi(x_k)\nabla^2\varphi(x_k)\,(\nabla\psi(x_k))^\top\right\|_{\mathrm{op}} \\
&\leqslant \left\|\sqrt{K}\Pi_k\sqrt{K}\right\|_* \nabla_{r_1} f\,(r_k)^2\, \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}}
\end{aligned}$$

$$\leqslant \|K\|_{\mathrm{op}} \|\Pi_k\|_* \nabla_{r_1} f(r_k)^2 \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}}.$$

From Lemma B.6, we have

$$\mathbb{E}_{a_{k+1}}[|\nabla_{r_1} f(r_k)|^2 \,|\mathcal{F}_k] \leqslant C(L(f))^2 \left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W_k\|\right)^{\max\{1,2\alpha\}}.$$

Using (90), we have that $\|\nabla\psi(x_k)\|_{\mathrm{op}} \leqslant C(\|x_k\|_\infty)$. Moreover, we also, by Lemma D.6, have $\left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}} = \left\|\nabla_x^2 S_{ij}(x_k, z)\right\|_{\mathrm{op}} \leqslant \left\|\nabla_x^2 S(x_k, \cdot)\right\|_\Gamma \leqslant \frac{C(\|x_k\|_\infty)}{d}$. Since $k \leqslant (t \wedge \vartheta)d$, we get

$$\mathbb{E}_{a_{k+1}}\left[\left|\langle \mathcal{B}, \sqrt{K}\Pi_k\sqrt{K}\rangle\right| \,|\mathcal{F}_k\right] \leqslant \frac{C(\|x_k\|_\infty)}{d}(L(f))^2 \left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W_k\|\right)^{\max\{1,2\alpha\}}. \tag{112}$$

Similarly we get that

$$\begin{aligned}
\left|\langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle\right| &= \left|\langle \nabla_{r_1} f(r_k)^2 \nabla\psi(x_k)\nabla^2\varphi(x_k) (\nabla\psi(x_k))^\top, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle\right| \\
&\leqslant \left\|\left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\right\|_* \left\|\nabla_{r_1} f(r_k)^2 \nabla\psi(x_k)\nabla^2\varphi(x_k) (\nabla\psi(x_k))^\top\right\|_{\mathrm{op}} \\
&\leqslant \left\|\sqrt{K}\Pi_k v_k\right\|^2 \nabla_{r_1} f(r_k)^2 \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}} \\
&\leqslant \|K\|_{\mathrm{op}} \|\Pi_k v_k\|^2 \nabla_{r_1} f(r_k)^2 \|\nabla\psi(x_k)\|_{\mathrm{op}}^2 \left\|\nabla^2\varphi(x_k)\right\|_{\mathrm{op}}.
\end{aligned}$$

Upon taking expectations, using Cauchy–Schwarz we have

$$\mathbb{E}_{a_{k+1}}[\|\Pi_k v_k\|^2 \nabla_{r_1} f(r_k)^2 |\mathcal{F}_k] \leqslant \left(\mathbb{E}[\|\Pi_k v_k\|^4 |\mathcal{F}_k]\right)^{1/2} \left(\mathbb{E}[|\nabla_{r_1} f(r_k)|^4 |\mathcal{F}_k]\right)^{1/2}.$$

Now by Lemma B.6 we know

$$\mathbb{E}_{a_{k+1}}[|\nabla_{r_1} f(r_k)|^4 \,|\mathcal{F}_k] \leqslant C(L(f))^4 \left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W_k\|\right)^{\max\{1,4\alpha\}},$$

and $\mathbb{E}_{a_{k+1}}\left[\|\Pi_k v_k\|^4 \middle| \mathcal{F}_k\right] = 8$, as Π_k is a projection. Using Lemma D.6 on the growth of φ, yields

$$\mathbb{E}_{a_{k+1}}\left[\left|\langle \mathcal{B}, \left(\sqrt{K}\Pi_k v_k\right)^{\otimes 2}\rangle\right| \middle| \mathcal{F}_k\right] \leqslant \frac{C(\|x_k\|_\infty)}{d}(L(f))^4 \left(1 + \frac{1}{\sqrt{d}} \|K\|_{\mathrm{op}}^{1/2} \|W_k\|\right)^{\max\{1,4\alpha\}}. \tag{113}$$

As $k \leqslant (t \wedge \vartheta)d$, then $\frac{1}{\sqrt{d}} \|W_k\| \leqslant \sqrt{M}$. The result then immediately follows by combining (112) and (113) and summing up with the extra factor γ_k^2/d. Applying the scalar bound to the real and imaginary parts of all nine entries of $S(\cdot, z)$ and taking a finite union bound gives the displayed matrix norm bound. Since the matrix dimension is fixed, this only changes the constant. □

Note that since $|\Gamma^\delta| \leqslant C_\Gamma d^{4\delta}$, all of the same estimates hold uniformly over $z \in \Gamma^\delta$ by a union bound.

E Entropy Barrier and High-probability Exponential Decay

This section analyzes the homogenized dynamics in the isotropic squared parameterization setting. We prove that, for sufficiently small constant stepsize, the risk decays exponentially with high probability and the dynamics exist globally. The argument is based on an entropy barrier tailored to the coordinatewise structure of the SDE. First, we derive an exact product identity for each coordinate pair $(\mathcal{U}_{t,i}, \mathcal{V}_{t,i})$ and introduce logarithmic coordinates in which the residual $\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1$ has an entropy representation. We then derive an exact SDE for the empirical entropy and prove deterministic barrier estimates comparing the entropy to both the coordinate size and the risk. These estimates yield an exponential supermartingale argument, giving explicit confidence bounds and a high-probability

exponential decay theorem. Finally, we record dynamical consequences of this decay, including integrability of the risk and uniform separation from the saddle at the origin.

In this section, we specialize to the isotropic setting $\beta^* = \mathbb{1}_d$ and $a \sim \mathcal{N}(0, I_d)$, so that the risk in (3) becomes

$$\mathcal{R}(x) = \frac{1}{4d} \sum_{i=1}^{d} \left(u_i^2 - v_i^2 - 1\right)^2.$$

We write $\mathcal{X}_t = (\mathcal{U}_t, \mathcal{V}_t)$ for the corresponding homogenized process. For constant stepsize $\gamma > 0$, the coordinatewise SDE in (4) takes the form

$$\mathrm{d}\mathcal{U}_{t,i} = -\gamma \mathcal{U}_{t,i} \left(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1\right) \mathrm{d}t + 2\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{U}_{t,i}\, \mathrm{d}\mathfrak{B}_{t,i}, \tag{114}$$

$$\mathrm{d}\mathcal{V}_{t,i} = \gamma \mathcal{V}_{t,i} \left(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1\right) \mathrm{d}t - 2\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{V}_{t,i}\, \mathrm{d}\mathfrak{B}_{t,i}, \tag{115}$$

for $i = 1, \ldots, d$, where $\mathfrak{B}_t$ is a standard Brownian motion in $\mathbb{R}^d$. We initialize the dynamics at

$$\mathcal{U}_{0,i} = \mathcal{V}_{0,i} = 1, \qquad i = 1, \ldots, d.$$

Since the coefficients are locally Lipschitz, the SDE admits a unique maximal local strong solution on a random interval $[0, \zeta)$, where $\zeta \in (0, \infty]$ denotes the explosion time.

E.1 Exact Product Identity and Logarithmic Coordinates

We first record an exact product identity for each coordinate pair.

Lemma E.1 (Exact product identity)**.** *For each $i = 1, \ldots, d$ and all $0 \leqslant t < \zeta$, we have*

$$\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2 = \exp\Big(-8\gamma^2 I_t\Big), \quad \textit{where} \quad I_t := \int_0^t \mathcal{R}(\mathcal{X}_s)\, \mathrm{d}s.$$

In particular, $\mathcal{U}_{t,i}^2 > 0$ and $\mathcal{V}_{t,i}^2 > 0$ for all $0 \leqslant t < \zeta$.

