SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality

Courtney Paquette*† Kiwon Lee† Fabian Pedregosa* Elliot Paquette†

February 9, 2021

Abstract

We propose a new framework, inspired by random matrix theory, for analyzing the dynamics of stochastic gradient descent (SGD) when both number of samples and dimensions are large. This framework applies to any fixed stepsize and the finite sum setting. Using this new framework, we show that the dynamics of SGD on a least squares problem with random data become deterministic in the large sample and dimensional limit. Furthermore, the limiting dynamics are governed by a Volterra integral equation. This model predicts that SGD undergoes a phase transition at an explicitly given critical stepsize that ultimately affects its convergence rate, which we also verify experimentally. Finally, when input data is isotropic, we provide explicit expressions for the dynamics and average-case convergence rates (*i.e.,* the complexity of an algorithm averaged over all possible inputs). These rates show significant improvement over the worst-case complexities.

1 Introduction

Stochastic gradient descent (SGD) Robbins and Monro [1951] is one of the most popular and important stochastic optimization methods for use in large-scale problems. There are well-established worst-case convergence rates, but SGD lacks a detailed theory that encompasses both its successes and its empirically observed peculiarities. For example, the solutions to which SGD converges have qualitative differences that seemingly depend on how SGD is tuned [Jastrzebski et al., 2017, Keskar et al., 2016]. Furthermore, the dependence of the runtime of SGD on its stepsize is complicated, and stepsize selection is an active area of research [Bollapragada et al., 2018, Friedlander and Schmidt, 2012, Mahsereci and Hennig, 2017, Schaul et al., 2013, Vaswani et al., 2019]. Beyond the confines of SGD, the behavior of other stochastic optimization algorithms is even more poorly understood [Sutskever et al., 2013]. Because of these challenges, *making good quantitative predictions for the dynamics of stochastic algorithms remains a difficult, broad and deep problem.*

A prolific technique for analyzing optimizations methods, both stochastic and deterministic, is the stochastic differential equations (SDE) paradigm [Chaudhari and Soatto, 2018, Hu et al., 2017, Jastrzebski et al., 2017, Kushner and Yin, 2003, Li et al., 2017, Ljung, 1977, Mandt et al., 2016, Su et al., 2016]. These SDEs relate to the dynamics of the optimization method by taking the limit when the stepsize goes to zero, so that the trajectory of the objective function over the lifetime of the algorithm converges to the solution of an SDE. Naturally, in practice, the stepsize is taken as large as possible, which limits the predictive power of the SDE method.

A related popular paradigm for analyzing the behavior of SGD is the noisy gradient model. Often used in conjunction with the SDE approach, in this model, one supposes that the stochastic gradient estimators in SGD are the true gradient plus some independent noise (typically assumed to be Gaussian with some covariance structure) [Jastrzebski et al., 2017, Li et al., 2017, Mandt et al., 2016, Simsekli et al., 2019] or more generally the gradient estimators are independent with a common distribution, see for *e.g.* Huang et al. [2020]. The latter is equivalent to the "streaming setting" Jain et al. [2018] or the "one-pass" assumption on the data [Gurbuzbalaban et al., 2020].

*Google Research, Brain Team

†Department of Mathematics and Statistics, McGill University, Montreal, QC, Canada, H3A 0B9; CP is a CIFAR AI chair; https://cypaquette.github.io/. Research by EP was supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada.; https://elliotpaquette.github.io/.

Here one generates a new sample at each iteration and does not reuse any past data. In practice, SGD is typically implemented on a finite dataset with multiple passes over the data. Such modeling assumptions on the stochastic gradient estimators can not capture the full dynamics of SGD (see Figure 1).

We offer a new alternative, inspired by the phenomenology of random matrix theory. We prove that SGD with a *fixed* stepsize γ has deterministic dynamics, when run on the least squares problem with *high–dimensional* random data, and, we analyze these dynamics to provide stepsize selection and convergence properties (see Figure 1 for a comparison). We neither impose assumptions on the gradient estimators nor take the stepsize to 0 and we work in the *non-streaming* or *finite sum* setting (a.k.a. incremental gradient). These deterministic dynamics are governed by a Volterra integral equation, that is, the function values converge to the solution ψ_0 of

$$\psi_0(t) = z(t) + r\gamma^2 \int_0^t h_2(t-s)\psi_0(s)\ \mathrm{d}s, \qquad \text{where} \quad h_2(t) = \int_0^\infty x^2 e^{-2\gamma t x}\ \mathrm{d}\mu(x). \tag{1}$$

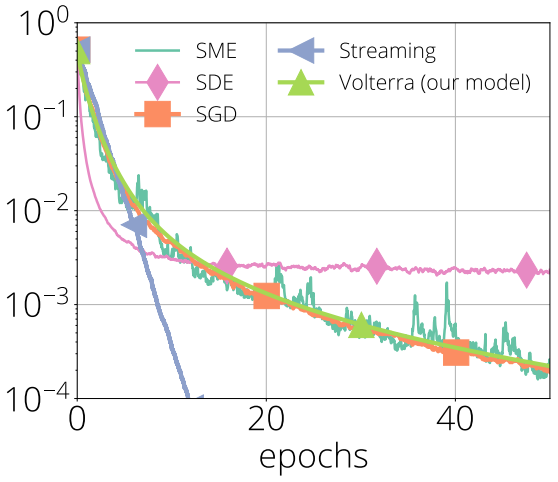

Figure 1: The proposed Volterra equation model **accurately** tracks SGD on a random least-squares problem for any choice of stepsize. Other models introduce biases that substantially impact model fidelity.

Here, r is the ratio of the number of parameters to sample size, and μ is the distribution of the eigenvalues of the Hessian's objective. The function z is an explicit forcing function, which has dependence on all parts of the problem, including the initialization $\boldsymbol{x}_0$ and the target $\boldsymbol{b}$. See Theorem 1.1 for the precise statement. The value of the stepsize γ can be as large as the convergence threshold which we explicitly provide. This Volterra equation (1) has rich behavior; the asymptotic suboptimality of ψ_0 has a discontinuity in γ at a critical stepsize (see Theorem 1.2).

Notation. We denote vectors in lowercase boldface ($\boldsymbol{x}$) and matrices in uppercase boldface ($\boldsymbol{H}$). A sequence of random variables $\{y_d\}_{d=0}^\infty$ converges in probability to y, indicated by $y_d \xrightarrow[d\to\infty]{\Pr} y$, if for any $\varepsilon > 0$, $\lim_{d\to\infty} \Pr(|y_d - y| > \varepsilon) = 0$. Probability measures are denoted by μ and their densities by $\mathrm{d}\mu$. We say a sequence of random measures μ_d converges to μ weakly in probability if for any bounded continuous function f, we have $\int f\,\mathrm{d}\mu_d \to \int f\,\mathrm{d}\mu$ in probability. For two random variables x and y we write $x \overset{\text{law}}{=} y$ to mean they have the same distribution.

1.1 Problem setting.

We consider the least–squares problem when the number of samples (n) and features (d) are large:

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) \overset{\text{def}}{=} \frac{1}{2n}\sum_{i=1}^n (\boldsymbol{a}_i\boldsymbol{x} - b_i)^2 \right\}, \quad \text{with } \boldsymbol{b} \overset{\text{def}}{=} \boldsymbol{A}\widetilde{\boldsymbol{x}} + \sqrt{n}\,\boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is a random data matrix whose i-th row is denoted by $\boldsymbol{a}_i \in \mathbb{R}^{d\times 1}$, $\widetilde{\boldsymbol{x}} \in \mathbb{R}^d$ is the signal vector, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is a source of noise. The target $\boldsymbol{b} = \boldsymbol{A}\widetilde{\boldsymbol{x}} + \sqrt{n}\,\boldsymbol{\eta}$ comes from a generative model corrupted by noise.