Proof. By Itô's formula,

$$\mathrm{d}\mathcal{U}_{t,i}^2 = \mathcal{U}_{t,i}^2 \Big(-2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2 \mathcal{R}(\mathcal{X}_t)\Big)\, \mathrm{d}t + 4\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{U}_{t,i}^2\, \mathrm{d}\mathfrak{B}_{t,i},$$

$$\mathrm{d}\mathcal{V}_{t,i}^2 = \mathcal{V}_{t,i}^2 \Big(+2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2 \mathcal{R}(\mathcal{X}_t)\Big)\, \mathrm{d}t - 4\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathcal{V}_{t,i}^2\, \mathrm{d}\mathfrak{B}_{t,i}.$$

Applying Itô's product rule to $\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2$, the martingale terms cancel and the quadratic covariation contributes

$$\mathrm{d}\langle \mathcal{U}_{\cdot,i}^2, \mathcal{V}_{\cdot,i}^2 \rangle_t = -16\gamma^2 \mathcal{R}(\mathcal{X}_t) \mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2\, \mathrm{d}t,$$

so that

$$\mathrm{d}(\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2) = -8\gamma^2 \mathcal{R}(\mathcal{X}_t) \mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2\, \mathrm{d}t.$$

Solving this scalar ODE and using $\mathcal{U}_{0,i}^2 \mathcal{V}_{0,i}^2 = 1$ gives

$$\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2 = \exp\Big(-8\gamma^2 I_t\Big) > 0.$$

Because both factors are nonnegative, each must be strictly positive. □

Define

$$\rho(I_t) := \sqrt{1 + 4e^{-8\gamma^2 I_t}}.$$

Then

$$\frac{\rho(I_t) + 1}{2} > 0, \qquad \frac{\rho(I_t) - 1}{2} > 0, \qquad \frac{\rho(I_t) + 1}{2} \cdot \frac{\rho(I_t) - 1}{2} = e^{-8\gamma^2 I_t}.$$

By Lemma E.1, the product of the two normalizing factors matches $\mathcal{U}_{t,i}^2 \mathcal{V}_{t,i}^2$. Hence, for each i, we may define a unique real-valued process

$$\mathcal{Z}_{t,i} := \log\left(\frac{\mathcal{U}_{t,i}^2}{(\rho(I_t) + 1)/2}\right) = -\log\left(\frac{\mathcal{V}_{t,i}^2}{(\rho(I_t) - 1)/2}\right),$$

such that

$$\mathcal{U}_{t,i}^2 = \frac{\rho(I_t)+1}{2} e^{\mathcal{Z}_{t,i}}, \qquad \mathcal{V}_{t,i}^2 = \frac{\rho(I_t)-1}{2} e^{-\mathcal{Z}_{t,i}}.$$

For $c > 0$ and $x > 0$, define

$$f_c(x) := x - c - c\log(x/c),$$

and for $\rho \geqslant 1$ define

$$\Phi_\rho(z) := \rho(\cosh z - 1) + \sinh z - z.$$

Lemma E.2 (Logarithmic-coordinate identities)**.** *For every $0 \leqslant t < \zeta$ and every $i = 1, \dots, d$,*

$$\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1 = \rho(I_t)\sinh(\mathcal{Z}_{t,i}) + \cosh(\mathcal{Z}_{t,i}) - 1, \tag{116}$$

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 = \rho(I_t)\cosh(\mathcal{Z}_{t,i}) + \sinh(\mathcal{Z}_{t,i}), \tag{117}$$

$$\Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2). \tag{118}$$

Moreover, $\mathcal{U}_{t,i} > 0$ and $\mathcal{V}_{t,i} > 0$ for all $0 \leqslant t < \zeta$, and hence

$$\mathrm{d}\log \mathcal{U}_{t,i} = \Big(-\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) - 2\gamma^2 \mathcal{R}(\mathcal{X}_t)\Big)\,\mathrm{d}t + 2\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\,\mathrm{d}\mathfrak{B}_{t,i}, \tag{119}$$

$$\mathrm{d}\log \mathcal{V}_{t,i} = \Big(+\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) - 2\gamma^2 \mathcal{R}(\mathcal{X}_t)\Big)\,\mathrm{d}t - 2\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\,\mathrm{d}\mathfrak{B}_{t,i}. \tag{120}$$

Proof. The identities (116)–(118) follow directly from the parametrization

$$\mathcal{U}_{t,i}^2 = \frac{\rho(I_t)+1}{2} e^{\mathcal{Z}_{t,i}}, \qquad \mathcal{V}_{t,i}^2 = \frac{\rho(I_t)-1}{2} e^{-\mathcal{Z}_{t,i}}.$$

Since $\mathcal{U}_{0,i} = \mathcal{V}_{0,i} = 1$ and neither process can hit 0 before ζ by Lemma E.1, continuity gives strict positivity. The logarithmic SDEs then follow from Itô's formula applied to $\log \mathcal{U}_{t,i}$ and $\log \mathcal{V}_{t,i}$. □

For each coordinate, define the entropy density

$$h_{t,i} := \Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2), \qquad 0 \leqslant t < \zeta. \tag{121}$$

The empirical entropy is then the average of these coordinate densities:

$$H_t := \frac{1}{d}\sum_{i=1}^d h_{t,i}. \tag{122}$$

Equivalently,

$$H_t = \frac{1}{d}\sum_{i=1}^d \Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = \frac{1}{d}\sum_{i=1}^d \left[f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2)\right].$$

At time $t = 0$,

$$\rho(I_0) = \sqrt{5}, \qquad \mathcal{Z}_{0,i} = -\log\Big(\tfrac{\sqrt{5}+1}{2}\Big),$$

and hence

$$h_{0,i} = 2 - \sqrt{5} + \log\Big(\tfrac{\sqrt{5}+1}{2}\Big) \qquad \text{for every } i = 1, \dots, d.$$

Therefore,

$$H_0 = 2 - \sqrt{5} + \log\Big(\tfrac{\sqrt{5}+1}{2}\Big),$$

which is independent of d.

E.2 Exact Dynamics of the Empirical Entropy

Recall that

$$H_t = \frac{1}{d}\sum_{i=1}^{d} h_{t,i}, \qquad h_{t,i} := \Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}).$$

We next derive the SDE satisfied by the empirical entropy H_t.

Proposition E.3 (Empirical entropy SDE)**.** *For $0 \leqslant t < \zeta$,*

$$\mathrm{d}H_t = \left(-8\gamma\mathcal{R}(\mathcal{X}_t) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)\left[\frac{1}{d}\sum_{i=1}^{d}(\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2) + \rho(I_t)\right]\right)\mathrm{d}t + \mathrm{d}M_t, \tag{123}$$

where

$$M_t = 4\gamma\frac{1}{d}\sum_{i=1}^{d}\int_0^t \mathbb{1}_{\{s<\zeta\}}\sqrt{\mathcal{R}(\mathcal{X}_s)}\,(\mathcal{U}_{s,i}^2 - \mathcal{V}_{s,i}^2 - 1)\,\mathrm{d}\mathfrak{B}_{s,i},$$

and

$$\mathrm{d}\langle M\rangle_t = \mathbb{1}_{\{t<\zeta\}}\frac{64\gamma^2}{d}\mathcal{R}(\mathcal{X}_t)^2\,\mathrm{d}t. \tag{124}$$

Proof. Since

$$\partial_x f_c(x) = 1 - \frac{c}{x}, \qquad \partial_{xx} f_c(x) = \frac{c}{x^2}, \qquad \partial_c f_c(x) = \log(c/x),$$

Itô's formula gives

$$\begin{aligned}\mathrm{d}f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) = &\left[\left(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\right)\Big(-2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)\Big) + 8\gamma^2\mathcal{R}(\mathcal{X}_t)\tfrac{\rho(I_t)+1}{2}\right]\mathrm{d}t\\ &+ 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\Big(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\Big)\mathrm{d}\mathfrak{B}_{t,i} + \log\left(\frac{(\rho(I_t)+1)/2}{\mathcal{U}_{t,i}^2}\right)\mathrm{d}\Big(\tfrac{\rho(I_t)+1}{2}\Big),\end{aligned}$$

and similarly

$$\begin{aligned}\mathrm{d}f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2) = &\left[\left(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\right)\Big(+2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)\Big) + 8\gamma^2\mathcal{R}(\mathcal{X}_t)\tfrac{\rho(I_t)-1}{2}\right]\mathrm{d}t\\ &- 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\Big(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\Big)\mathrm{d}\mathfrak{B}_{t,i} + \log\left(\frac{(\rho(I_t)-1)/2}{\mathcal{V}_{t,i}^2}\right)\mathrm{d}\Big(\tfrac{\rho(I_t)-1}{2}\Big).\end{aligned}$$

Now we compute

$$\begin{aligned}\Big(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\Big) - \Big(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\Big) &= \mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1,\\ \Big(\mathcal{U}_{t,i}^2 - \tfrac{\rho(I_t)+1}{2}\Big) + \Big(\mathcal{V}_{t,i}^2 - \tfrac{\rho(I_t)-1}{2}\Big) &= \mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 - \rho(I_t).\end{aligned}$$

Moreover, the two moving-target terms have the same finite-variation factor

$$\mathrm{d}\left(\frac{\rho(I_t)+1}{2}\right) = \mathrm{d}\left(\frac{\rho(I_t)-1}{2}\right) = \frac{1}{2}\,\mathrm{d}\rho(I_t),$$

and also we have

$$\mathcal{U}_{t,i}^2\mathcal{V}_{t,i}^2 = \frac{\rho(I_t)+1}{2}\cdot\frac{\rho(I_t)-1}{2}.$$