We apply SGD (incremental gradient) to the finite sum, quadratic problem above. On the k-th iteration it selects a uniformly random subset $B_k \subset \{1, 2, \cdots, n\}$, of batch-size β and makes the updates

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\gamma}{n}\sum_{i\in B_k} \nabla f_i(\boldsymbol{x}_k) = \boldsymbol{x}_k - \frac{\gamma}{n}\boldsymbol{A}^T\boldsymbol{P}_k(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}), \quad \text{where} \quad \boldsymbol{P}_k \overset{\text{def}}{=} \sum_{i\in B_k} \mathbb{e}_i\mathbb{e}_i^T. \tag{3}$$

Here $\boldsymbol{P}_k$ is a random orthogonal projection matrix with $\mathbb{e}_i$ the i-th standard basis vector, $\beta \in \mathbb{N}$ is a batch-size parameter, which we will allow to depend on n, $\gamma > 0$ is a stepsize parameter, and the function f_i is the i-th element of the sum in (2). Typical stepsizes for SGD (see *e.g* [Bottou et al., 2018, Thm 4.6]) include the second moment of the stochastic gradients, which under our problem setting grows like n. This explains the dependency on n in (3). We remark that β can equal 1 in which case (3) reduces to the simple SGD setting.

To perform our analysis we make the following explicit assumptions on the signal $\widetilde{\boldsymbol{x}}$, the noise $\boldsymbol{\eta}$, and the data matrix $\boldsymbol{A}$.

Assumption 1.1 (Initialization, signal, and noise)**.** *The initial vector $\boldsymbol{x}_0 \in \mathbb{R}^d$, the signal $\widetilde{\boldsymbol{x}} \in \mathbb{R}^d$, and noise vector $\boldsymbol{\eta} \in \mathbb{R}^n$ satisfy the following conditions:*

1. *The difference $\boldsymbol{x}_0 - \widetilde{\boldsymbol{x}}$ is any deterministic vector such that $\|\boldsymbol{x}_0 - \widetilde{\boldsymbol{x}}\|_2^2 = R$.*
2. *The entries of the noise vector $\boldsymbol{\eta}$ are i.i.d. random variables that verify for some constant $\widetilde{R} > 0$*

$$\mathbb{E}\left[\boldsymbol{\eta}\right] = \boldsymbol{0}, \quad \mathbb{E}\left[\|\boldsymbol{\eta}\|_2^2\right] = \widetilde{R}, \quad \textit{and} \quad \mathbb{E}\left[\|\boldsymbol{\eta}\|_\infty^p\right] = \mathcal{O}(n^{\epsilon - p/2}) \quad \textit{for any } \epsilon, p > 0. \tag{4}$$

Any subexponential law for the entries of $\boldsymbol{\eta}$ (say, uniform or Gaussian with variance $\widetilde{R}/n$) will satisfy (4). The scalings of the vectors $\boldsymbol{x}_0 - \widetilde{\boldsymbol{x}}$ and $\boldsymbol{\eta}$ arise as a result of preserving a constant signal-to-noise ratio in the generative model. Such generative models with this scaling have been used in numerous works [Gerbelot et al., 2020, Hastie et al., 2019, Mei and Montanari, 2019].

Next we state an assumption on the eigenvalue and eigenvector distribution of the data matrix $\boldsymbol{A}$. We then review practical scenarios in which this is verified.