Therefore the two moving-target terms cancel:

$$\log\left(\frac{(\rho(I_t)+1)/2}{\mathcal{U}_{t,i}^2}\right) + \log\left(\frac{(\rho(I_t)-1)/2}{\mathcal{V}_{t,i}^2}\right) = 0.$$

Summing the two Itô identities therefore yields

$$\begin{aligned}\mathrm{d}\Phi_{\rho(I_t)}(\mathcal{Z}_{t,i}) = \Big(&- 2\gamma(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2 + 4\gamma^2\mathcal{R}(\mathcal{X}_t)(\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 + \rho(I_t))\Big)\mathrm{d}t \\ &+ 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)\mathrm{d}\mathfrak{B}_{t,i}.\end{aligned}$$

Averaging over i and using

$$\frac{1}{d}\sum_{i=1}^{d}(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2 = 4\mathcal{R}(\mathcal{X}_t)$$

gives (123). The bracket identity (124) follows from independence of the Brownian motions:

$$\mathrm{d}\langle M\rangle_t = \frac{16\gamma^2}{d^2}\sum_{i=1}^{d}\mathcal{R}(\mathcal{X}_t)(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2\,\mathrm{d}t = \frac{64\gamma^2}{d}\mathcal{R}(\mathcal{X}_t)^2\,\mathrm{d}t.$$

□

E.3 Barrier Estimates

We next derive deterministic barrier estimates. The first shows that the empirical entropy controls the average size of $\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2$. The second gives a dimension-free comparison between the entropy density $\Phi_{\rho(I_t)}(\mathcal{Z}_{t,i})$ and the squared residual,

$$\left(\rho(I_t)\sinh(\mathcal{Z}_{t,i}) + \cosh(\mathcal{Z}_{t,i}) - 1\right)^2,$$

provided we work below a coordinatewise entropy barrier.

Lemma E.4 (Pointwise control of $\mathcal{U}^2 + \mathcal{V}^2$ by the entropy)**.** *For every $\rho \geqslant 1$ and every $z \in \mathbb{R}$,*

$$\rho\cosh z + \sinh z \leqslant 2\Phi_\rho(z) + 2\rho\log 2. \tag{125}$$

Consequently, for all $0 \leqslant t < \zeta$,

$$\frac{1}{d}\sum_{i=1}^{d}(\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2) \leqslant 2H_t + 2\rho(I_t)\log 2 \leqslant 2H_t + 2\sqrt{5}\log 2. \tag{126}$$

Proof. It suffices to prove the claim for $\rho > 1$, since both sides of (125) are continuous in ρ and the case $\rho = 1$ follows by taking the limit $\rho \downarrow 1$.

Fix $c > 0$ and define

$$g_c(x) := f_c(x) - \frac{x}{2} = \frac{x}{2} - c - c\log(x/c), \qquad x > 0.$$

Then $g_c'(x) = \frac{1}{2} - \frac{c}{x}$, so g_c is minimized at $x = 2c$, where $g_c(2c) = -c\log 2$. Hence

$$f_c(x) \geqslant \frac{x}{2} - c\log 2 \qquad \text{for all } x > 0.$$

Applying this with

$$(c, x) = \left(\tfrac{\rho+1}{2},\ \tfrac{\rho+1}{2}e^z\right), \qquad (c, x) = \left(\tfrac{\rho-1}{2},\ \tfrac{\rho-1}{2}e^{-z}\right),$$

and using (118), we obtain

$$\begin{aligned}\Phi_\rho(z) &\geqslant \frac{1}{2}\left(\tfrac{\rho+1}{2}e^z + \tfrac{\rho-1}{2}e^{-z}\right) - \left(\tfrac{\rho+1}{2} + \tfrac{\rho-1}{2}\right)\log 2 \\ &= \frac{1}{2}\left(\rho\cosh z + \sinh z\right) - \rho\log 2.\end{aligned}$$

Rearranging gives (125). Averaging with $(\rho, z) = (\rho(I_t), \mathcal{Z}_{t,i})$ and using $\rho(I_t) \leqslant \rho(0) = \sqrt{5}$ yields (126). □

Lemma E.5 (Coercivity under a coordinatewise entropy barrier)**.** *Fix $L_* > H_0$ and define*

$$K_{L_*} := \Big\{(\rho, z) : 1 \leqslant \rho \leqslant \sqrt{5},\ \Phi_\rho(z) \leqslant L_*\Big\}.$$

Then K_{L_*} is compact. Moreover, there exist constants

$$0 < m_{L_*} \leqslant M_{L_*} < \infty$$

depending only on L_* such that, for every $(\rho, z) \in K_{L_*}$,

$$m_{L_*}\Phi_\rho(z) \leqslant \big(\rho \sinh z + \cosh z - 1\big)^2 \leqslant M_{L_*}\Phi_\rho(z). \tag{127}$$

Proof. Since $\rho \in [1, \sqrt{5}]$ and $\Phi_\rho(z) \to \infty$ as $|z| \to \infty$ uniformly over that interval, K_{L_*} is compact by Heine–Borel. Consider

$$Q(\rho, z) := \frac{(\rho \sinh z + \cosh z - 1)^2}{\Phi_\rho(z)} \qquad (z \neq 0).$$

As $z \to 0$, we have

$$\rho \sinh z + \cosh z - 1 = \rho z + \frac{z^2}{2} + O(z^3),$$
$$\Phi_\rho(z) = \rho(\cosh z - 1) + \sinh z - z = \frac{\rho}{2} z^2 + O(z^3),$$

and therefore

$$Q(\rho, z) \to 2\rho \qquad \text{as } z \to 0.$$

Thus Q extends continuously across $\{z = 0\}$ by setting $Q(\rho, 0) := 2\rho$. Next,

$$\partial_z\big(\rho \sinh z + \cosh z - 1\big) = \rho \cosh z + \sinh z > 0 \qquad (\rho \geqslant 1),$$

so $\rho \sinh z + \cosh z - 1$ has the unique zero $z = 0$. Also,

$$\Phi'_\rho(z) = \rho \sinh z + \cosh z - 1, \qquad \Phi''_\rho(z) = \rho \cosh z + \sinh z,$$

with $\Phi_\rho(0) = \Phi'_\rho(0) = 0$ and $\Phi''_\rho(0) = \rho > 0$. Hence $\Phi_\rho(z) > 0$ for all $z \neq 0$. Therefore the continuous extension of Q is strictly positive on the compact set K_{L_*}, so it attains a positive minimum and finite maximum:

$$0 < m_{L_*} := \min_{K_{L_*}} Q \leqslant \max_{K_{L_*}} Q =: M_{L_*} < \infty.$$

This is exactly (127). □

Lemma E.6 (Risk–entropy comparison below the coordinate barrier)**.** *Fix $L_* > H_0$. On any time interval on which*

$$\max_{1 \leqslant i \leqslant d} h_{t,i} < L_*,$$

we have

$$\frac{m_{L_*}}{4} H_t \leqslant \mathcal{R}(\mathfrak{X}_t) \leqslant \frac{M_{L_*}}{4} H_t. \tag{128}$$

Proof. If $\max_i h_{t,i} < L_*$, then

$$(\rho(I_t), \mathfrak{Z}_{t,i}) \in K_{L_*} \qquad \text{for every } i.$$

By Lemma E.5 and (116),

$$m_{L_*} h_{t,i} \leqslant (\mathfrak{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2 \leqslant M_{L_*} h_{t,i}.$$

Averaging over i and using

$$\mathcal{R}(\mathfrak{X}_t) = \frac{1}{4d} \sum_{i=1}^{d} (\mathfrak{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2$$

gives (128). □

E.4 Exponential Decay with Explicit Confidence

We now prove exponential decay of the empirical entropy, and hence of the risk, using a coordinate-wise entropy barrier.

Fix $H_* > H_0$ and $L_* > H_0$, and define the stopping time

$$\tau_{H,L} := \inf\left\{t \in [0,\zeta) : H_t \geqslant H_* \text{ or } \max_{1\leqslant i\leqslant d} h_{t,i} \geqslant L_*\right\}, \tag{129}$$

with the convention $\inf \emptyset = \infty$.

Lemma E.7 (The entropy barrier is hit before explosion)**.** *Almost surely,*

$$\tau_{H,L} \leqslant \zeta.$$

Proof. Suppose, toward a contradiction, that $\zeta < \tau_{H,L}$ on some sample path. Then $H_t < H_*$ and $h_{t,i} < L_*$ for every i and all $t < \zeta$. Hence

$$(\rho(I_t), \mathcal{Z}_{t,i}) \in K_{L_*} \qquad \text{for all } i \text{ and all } t < \zeta.$$

Since K_{L_*} is compact, the logarithmic coordinates $\mathcal{Z}_{t,i}$ remain bounded on $[0,\zeta)$. By the parametrization

$$\mathcal{U}_{t,i}^2 = \frac{\rho(I_t)+1}{2} e^{\mathcal{Z}_{t,i}}, \qquad \mathcal{V}_{t,i}^2 = \frac{\rho(I_t)-1}{2} e^{-\mathcal{Z}_{t,i}},$$

the coordinates $\mathcal{U}_{t,i}$ and $\mathcal{V}_{t,i}$ remain bounded on every finite interval $[0,T] \subset [0,\zeta)$. Therefore the coefficients in (114)–(115) remain bounded and locally Lipschitz on a neighborhood of the attained path. The standard continuation criterion for SDEs then extends the solution beyond ζ, contradicting maximality. Thus $\tau_{H,L} \leqslant \zeta$ almost surely. □

On $[0, \tau_{H,L})$, Lemma E.6 gives

$$\frac{m_{L_*}}{4} H_t \leqslant \mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_t,$$

and Lemma E.4 gives

$$\frac{1}{d}\sum_{i=1}^{d} (\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2) \leqslant 2H_* + 2\sqrt{5}\log 2, \qquad t < \tau_{H,L}. \tag{130}$$

Define

$$\lambda_{H,L} := \gamma\Big(8 - 4\gamma(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big)\frac{m_{L_*}}{4}, \tag{131}$$

and

$$\sigma_{H,L}^2 := \frac{4\gamma^2 M_{L_*}^2}{d}. \tag{132}$$

Lemma E.8 (Drift and quadratic-variation bounds before the barrier)**.** *Assume*

$$0 < \gamma < \frac{2}{2H_* + 2\sqrt{5}\log 2 + \sqrt{5}}, \tag{133}$$

so that $\lambda_{H,L} > 0$. Then, for $t < \tau_{H,L}$,

$$\mathrm{d}H_t \leqslant -\lambda_{H,L} H_t \,\mathrm{d}t + \mathrm{d}M_t, \tag{134}$$

and

$$\mathrm{d}\langle M\rangle_t \leqslant \sigma_{H,L}^2 H_t^2 \,\mathrm{d}t. \tag{135}$$

Proof. Starting from Proposition E.3 and using (130),

$$\begin{aligned}\mathrm{d}H_t &\leqslant \Big(-8\gamma\mathcal{R}(\mathcal{X}_t) + 4\gamma^2\mathcal{R}(\mathcal{X}_t)(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big)\mathrm{d}t + \mathrm{d}M_t \\ &= -\gamma\Big(8 - 4\gamma(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big)\mathcal{R}(\mathcal{X}_t)\,\mathrm{d}t + \mathrm{d}M_t.\end{aligned}$$

Using the lower bound in (128) gives (134). Similarly, from (124) and the upper bound in (128),

$$\mathrm{d}\langle M\rangle_t = \frac{64\gamma^2}{d}\mathcal{R}(\mathcal{X}_t)^2\,\mathrm{d}t \leqslant \frac{64\gamma^2}{d}\left(\frac{M_{L_*}}{4}H_t\right)^2 \mathrm{d}t = \frac{4\gamma^2 M_{L_*}^2}{d} H_t^2\,\mathrm{d}t.$$

This is (135). □

The next result gives exponential decay up to the coordinatewise barrier.

Theorem E.9 (Stopped exponential decay with explicit confidence)**.** *Let $H_*, L_* > H_0$, and define $\lambda_{H,L}$ and $\sigma^2_{H,L}$ by (131)–(132). Assume*

$$0 < \gamma < \frac{2}{2H_* + 2\sqrt{5}\log 2 + \sqrt{5}}.$$

Fix $\theta > 0$ such that

$$0 < \theta < \frac{2\lambda_{H,L}}{\sigma^2_{H,L}},$$

and define

$$\mu_{H,L}(\theta) := \lambda_{H,L} - \frac{\theta}{2}\sigma^2_{H,L} > 0.$$

Then

$$\mathbb{P}\left(\sup_{0 \leqslant s \leqslant \tau_{H,L}} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_*\right) \leqslant \left(\frac{H_0}{H_*}\right)^{\theta}. \tag{136}$$

Proof. Fix $0 < \varepsilon < H_0$ and define

$$\tau_\varepsilon := \inf\{t \in [0, \zeta) : H_t \leqslant \varepsilon\}, \qquad \tau := \tau_{H,L} \wedge \tau_\varepsilon.$$

On $[0, \tau]$, we have $H_t \in [\varepsilon, H_*]$, so Itô's formula gives

$$\log H_{t\wedge\tau} = \log H_0 + \int_0^{t\wedge\tau} \frac{1}{H_s}\,\mathrm{d}H_s - \frac{1}{2}\int_0^{t\wedge\tau} \frac{1}{H_s^2}\,\mathrm{d}\langle M\rangle_s.$$

By Lemma E.8,

$$\log H_{t\wedge\tau} \leqslant \log H_0 - \lambda_{H,L}(t\wedge\tau) - \frac{1}{2}\int_0^{t\wedge\tau} \frac{1}{H_s^2}\,\mathrm{d}\langle M\rangle_s + N_t,$$

where

$$N_t := \int_0^{t\wedge\tau} \frac{1}{H_s}\,\mathrm{d}M_s.$$

Moreover,

$$\langle N\rangle_t = \int_0^{t\wedge\tau} \frac{1}{H_s^2}\,\mathrm{d}\langle M\rangle_s \leqslant \sigma^2_{H,L}(t\wedge\tau).$$

Thus, for every $\theta > 0$,

$$\mathcal{E}_t := \exp\left(\theta N_t - \frac{\theta^2}{2}\langle N\rangle_t\right)$$

is a positive local martingale, hence a supermartingale. Since

$$\mu_{H,L}(\theta) = \lambda_{H,L} - \frac{\theta}{2}\sigma^2_{H,L},$$

we obtain

$$\begin{aligned} e^{\theta\mu_{H,L}(\theta)(t\wedge\tau)} H^\theta_{t\wedge\tau} &\leqslant H_0^\theta \exp\Big(\theta N_t - \theta(\lambda_{H,L} - \mu_{H,L}(\theta))(t\wedge\tau)\Big) \\ &= H_0^\theta \exp\Big(\theta N_t - \frac{\theta^2}{2}\sigma^2_{H,L}(t\wedge\tau)\Big) \\ &\leqslant H_0^\theta \mathcal{E}_t. \end{aligned}$$

Ville's maximal inequality gives

$$\mathbb{P}\left(\sup_{0 \leqslant s \leqslant \tau} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_*\right) \leqslant \left(\frac{H_0}{H_*}\right)^{\theta}.$$

It remains to remove the auxiliary lower stopping time τ_ε. Define

$$\tau_0 := \inf\{t \in [0, \zeta) : H_t = 0\}, \qquad S_\varepsilon := \tau_{H,L} \wedge \tau_\varepsilon, \qquad S_0 := \tau_{H,L} \wedge \tau_0.$$

Since H_t is continuous and nonnegative, $S_\varepsilon \uparrow S_0$ pathwise as $\varepsilon \downarrow 0$. Therefore,

$$\left\{ \sup_{0\leqslant s\leqslant S_\varepsilon} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_* \right\} \uparrow \left\{ \sup_{0\leqslant s\leqslant S_0} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_* \right\}.$$

By continuity from below of probability, the estimate obtained above gives

$$\mathbb{P}\left(\sup_{0\leqslant s\leqslant S_0} e^{\mu_{H,L}(\theta)s} H_s \geqslant H_* \right) \leqslant \left(\frac{H_0}{H_*} \right)^\theta.$$

Finally, $H_t = 0$ is absorbing. Indeed, if $H_t = 0$, then every coordinate entropy density $h_{t,i}$ vanishes, hence $\mathcal{Z}_{t,i} = 0$ for every i. Therefore

$$\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1 = 0 \qquad \text{for all } i,$$

so $\mathcal{R}(\mathcal{X}_t) = 0$. The drift and diffusion coefficients in (114)–(115) then vanish, and by pathwise uniqueness the solution remains constant thereafter. Thus

$$\sup_{0\leqslant s\leqslant S_0} e^{\mu_{H,L}(\theta)s} H_s = \sup_{0\leqslant s\leqslant \tau_{H,L}} e^{\mu_{H,L}(\theta)s} H_s.$$