Assumption 1.2 (Data matrix)**.** *Let $\boldsymbol{A}$ be a random $n \times d$ matrix such that the number of features, d, tends to infinity proportionally to the size of the data set, n, so that $\frac{d}{n} \to r \in (0, \infty)$. Let $\boldsymbol{H} \overset{def}{=} \frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}$ with eigenvalues $\lambda_1 \leq \ldots \leq \lambda_d$ and let δ_{λ_i} denote the Dirac delta with mass at λ_i. We make the following assumptions on the eigenvalue distribution of this matrix:*

1. *The eigenvalue distribution of $\boldsymbol{H}$ converges to a deterministic limit μ with compact support. Formally, the empirical spectral measure (ESM) satisfies*

$$\mu_{\boldsymbol{H}} = \frac{1}{d}\sum_{i=1}^{d} \delta_{\lambda_i} \to \mu \quad \textit{weakly in probability}\,. \tag{5}$$

2. *The largest eigenvalue $\lambda_{\boldsymbol{H}}^+$ of $\boldsymbol{H}$ converges in probability to the largest element λ^+ in the support of μ, i.e.*

$$\lambda_{\boldsymbol{H}}^+ \xrightarrow[d\to\infty]{\Pr} \lambda^+. \tag{6}$$

3. *(Orthogonal invariance) Let $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{O} \in \mathbb{R}^{n\times n}$ be orthogonal matrices. The matrix $\boldsymbol{A}$ is orthogonally invariant in the sense that*

$$\boldsymbol{A}\boldsymbol{U} \overset{law}{=} \boldsymbol{A} \quad \textit{and} \quad \boldsymbol{O}\boldsymbol{A} \overset{law}{=} \boldsymbol{A} \tag{7}$$

Assumption 1.2 characterizes the distribution of eigenvalues for the random matrix $\boldsymbol{H}$ which approximately equals the distribution μ. The ESM and its convergence to the limiting spectral distribution μ are well studied in random matrix theory, and for many random matrix ensembles the limiting spectral distribution is known. In machine learning literature, it has been shown that the spectrum of the Hessians of neural networks share many characteristics with the limiting spectral distributions found in classical random matrix theory [Behrooz et al., 2019, Dauphin et al., 2014, Martin and Mahoney, 2018, Papyan, 2018, Sagun et al., 2016].

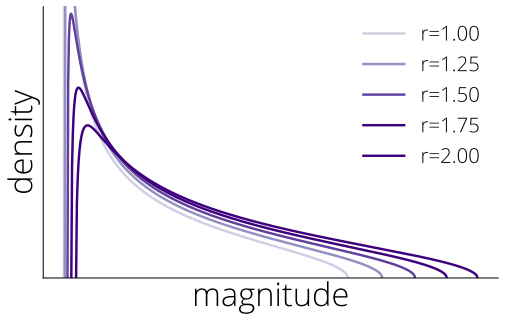

Figure 2: The ESM of matrices $\frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}$ with i.i.d. entries converges as $n, d \to \infty$ to the *Marchenko-Pastur* distribution, shown here for different values of $r = d/n$.

The last assumption, orthogonal invariance, is a rather strong condition as it implies that the singular vectors of $\boldsymbol{A}$ are uniformly distributed on the sphere. The classic example of a matrix which satisfies this property are matrices whose entries are generated from standard Gaussians. There is however, a large body of literature [Cipolloni et al., 2020, Knowles and Yin, 2017] showing that other classes of large dimensional random matrices behave like orthogonally

invariant ensembles; weakening the orthogonal invariance assumption is an interesting future direction of research which is beyond the scope of this paper. Moreover our numerical simulations suggest that (7) is unnecessary as our Volterra equation holds for ensembles without this orthogonal invariance property (see one-hidden layer networks in Section 4). For a more thorough review of random matrix theory see [Bai and Silverstein, 2010, Tao, 2012].

Examples of data distributions. In this section we review examples of data-generating distributions that verify Assumption 1.2.

Example 1: Isotropic features with Gaussian entries. The first model we consider has entries of $\boldsymbol{A}$ which are i.i.d. standard Gaussian random variables, that is, $A_{ij} \sim N(0,1)$ for all i, j. This ensemble has a rich history in random matrix theory. When the number of features d tends to infinity proportionally to the size of the data set n, $\frac{d}{n} \to r \in (0, \infty)$, the seminal work of Marčenko and Pastur [1967] showed that the spectrum of $\boldsymbol{H} = \frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}$ asymptotically approaches a deterministic measure μ_{MP}, verifying Assumption 1.2. This measure, μ_{MP}, is given by the Marchenko-Pastur law:

$$\mathrm{d}\mu_{\mathrm{MP}}(\lambda) \stackrel{\text{def}}{=} \delta_0(\lambda)\max\{1-\tfrac{1}{r}, 0\} + \frac{\sqrt{(\lambda-\lambda^-)(\lambda^+-\lambda)}}{2\pi\lambda r} 1_{[\lambda^-,\lambda^+]}\,, \qquad \text{where} \quad \lambda^- \stackrel{\text{def}}{=} (1-\sqrt{r})^2 \quad \text{and} \quad \lambda^+ \stackrel{\text{def}}{=} (1+\sqrt{r})^2\,. \tag{8}$$

Example 2: Planted spectrum One may wonder if there are limits to what singular value distributions can appear for orthogonally invariant random matrices, but as it turns out, any singular value distribution is attainable. Suppose that

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \tag{9}$$

where $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ are random matrices, uniformly chosen from the orthogonal group and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times d}$ is any deterministic matrix such that the squared singular values of $\boldsymbol{\Sigma}$ have an empirical distribution that converges to a desired limit μ. Then $\boldsymbol{A}$ is orthogonally invariant. As in the previous case, we assume that the dimensions of the matrix $\boldsymbol{A}$ grow at a comparable rate given by $\frac{d}{n} \to r \in (0, \infty)$. Constructions like this appear in neural networks initialized with random orthogonal weight matrices, and they produce exotic singular value distributions [Saxe et al., 2013, Figure 7].

Example 3: Linear neural networks. This model encompasses linear neural networks with a squared loss, where the m layers have random weights ($\boldsymbol{W}_i$ with $i = 1, \ldots, m$) and the final layer's weights are given by the regression coefficient $\boldsymbol{x}$. The entries of these random weight matrices are sampled i.i.d. from a standard Gaussian. The optimization problem in (2) becomes

$$\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) = \frac{1}{2n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \right\}, \quad \text{where } \boldsymbol{A} = \boldsymbol{W}_1\boldsymbol{W}_2\boldsymbol{W}_3\cdots\boldsymbol{W}_m. \tag{10}$$

It is known that products of Gaussian matrices satisfy (7) with a limiting spectral measure in the large n limit and fixed depth given by the Fuss-Catalan law Alexeev et al. [2010], Liu et al. [2011].

1.2 Main contributions

A new paradigm for analyzing the dynamics of SGD. We propose a framework for the analysis of SGD that exploits the fact that when increasing the problem size (*i.e.* n and d large), statistics that are driven by the full population converge to deterministic processes; the spirit of which is behind law of large numbers and concentration of measure. A practical outcome of this framework is a new expression for the function values of SGD as a Volterra equation:

Theorem 1.1 (Concentration of SGD). *Suppose the stepsize satisfies* $\gamma < \frac{2}{r}\big(\int_0^\infty x\,\mathrm{d}\mu(x)\big)^{-1}$ *and the batchsize satisfies* $\beta(n) \le n^{1/5-\delta}$ *for some* $\delta > 0$. *Under Assumptions 1.1 and 1.2,*

$$\sup_{0\le t\le T} \left| f\big(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}\big) - \psi_0(t) \right| \xrightarrow[n\to\infty]{\Pr} 0\,, \tag{11}$$

where the function ψ_0 is the solution to the Volterra equation

$$\psi_0(t) = \tfrac{R}{2} h_1(t) + \tfrac{\widetilde{R}}{2}\big(r h_0(t) + (1-r)\big) + \int_0^t \gamma^2 r h_2(t-s)\psi_0(s)\,\mathrm{d}s,$$
$$\text{and} \quad h_k(t) = \int_0^\infty x^k e^{-2\gamma t x}\,\mathrm{d}\mu(x) \quad \text{for all} \quad k \geq 0. \tag{12}$$

The expression highlights how the algorithm, stepsize, signal and noise levels interact with each other to produce different dynamics. For instance, our framework allows one to see the effect of the *entire* spectrum of the data matrix on the dynamics. Also we note that the batch-size β does not appear in the limiting Volterra equation. Numerical simulations in Section 4 confirm that ψ accurately predicts the behavior of SGD.