This proves (136). □

We now give a pathwise criterion for removing the coordinatewise barrier. Let

$$V_{H,L}(\theta) := \frac{M_{L_*}}{4} \frac{H_*}{\mu_{H,L}(\theta)}.$$

For $\eta \in (0,1)$, define

$$R_{H,L}(\theta,\eta) := \sqrt{2V_{H,L}(\theta) \log\left(\frac{4d}{\eta} \right)},$$

and

$$p_{H,L}(\theta) := \exp\left(-4\gamma^2 V_{H,L}(\theta)\right).$$

Set

$$A_{H,L}(\theta) := \operatorname{arcsinh}\left(\frac{1}{2p_{H,L}(\theta)} \right), \qquad B_{H,L}(\theta,\eta) := 2\cosh(A_{H,L}(\theta) + 8\gamma R_{H,L}(\theta,\eta)).$$

Finally, define

$$c_{H,L}(\theta,\eta) := \frac{p_{H,L}(\theta)^2}{B_{H,L}(\theta,\eta)}, \qquad \underline{c}_{H,L}(\theta) := \frac{\sqrt{1+4p_{H,L}(\theta)^2}-1}{2}, \qquad \overline{c} := \frac{\sqrt{5}+1}{2},$$

and

$$\mathfrak{L}_{H,L}(\theta,\eta) := 2 \sup_{\substack{c\in[\underline{c}_{H,L}(\theta),\overline{c}] \\ x\in[c_{H,L}(\theta,\eta),B_{H,L}(\theta,\eta)]}} f_c(x). \tag{137}$$

Lemma E.10 (Removal of the coordinatewise entropy barrier)**.** *Assume the hypotheses of Theorem E.9. Fix $\eta \in (0,1)$ and suppose that*

$$L_* > \mathfrak{L}_{H,L}(\theta,\eta). \tag{138}$$

Then

$$\mathbb{P}\Big(\zeta = \infty \text{ and } H_t \leqslant H_* e^{-\mu_{H,L}(\theta)t} \text{ for all } t \geqslant 0\Big) \geqslant 1 - \left(\frac{H_0}{H_*} \right)^\theta - \eta. \tag{139}$$

Moreover,

$$\mathbb{P}\left(\zeta = \infty \text{ and } \mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_* e^{-\mu_{H,L}(\theta)t} \text{ for all } t \geqslant 0\right) \geqslant 1 - \left(\frac{H_0}{H_*} \right)^\theta - \eta. \tag{140}$$

Proof. Let

$$E_H := \left\{ \sup_{0 \leqslant s \leqslant \tau_{H,L}} e^{\mu_{H,L}(\theta) s} H_s < H_* \right\}.$$

By Theorem E.9,

$$\mathbb{P}(E_H) \geqslant 1 - \left(\frac{H_0}{H_*} \right)^{\theta}.$$

On E_H, for all $t < \tau_{H,L}$,

$$H_t \leqslant H_* e^{-\mu_{H,L}(\theta) t}.$$

Using Lemma E.6, we get

$$\int_0^{t \wedge \tau_{H,L}} \mathcal{R}(\mathcal{X}_s)\, \mathrm{d}s \leqslant V_{H,L}(\theta).$$

For each coordinate define the martingale

$$N_{t,i} := \int_0^{t \wedge \tau_{H,L}} \sqrt{\mathcal{R}(\mathcal{X}_s)}\, \mathrm{d}\mathfrak{B}_{s,i}.$$

On E_H, its quadratic variation is bounded by $V_{H,L}(\theta)$. Therefore, by the reflection principle and a union bound over $i = 1, \ldots, d$,

$$\mathbb{P}\left(E_H \cap \left\{ \max_{1 \leqslant i \leqslant d} \sup_{t \geqslant 0} |N_{t,i}| > R_{H,L}(\theta, \eta) \right\} \right) \leqslant \eta.$$

Let E_N denote the complementary martingale event. On $E_H \cap E_N$, define

$$\widetilde{\mathcal{Z}}_{t,i} := \log\left(\frac{\mathcal{U}_{t,i}}{\mathcal{V}_{t,i}} \right).$$

Since

$$\mathcal{U}_{t,i} \mathcal{V}_{t,i} = \exp\left(-4\gamma^2 I_t \right),$$

we have, for $t < \tau_{H,L}$,

$$\mathcal{U}_{t,i} \mathcal{V}_{t,i} \geqslant p_{H,L}(\theta).$$

Moreover, Itô's formula gives

$$\mathrm{d}\widetilde{\mathcal{Z}}_{t,i} = \left(-4\gamma \mathcal{U}_{t,i} \mathcal{V}_{t,i} \sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma \right) \mathrm{d}t + 4\gamma \sqrt{\mathcal{R}(\mathcal{X}_t)}\, \mathrm{d}\mathfrak{B}_{t,i}.$$

Thus the noise-compensated process

$$\mathcal{Y}_{t,i} := \widetilde{\mathcal{Z}}_{t,i} - 4\gamma N_{t,i}$$

satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{Y}_{t,i} = -4\gamma \mathcal{U}_{t,i} \mathcal{V}_{t,i} \sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma.$$

A deterministic barrier argument gives, for every $t < \tau_{H,L}$,

$$|\widetilde{\mathcal{Z}}_{t,i}| \leqslant A_{H,L}(\theta) + 8\gamma R_{H,L}(\theta, \eta).$$

Consequently,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 = 2\mathcal{U}_{t,i} \mathcal{V}_{t,i} \cosh(\widetilde{\mathcal{Z}}_{t,i}) \leqslant B_{H,L}(\theta, \eta),$$

and, since

$$\mathcal{U}_{t,i} \mathcal{V}_{t,i} \geqslant p_{H,L}(\theta),$$

we also have

$$\mathcal{U}_{t,i}^2, \mathcal{V}_{t,i}^2 \geqslant c_{H,L}(\theta, \eta).$$

Moreover,

$$\frac{\rho(I_t) - 1}{2} \geqslant \underline{c}_{H,L}(\theta), \qquad \frac{\rho(I_t) + 1}{2} \leqslant \overline{c}.$$

Therefore, by the definition of $\mathfrak{L}_{H,L}(\theta, \eta)$, we have

$$h_{t,i} = f_{\frac{\rho(I_t)+1}{2}}(\mathcal{U}_{t,i}^2) + f_{\frac{\rho(I_t)-1}{2}}(\mathcal{V}_{t,i}^2) \leqslant \mathfrak{L}_{H,L}(\theta, \eta) < L_*.$$

Thus the coordinatewise entropy barrier cannot be hit on $E_H \cap E_N$. The entropy barrier $H_t = H_*$ cannot be hit either, because on E_H we have $H_t < H_* e^{-\mu_{H,L}(\theta)t} < H_*$ for all $t < \tau_{H,L}$. Hence $\tau_{H,L} = \infty$ on $E_H \cap E_N$. By Lemma E.7, this also implies $\zeta = \infty$.

The entropy decay estimate follows from the definition of E_H, and the risk decay follows from the upper bound in (128). The probability bound follows from

$$\mathbb{P}(E_H \cap E_N) \geqslant 1 - \left(\frac{H_0}{H_*}\right)^{\theta} - \eta.$$

□

E.5 High-probability Decay at Prescribed Confidence

The previous theorem gives an explicit confidence estimate. For fixed d, H_*, L_*, and γ, the admissible range of θ is finite. By choosing the stepsize sufficiently small, we can achieve a prescribed confidence level.

Corollary E.11 (High-probability decay for sufficiently small stepsize)**.** *Fix $H_*, L_* > H_0$, and $\delta \in (0,1)$. Define*

$$\theta_\delta := \frac{\log(2/\delta)}{\log(H_*/H_0)}$$

and

$$\bar{\gamma}_{H,L,\delta} := \frac{2}{2H_* + 2\sqrt{5}\log 2 + \sqrt{5} + \frac{2M_{L_*}^2}{d\, m_{L_*}}\theta_\delta}.$$

Assume

$$0 < \gamma < \bar{\gamma}_{H,L,\delta}.$$

Then

$$\mu_{H,L}(\delta) := m_{L_*}\gamma\Big(2 - \gamma(2H_* + 2\sqrt{5}\log 2 + \sqrt{5})\Big) - \frac{2\gamma^2 M_{L_*}^2}{d}\theta_\delta > 0.$$

Suppose further that the coordinatewise barrier is chosen so that

$$L_* > \mathfrak{L}_{H,L}\left(\theta_\delta, \frac{\delta}{2}\right), \tag{141}$$

where $\mathfrak{L}_{H,L}$ is defined in (137). Then

$$\mathbb{P}\left(\zeta = \infty \text{ and } \mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_* e^{-\mu_{H,L}(\delta)t} \text{ for all } t \geqslant 0\right) \geqslant 1 - \delta. \tag{142}$$