Phase transition of SGD dynamics and critical stepsize. We prove a surprising dichotomy in the dynamics of SGD for a general measure: SGD undergoes a phase transition at a critical stepsize which we denote by γ_*

$$\gamma_* \stackrel{\text{def}}{=} \frac{1}{\frac{r}{2}\int_0^\infty \frac{x^2}{x-\lambda^-}\,\mathrm{d}\mu(x)}\,. \tag{13}$$

Starting at small stepsizes, we see that the linear rate of convergence for SGD *freezes* on the smallest eigenvalue of $\boldsymbol{H}$, that is $f(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor})$ decreases like $e^{-2\gamma\lambda^- t}$. However when γ passes the transition point γ_*, the dynamics of SGD have a more complicated dependency on the stepsize (in particular it is no longer log-linear in γ). This is strongly reminiscent of a *freezing transition*, often seen in the free energies of random energy models (see Derrida [1981]), with γ playing the role of temperature. This is summarized in our second main contribution – the asymptotic rates for SGD under a general spectral measure μ (see Appendix E.1).

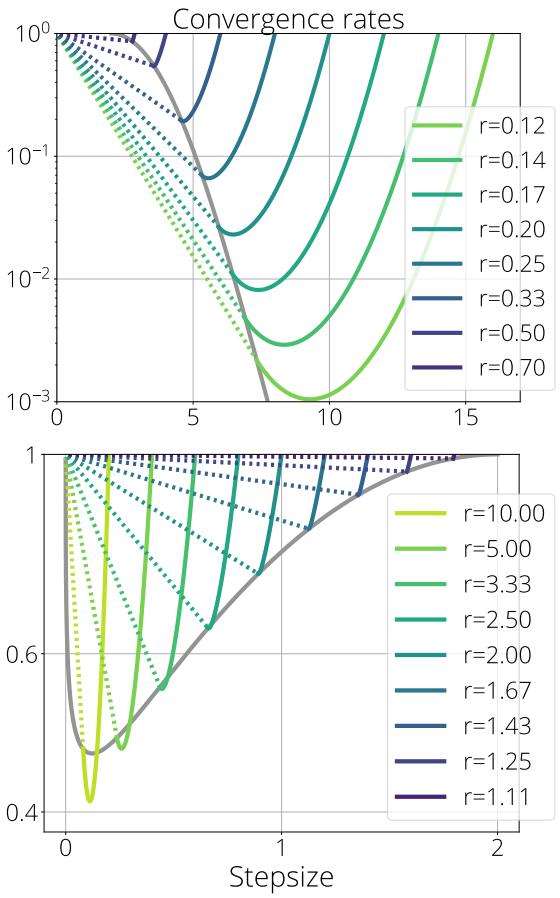

Figure 3: **Phase transition of the convergence rate** (y-axis) as a function of the stepsize (x-axis, γ) for the isotropic features model. Smaller stepsizes (dotted) yield convergence rates which depend linearly on γ with a slope that is always frozen on λ^-. The convergence rate abruptly changes behavior once it hits the *critical stepsize* (solid gray, γ_*), becoming a non-linear function of γ. The critical stepsize appears to be a good predictor for the optimal stepsize. In addition, the more over-parameterized the data matrix ($r \to 0$) is, the smaller the window of convergent stepsizes and as $\boldsymbol{H}$ becomes ill-conditioned ($r \to 1$), the linear rate degenerates and the high temperature phase disappears.

Theorem 1.2 (Critical stepsize, asymptotic rates)**.** *Suppose $r \neq 1$ (i.e. strongly convex regime). For $\gamma_* < \gamma < \frac{2}{r}\big(\int_0^\infty x\,\mathrm{d}\mu(x)\big)^{-1}$, the value of $\lambda^*(\gamma)$ is given as the unique solution to*

$$r\gamma^2 \int_0^\infty e^{2\gamma\lambda^* t} h_2(t)\,\mathrm{d}t = 1\,. \tag{14}$$

The function $\psi_0(t)$ satisfies that for some explicit constant $c(R, \widetilde{R}, \mu) > 0$,

$$\psi_0(t) - \frac{\widetilde{R}}{2} \cdot \frac{r\mu(\{0\}) + (1-r)}{1 - \frac{\gamma r}{2}\big(\int_0^\infty x\,\mathrm{d}\mu(x)\big)} \sim \frac{c}{\gamma} e^{-2\gamma t \lambda^*(\gamma)}. \tag{15}$$

If in addition $\gamma_* > 0$, and $\mu([\lambda^-, \lambda^- + t]) \sim c_\mu t^\alpha$ as $t \to 0$ then there is a constant $c(R, \widetilde{R}, \gamma, \mu) > 0$ so that for $0 < \gamma < \gamma_*$,

$$\psi_0(t) - \frac{\widetilde{R}}{2} \cdot \frac{r\mu(\{0\}) + (1-r)}{1 - \frac{\gamma r}{2}\big(\int_0^\infty x\,\mathrm{d}\mu(x)\big)} \sim \frac{c}{t^\alpha} e^{-2\gamma t \lambda^-}. \tag{16}$$

We also give rates for the case of $r = 1$ in Thm E.3. See App. E.1 for further discussion and the derivation.

	Strongly convex, $\gamma < \gamma_*$	**Strongly convex, $\gamma = \gamma_*$**
Worst	$\exp(-\gamma t \lambda^- + \frac{\gamma^2}{2}(\lambda^+)^2 t)$	$\exp(-\gamma t \lambda^- + \frac{\gamma^2}{2}(\lambda^+)^2 t)$
Average	$\exp(-2\gamma\lambda^- t) \cdot \frac{1}{t^{3/2}}$	$\exp(-2\gamma\lambda^- t) \cdot \frac{1}{t^{1/2}}$
	Strongly convex, $\gamma > \gamma_*$	**Non-strongly convex w/ noise, $r = 1, \widetilde{R} > 0$**
Worst	$\exp(-\gamma t \lambda^- + \frac{\gamma^2}{2}(\lambda^+)^2 t)$	$(R + \widetilde{R} \cdot d) \cdot \frac{1}{t}$
Average	$\exp[-\gamma t(1 - \frac{r\gamma}{2})(1 + r + \sqrt{(1+r)^2 - \frac{8}{\gamma}})]$	$R \cdot \frac{1}{t^{1/2}} + \widetilde{R} \cdot \frac{1}{t^{1/2}}$

Table 1: **Asymptotic convergence guarantees** for $f(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}) - \frac{\widetilde{R}}{2}(1 - \frac{r\gamma}{2})^{-1} \max\{0, 1 - r\}$ on the isotropic features model. Stepsizes smaller than γ_* have linear rates based only on λ^- multiplied by a polynomial term (t^α in Theorem 1.2). Larger stepsizes have linear rates with factor λ^* made explicit here. Average-case complexity are strictly better than the worst-case complexity, in some cases by a factor γ vs γ^2. Note also how the rates highlight the freezing transition in the strongly convex regime. For worst-case rates, see [Bottou et al., 2018, Theorem 4.6] [Ghadimi and Lan, 2013, Theorem 2.1]; λ^+ can be replaced by the max-ℓ^2-row-norm in the worst-case bounds below.

Average-case complexity for SGD. Our last contribution is one of the first average-case complexity results for any stochastic optimization algorithm. The value $\psi_0(t)$ is the average function value at iteration t after first taking the model size to infinity. Consequently, this yields a notion of average complexity for SGD to a neighborhood. When the data matrix $\boldsymbol{A}$ satisfies the isotropic features model $\boldsymbol{A}$, we give an explicit formula for the expected function values $\psi_0(t)$, the critical stepsize γ_*, and the corresponding λ^* (Appendix E.2, Theorem E.5 and Section 3, Theorem 3.1). Table 1 summarizes our average rates.