Proof. By construction,

$$\left(\frac{H_0}{H_*}\right)^{\theta_\delta} = \frac{\delta}{2}.$$

The condition $\gamma < \bar{\gamma}_{H,L,\delta}$ is exactly equivalent to

$$\theta_\delta < \frac{2\lambda_{H,L}}{\sigma_{H,L}^2},$$

where $\lambda_{H,L}$ and $\sigma_{H,L}^2$ are defined in (131)–(132). Moreover,

$$\mu_{H,L}(\delta) = \lambda_{H,L} - \frac{\theta_\delta}{2}\sigma_{H,L}^2.$$

Thus Lemma E.10 applies with $\theta = \theta_\delta$ and $\eta = \delta/2$. Substituting these values into (140) gives (142). □

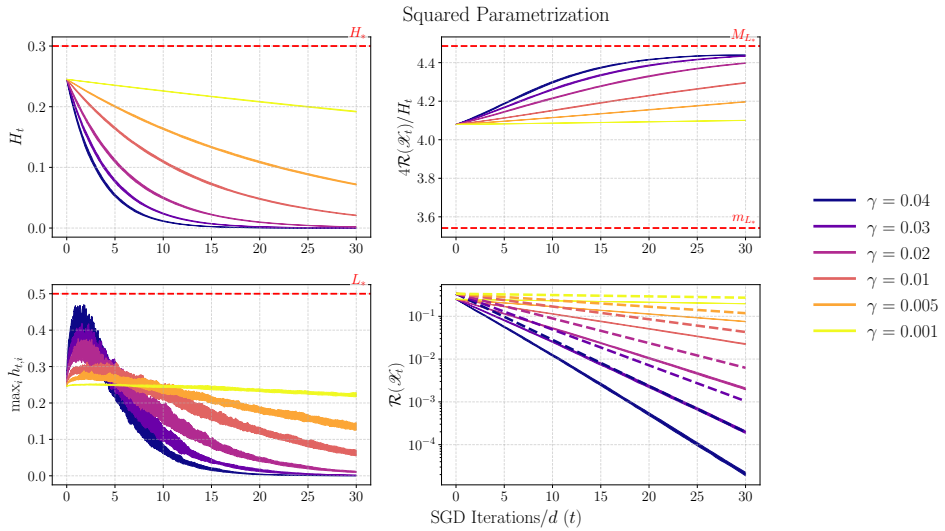

Figure 5: **Coordinatewise entropy barriers and exponential risk decay on a diagonal linear network.** The figure illustrates the entropy-barrier mechanism from Appendix E. The top-left panel shows the empirical entropy H_t, while the bottom-left panel shows the largest coordinatewise entropy density $\max_i h_{t,i}$; the red dashed lines mark the barriers H_* and L_*. The top-right panel plots the risk–entropy ratio $4\mathcal{R}(\mathcal{X}_t)/H_t$, with red dashed lines marking empirical coordinatewise coercivity constants m_{L_*} and M_{L_*} estimated from the coordinate quotients $q_{t,i}$. The bottom-right panel shows the risk $\mathcal{R}(\mathcal{X}_t)$ on a logarithmic scale, together with exponential envelopes of the form $(M_{L_*}/4)H_* e^{-\mu t}$. We plot these envelopes for all tested stepsizes, including those beyond the theorem's certified small-stepsize regime, illustrating that the predicted exponential decay persists empirically beyond the sufficient condition.

E.6 Risk Integrability and Dynamical Consequences

We now record two consequences of the high-probability exponential decay: integrability of the risk and uniform separation from the saddle.

Lemma E.12 (Finite integral of the risk)**.** *On the event in Corollary E.11,*

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s \leqslant \frac{M_{L_*}}{4}\frac{H_*}{\mu_{H,L}(\delta)} < \infty. \tag{143}$$

Proof. On the event in Corollary E.11,

$$\mathcal{R}(\mathcal{X}_t) \leqslant \frac{M_{L_*}}{4} H_* e^{-\mu_{H,L}(\delta)t} \qquad \text{for all } t \geqslant 0.$$

Therefore

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s \leqslant \frac{M_{L_*}}{4} H_* \int_0^\infty e^{-\mu_{H,L}(\delta)s}\,\mathrm{d}s = \frac{M_{L_*}}{4}\frac{H_*}{\mu_{H,L}(\delta)}.$$

□

E.6.1 Non-explosion via Risk Integrability

Although Corollary E.11 already includes $\zeta = \infty$, the following pathwise argument explains why risk integrability prevents finite-time explosion.

For each coordinate define

$$\widetilde{z}_{t,i} := \log\left(\frac{\mathcal{U}_{t,i}}{\mathcal{V}_{t,i}}\right), \qquad 0 \leqslant t < \zeta.$$

By Lemma E.1,

$$\mathcal{U}_{t,i}\mathcal{V}_{t,i} = \exp\left(-4\gamma^2 I_t\right).$$

Applying Itô's formula to $\widetilde{\mathcal{Z}}_{t,i} = \log \mathcal{U}_{t,i} - \log \mathcal{V}_{t,i}$ gives

$$\mathrm{d}\widetilde{\mathcal{Z}}_{t,i} = \Big(-4\gamma \mathcal{U}_{t,i}\mathcal{V}_{t,i} \sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma\Big)\,\mathrm{d}t + 4\gamma\sqrt{\mathcal{R}(\mathcal{X}_t)}\,\mathrm{d}\mathfrak{B}_{t,i}. \tag{144}$$

Lemma E.13 (Pathwise coordinate bound from risk integrability)**.** *Assume*

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s < \infty.$$

Then, for every coordinate i,

$$\sup_{0\leqslant t<\zeta} (\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2) < \infty \qquad \textit{almost surely}.$$

Consequently, finite risk integral is incompatible with finite-time explosion.

Proof. Define the continuous martingale

$$N_{t,i} := \int_0^t \sqrt{\mathcal{R}(\mathcal{X}_s)}\,\mathrm{d}\mathfrak{B}_{s,i}.$$

Its quadratic variation satisfies

$$\langle N_i\rangle_\infty = \int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s < \infty.$$

Hence $N_{t,i}$ converges almost surely as $t \to \infty$, and therefore

$$N_i^* := \sup_{t\geqslant 0} |N_{t,i}| < \infty \qquad \text{almost surely}.$$

Let

$$P_t := \mathcal{U}_{t,i}\mathcal{V}_{t,i} = \exp(-4\gamma^2 I_t).$$

Since $I_\infty < \infty$, there exists

$$p_\infty := \exp(-4\gamma^2 I_\infty) > 0$$

such that $P_t \geqslant p_\infty$ for all t. Define

$$\mathcal{Y}_{t,i} := \widetilde{\mathcal{Z}}_{t,i} - 4\gamma N_{t,i}.$$

By (144),

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{Y}_{t,i} = -4\gamma P_t \sinh(\widetilde{\mathcal{Z}}_{t,i}) + 2\gamma.$$

A deterministic barrier argument gives

$$|\widetilde{\mathcal{Z}}_{t,i}| \leqslant 8\gamma N_i^* + \operatorname{arcsinh}\left(\frac{1}{2p_\infty}\right) \qquad \text{for all } t.$$

Therefore,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 = 2P_t\cosh(\widetilde{\mathcal{Z}}_{t,i}) \leqslant 2\cosh\left(8\gamma N_i^* + \operatorname{arcsinh}\left(\frac{1}{2p_\infty}\right)\right) < \infty.$$

Thus all coordinates remain bounded on finite time intervals. Since the SDE coefficients are locally Lipschitz, the standard continuation criterion rules out finite-time explosion. □

E.6.2 Uniform Separation from the Saddle via Risk Integrability

We now show that risk integrability also prevents the dynamics from approaching the saddle at the origin.

Lemma E.14 (Uniform separation from the saddle)**.** *If*

$$\int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s < \infty,$$

then, for every coordinate i and all $t \geqslant 0$,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant m,$$

where

$$m := 2\exp\left(-4\gamma^2 \int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right) > 0.$$

Proof. By the arithmetic–geometric mean inequality,

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant 2\mathcal{U}_{t,i}\mathcal{V}_{t,i}.$$

Using Lemma E.1,

$$\mathcal{U}_{t,i}\mathcal{V}_{t,i} = \exp\left(-4\gamma^2 \int_0^t \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right).$$

Therefore

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant 2\exp\left(-4\gamma^2 \int_0^t \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right) \geqslant 2\exp\left(-4\gamma^2 \int_0^\infty \mathcal{R}(\mathcal{X}_s)\,\mathrm{d}s\right) =: m.$$

Since the integral is finite, $m > 0$. □

Combining Corollary E.11 with Lemma E.14, we obtain that, with probability at least $1-\delta$, the dynamics exist globally, the risk decays exponentially, and the coordinates remain uniformly separated from the saddle:

$$\mathcal{U}_{t,i}^2 + \mathcal{V}_{t,i}^2 \geqslant 2\exp\left(-\gamma^2 \frac{M_{L_*} H_*}{\mu_{H,L}(\delta)}\right) \qquad \text{for all } t \geqslant 0,\ i = 1, \dots, d.$$

F Examples

In this section, we provide further key learning problems within our family of models.

F.1 Mean-squared Error

One canonical example of (7) is the outer function $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by $f(r) = \frac{1}{2}(r_1 - r_2)^2$.

In this case, the problem is

$$\min_{x \in \mathbb{R}^{2d}} \left\{ \mathcal{R}(x) = \frac{1}{2d}\,\mathbb{E}_a \langle \psi(x) - \beta^*, a\rangle^2 = \frac{1}{2d}(\psi(x) - \beta^*)^\top K (\psi(x) - \beta^*) \right\}, \tag{145}$$

and satisfies

$$I\left(B(x)\right) = \mathbb{E}_a[\nabla_{r_1} f(r)^2] = 2\mathcal{R}(x) = 2h\left(B(x)\right) = B(x)_{11} - B(x)_{12} - B(x)_{21} + B(x)_{22}.$$

Moreover,

$$\frac{1}{d}\operatorname{Tr}\left(\nabla^2 \mathcal{R}(x)\right) = \frac{1}{d}\sum_{i=1}^d K_{ii}\left[\left(2q_{11}u_i + 2q_{12}v_i + l_1\right)^2 + \left(2q_{12}u_i + 2q_{22}v_i + l_2\right)^2\right]$$

$$+ \frac{2}{d}(q_{11} + q_{22}) \sum_{i=1}^d K_{ii}\left(\psi(x) - \beta^*\right)_i.$$

Remark F.1. Note that both $\mathcal{R}(x)$ and $\frac{1}{d}\operatorname{Tr}\left(\nabla^2 \mathcal{R}(x)\right)$ satisfy Assumption 3.6. Consequently, their SGD dynamics concentrate and can be analyzed via Theorem 3.7.

Example F.2. The simplest parametrization is to consider $\psi(x) = u$ which corresponds to the classical leas-squares framework,

$$\min_{u \in \mathbb{R}^d} \frac{1}{2d}\,\mathbb{E}_a \langle u - \beta^*, a\rangle^2. \tag{146}$$

Here, $\psi(u, v) = u$ for $\mathcal{Q} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, $l = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and $c = 0$.

Example F.3 (Diagonal Linear Network, Standard Parametrization)**.** This formulation arises in the problem of minimizing the expected squared risk over a standard depth-two diagonal linear neural network, represented by

$$\min_{u,v \in \mathbb{R}^d} \frac{1}{2d}\,\mathbb{E}_a \langle u \odot v - \beta^*, a\rangle^2. \tag{147}$$

Here, $\psi(u, v) = u \odot v$ for $\mathcal{Q} = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$, $l = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $c = 0$.

Example F.4 (Diagonal Linear Network, Squared Parametrization)**.** The most relevant example is the squared parametrization of the linear regression model. The optimization problem associated with this parametrization is formulated as

$$\min_{u,v\in\mathbb{R}^d} \frac{1}{4d}\,\mathbb{E}_a\langle u^2 - v^2 - \beta^*, a\rangle^2. \tag{148}$$

Here, $\psi(u,v) = u^2 - v^2$ for $\mathcal{Q} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$, $l = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $c = 0$.

F.2 Binary Logistic Regression

We consider the setting of (7) with outer function $f\colon\ \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(r) = -r_1 \cdot \frac{\exp(r_2)}{\exp(r_2)+1} + \log(\exp(r_1)+1).$$

In this case, the problem takes the form

$$\min_{x\in\mathbb{R}^{2d}} \left\{ \mathcal{R}(x) = -\,\mathbb{E}_a\langle \psi(x), a\rangle/\sqrt{d} \cdot \frac{\exp(\langle\beta^*, a\rangle/\sqrt{d})}{\exp(\langle\beta^*, a\rangle/\sqrt{d})+1} + \log(\exp(\langle\psi(x), a\rangle/\sqrt{d})+1) \right\},$$

and, using a multivariate normal distribution as in Collins-Woodfin et al. [2024a], satisfies

$$\mathcal{R}(x) = h\left(B(x)\right) = \frac{1}{\pi}\int_{\mathbb{R}^2} \left(-\tilde{u}\cdot\frac{\exp(\tilde{v})}{\exp(\tilde{v})+1} + \log(\exp(\tilde{u})+1)\right) \exp\left(-\begin{pmatrix} u \\ v \end{pmatrix}^\top \begin{pmatrix} u \\ v \end{pmatrix}\right) \mathrm{d}u\mathrm{d}v$$

where $\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = \sqrt{2}L \begin{pmatrix} u \\ v \end{pmatrix}$ and $B(x) = LL^\top$.

Analogously, since

$$\nabla_{r_1} f(r) = -\frac{\exp(r_2)}{\exp(r_2)+1} + \frac{\exp(r_1)}{\exp(r_1)+1},$$

we obtain

$$I\left(B(x)\right) = \mathbb{E}_a[\nabla_{r_1} f(r)^2] = \frac{1}{\pi}\int_{\mathbb{R}^2} \left(-\frac{\exp(\tilde{v})}{\exp(\tilde{v})+1} + \frac{\exp(\tilde{u})}{\exp(\tilde{u})+1}\right)^2 \exp\left(-\begin{pmatrix} u \\ v \end{pmatrix}^\top \begin{pmatrix} u \\ v \end{pmatrix}\right) \mathrm{d}u\mathrm{d}v.$$

G Numerical Simulation Details

Here we provide more details for the figures that appear in the main paper.

Figure 1: Three views of empirical risk dynamics for SGD on a diagonal linear network: concentration, spectral effects, and regime transitions. *Left:* Identity covariance $K = I_d$. As the dimension d increases, the risk trajectory of SGD concentrates around a deterministic limit (shown in red), as characterized in Theorem 3.7. *Middle:* Power-law covariance spectrum. Our homogenized SGD prediction (transparent) from Theorem 3.7 closely tracks SGD (opaque) over a range of power-law exponents β, at fixed dimension $d = 10^3$. *Right:* Identity covariance $K = I_d$ at fixed dimension $d = 10^3$. Varying the stepsize γ reveals distinct convergence/divergence regimes; the homogenized prediction remains accurate even for stepsizes above the convergence threshold.

The left and right panels use the squared parametrization (Example F.4), whereas the middle panel uses the standard parametrization (Example F.3). For the power-law spectrum, eigenvalues are generated via inverse-CDF sampling: draw $u_i \sim \mathcal{U}(0,1)$ and set $\lambda_i = u_i^{1/(1-\beta)}$, yielding i.i.d. $\lambda_i \in [0,1]$ with density $p(\lambda) = (1-\beta)\lambda^{-\beta}$. We set $\beta^* = \mathbb{1}_d$ and initialize SGD at $u_0 = v_0 = \alpha\,\mathbb{1}_d$ with $\alpha = 0.6$ (left and right), and at $u_0 = 2\,\mathbb{1}_d$, $v_0 = 0$ (middle). Curves are aggregated over 30 independent stochastic runs. The stepsize is fixed to $\gamma = 0.1$ (left) and $\gamma = 1.1$ (middle); in the right panel, γ is swept across the displayed range at fixed $d = 10^3$.

Numerical construction of the theoretical curve: The contour-integral evaluation used to recover the theory curve is very sensitive to the choice of contour radius. Under symmetric initialization ($u_0 = v_0$), a single contour $C_M(0)$ suffices; we empirically verify that $\|x_k\|_\infty < M$ along the trajectory and use $M = 1.3$ with $N = 100$ contour discretization points. If M is too small, the condition $\|x_k\|_\infty < M$ is violated, leading to incorrect contour integrals and a mismatch (notably at initialization) between SGD and the PDE prediction; if M is too large, the PDE solver may overflow. Accordingly, M must be calibrated for each initialization.

Figure 2: Curvature dynamics for SGD on a diagonal linear network. *Left:* The evolution of the curvature measured by the scaled trace of the Hessian $\frac{1}{d}\operatorname{Tr}(\nabla^2\mathcal{R})$ is shown alongside the empirical risk $\mathcal{R}$, illustrating "flat" progress in which the risk increases sharply accompanied by a marked drop in curvature as we vary the stepsize γ. *Right:* As the dimension d increases, the curvature dynamics of SGD concentrate around a deterministic limit (shown in red), as proven in Theorem 3.7. The left panel corresponds to the squared parametrization (Example F.4), while the right one corresponds to the standard parametrization (Example F.3).

Note that in the standard parametrization, the scaled Hessian trace simplifies to $\operatorname{Tr}(\nabla^2\mathcal{R}(x)) = \frac{1}{d}\sum_{i=1}^{d} K_{ii}(u_i^2 + v_i^2)$. We set $\beta^* = \mathbb{1}_d$ and $K = I_d$, and initialize SGD at $u_0 = v_0 = \alpha\,\mathbb{1}_d$ with $\alpha = 1.0$ (left) and $\alpha = 0.6$ (right). Curves are aggregated over 30 independent stochastic runs. *Left:* fixed $d = 10^3$ with a sweep over γ. *Right:* fixed $\gamma = 0.1$ with varying d. We use $M = 1.0$ and $N = 300$ for the contour construction.

Figure 3: Risk concentration of SGD and the homogenized SDE under non-diagonal covariance on a diagonal linear network. As the dimension d increases, the risk trajectories of SGD (opaque) concentrate around the prediction of the non-diagonal homogenized SDE (14) (transparent), suggesting that the same high-dimensional concentration phenomenon persists beyond the diagonal covariance setting. The covariance matrix K is sampled from a Marchenko–Pastur ensemble.

The left panel uses the standard parametrization (Example F.3) and is initialized at $u_0 = 2 \cdot \mathbb{1}_d$, while the right panel uses the squared parametrization (Example F.4) with initialization $u_0 = v_0 = \alpha\,\mathbb{1}_d$ and $\alpha = 0.6$. In both panels, we set $\beta^\star = \mathbb{1}_d$ and fix the stepsize to $\gamma = 0.2$. Curves are aggregated over 30 independent stochastic runs. The covariance matrix K is generated by drawing a Gaussian random matrix $X \in \mathbb{R}^{d\times d}$ with i.i.d. standard normal entries and setting $K = \frac{1}{d}XX^\top$.

Figure 4: Risk discrepancy between SGD and its continuous-time approximations on a diagonal linear network. For each stepsize γ, we report the absolute difference between the empirical risk of SGD after $T \cdot d$ iterations (with $T = 20$) and two approximations: (i) homogenized SGD (HSGD) (13) (blue), and (ii) stochastic gradient flow (SGF) (16) (pink). As γ increases, the proposed HSGD—which captures large stepsize effects—provides a progressively more accurate approximation of SGD, whereas SGF is derived under a vanishing stepsize regime and thus degrades for larger γ. Initialization scale α controls proximity to the saddle point $x = 0$: smaller α corresponds to a longer transient before learning accelerates.

We consider the squared parametrization (Example F.4) with mean-squared error, in the sparse recovery setting of Pesme et al. [2021]. The ground-truth signal satisfies $\beta_i^* = 1$ for $1 \leqslant i \leqslant 5$ and $\beta_i^* = 0$ for $i > 5$, with covariance $K = I_d$ and dimension $d = 100$. SGD is initialized at $u_0 = v_0 = \alpha\,\mathbb{1}_d$, with $\alpha = 0.01$ (left) and $\alpha = 0.1$ (right), and results are aggregated over 30 independent stochastic runs. Both HSGD and SGF trajectories are simulated using the Euler–Maruyama method with time step $\mathrm{d}t = 2^{-8}$.

Figure 5: Coordinatewise entropy barriers and exponential risk decay on a diagonal linear network. The figure illustrates the entropy-barrier mechanism developed in Appendix E. The left panels track the two quantities that define the controlled regime of the proof: the empirical entropy H_t and the largest coordinatewise entropy density $\max_i h_{t,i}$. The red dashed lines mark the corresponding barriers H_* and L_*. Across the tested stepsizes, the trajectories remain below these barriers, indicating that the dynamics stay in the regime where the coordinatewise entropy-barrier argument applies.

The top-right panel displays the risk–entropy ratio $4\mathcal{R}(\mathfrak{X}_t)/H_t$. The red dashed lines mark empirical estimates of the coordinatewise coercivity constants m_{L_*} and M_{L_*}. These constants are estimated

from the coordinate quotients

$$q_{t,i} = \frac{(\mathcal{U}_{t,i}^2 - \mathcal{V}_{t,i}^2 - 1)^2}{h_{t,i}},$$

rather than directly from the averaged ratio $4\mathcal{R}(\mathcal{X}_t)/H_t$. Specifically, we pool the lower and upper coordinate-quotient summaries across all stepsizes and stochastic runs, restricted to times satisfying $H_{\min} \leqslant H_t \leqslant H_*$ and $\max_i h_{t,i} \leqslant L_*$, and then take robust quantiles to estimate m_{L_*} and M_{L_*}. The panel visualizes the theorem's risk–entropy comparison: while the coordinatewise barrier holds, the averaged ratio $4\mathcal{R}(\mathcal{X}_t)/H_t$ should remain controlled by the same coercivity constants.

The bottom-right panel shows the risk $\mathcal{R}(\mathcal{X}_t)$ on a logarithmic scale. The dashed curves are the exponential envelopes suggested by the theorem, of the form

$$\frac{M_{L_*}}{4} H_* e^{-\mu t},$$

using the empirical constants m_{L_*} and M_{L_*} from the coordinatewise quotient estimates. In the theorem, this bound is guaranteed under a sufficient small-stepsize condition ensuring a positive decay rate. In the experiment, we plot the same exponential envelope for all tested stepsizes, rather than suppressing curves outside the certified small-stepsize range. This highlights that the predicted exponential decay behavior appears to persist empirically even for stepsizes larger than those directly covered by the sufficient condition.

Shaded regions denote the 10th–90th percentile range over independent runs, and solid curves denote the median trajectory. We simulate the isotropic squared-parametrization setting with $\beta^* = \mathbb{1}_d$, covariance $K = I_d$, Gaussian covariates $a \sim \mathcal{N}(0, I_d)$, and initialization $u_0 = v_0 = \mathbb{1}_d$. The dimension is $d = 10^3$, the time horizon is $T = 30$, and results are aggregated over 30 independent stochastic runs for each constant stepsize $\gamma \in \{0.04, 0.03, 0.02, 0.01, 0.005, 0.001\}$. The horizontal axis is the rescaled time $t = k/d$, i.e., SGD iterations divided by the dimension.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the problem class, contributions, and scope. We also state the main limitations, namely the high-dimensional diagonal linear network setting with Gaussian data, in Section 1.2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We state the model assumptions in Section 1.2 and algorithm assumptions in Section 1.3. We clearly indicate when results specialize to the squared parametrization; see, e.g., Theorem 3.9. We discuss limitations from the Gaussian data assumption in Remarks 1.6 and 1.7. We also note that non-explosiveness holds in our numerics, and in Remark 3.2 we point to Appendix E, where it is proved in the isotropic squared-parametrization setting for constant stepsize γ.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Sections 1.2 and 1.3 specify the model and algorithmic assumptions. Complete proofs of all main results are provided in the appendix, and each theorem statement includes the assumptions required for the result. We also provide an intuition-of-the-proof paragraph following our main result, Theorem 3.7.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information needed to reproduce the experiments is provided in the figure captions and Section G, including parameters, data covariances, and initialization. The simulations use synthetic data and standard algorithms. We also plan to release the code on GitHub.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All details needed to reproduce the experiments are provided in the figure captions and Section G. The data are synthetic. We will release the code with instructions to reproduce the results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All relevant details (parameters, data covariance, initialization, etc.) are provided in the figure captions and Section G. The experiments use synthetic data and standard algorithms. We also plan to release the code on GitHub.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figures 1 and 2 include error bars (shaded regions) computed across multiple runs. These illustrate the concentration predicted by our theory and the agreement with the deterministic predictions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments use synthetic data and simple mean-squared error models, requiring minimal compute. No specialized hardware is needed, and all details to run the experiments are provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and made our utmost attempt to adhere to its guidelines. Our work uses only synthetic data and standard algorithms, with no human subjects or sensitive data. We cite relevant prior work and, to the best of our knowledge, do not anticipate ethical risks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is purely theoretical and not tied to specific applications. It uses only synthetic data, standard algorithms (SGD), and classical models (e.g., least squares). We therefore do not anticipate direct societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable. The work is purely theoretical and uses only synthetic Gaussian data. It does not release models or datasets with misuse risk; the models considered (e.g., least squares) are standard.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite prior work introducing the algorithms we study. We do not use external datasets or third-party assets; our experiments use synthetic data and standard libraries (e.g., NumPy).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: No new assets are introduced. The models and algorithms studied are standard. We will release code for reproducibility, including generation of the deterministic PDEs.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Not applicable. The paper does not involve crowdsourcing or human subjects; it is purely theoretical.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable. The work involves no human subjects and uses only synthetic data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Not applicable. LLMs were used only for minor editing (e.g., grammar and wording) and did not affect the methodology or results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.