The average-case complexity in the strongly convex case has significantly better linear rates than the worst-case guarantees and, in particular, there is no dependence on λ^+. We additionally capture a second-order behavior, the *polynomial correction term* (green in Table 1). This polynomial term has little effect on the complexity compared to the linear rate. However as the matrix $\boldsymbol{H}$ becomes ill-conditioned ($r \to 1$), the polynomial correction starts to dominate the average-case complexity. The sublinear rates in Table 1 for $r = 1$ show this effect and it accounts for the improved average-case rates in the convex setting. This improvement in the average rate indeed highlights that *the support of the spectrum does not fully determine the rate.* Many eigenvalues contribute meaningfully to the average rate. Hence, our results are not and cannot be purely explained by the support of the spectrum. As noted in Paquette et al. [2020], the worst-case rates when $r = 1$ have dimension-dependent constants due to the distance to the optimum $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|^2 \approx d$, which appears in the bounds.

Related work. *Average-case versus worst-case complexity.* Traditional worst-case analysis of optimization algorithms provide complexity bounds no matter how unlikely [Nemirovski, 1995, Nesterov, 2004]. There are a plethora of results on the worst-case analysis of SGD [Bertsekas and Tsitsiklis, 2000, Bottou et al., 2018, Ghadimi and Lan, 2013, Gower et al., 2019, Robbins and Monro, 1951] and in particular, specific results for SGD applied to the least squares problem (see *e.g.* Bertsekas [1997], Jain et al. [2018]). Worst-case analysis gives convergence guarantees, but the bounds are not always representative of typical runtime.

Average-case analysis, in contrast, gives sharper runtime estimates when some or all of its inputs are random. This type of analysis has a long history in computer-science and numerical analysis and it is often used to justify the superior performances of algorithms such as QuickSort [Hoare, 1962] and the simplex method, see for *e.g.*, [Borgwardt, 1986, Smale, 1983, Spielman and Teng, 2004]. Despite its rich history, average-case is rarely used in optimization due to the ill-defined notion of a typical objective function. Recently Lacotte and Pilanci [2020], Pedregosa and Scieur [2020] derived a framework for average-case analysis of gradient-based methods on the least-squares problem with vanishing noise and it was later extended by Paquette et al. [2020]. Similar results for the conjugate gradient method were derived in Deift and Trogdon [2020], Paquette and Trogdon [2020]. Our work is in the same line of research–providing the first average-case complexity for SGD.

For stochastic algorithms, Sagun et al. [2017] showed empirical evidence that SGD on neural networks exhibits

concentration of the function values. Other works Gurbuzbalaban et al. [2020], Huang et al. [2020], Mei et al. [2018, 2019], Sirignano and Spiliopoulos [2020] have used random matrix theory to analyze stochastic algorithms, but only in online or one-pass settings ($n \gg d$). We emphasize that our work applies to the finite sum setting; as we allow for multiple passes over the data.

Continuous time processes. A popular approach [An et al., 2018, Jastrzebski et al., 2017, Li et al., 2017, Mandt et al., 2016, Nguyen et al., 2019, Zhu et al., 2019] is to model the dynamics of SGD by imposing some structure on the noise and, by sending stepsize to 0, relate the iterates of SGD to the stochastic differential equation (SDE):

$$\mathrm{d}\boldsymbol{X}_t = -\nabla f(\boldsymbol{X}_t)\,\mathrm{d}t + (\gamma \boldsymbol{\Sigma}(\boldsymbol{X}_t))^{1/2}\,\mathrm{d}\boldsymbol{B}_t. \tag{17}$$

Here one typically assumes the stochastic gradient noise $\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})$ is normally distributed (but not necessarily [Simsekli et al., 2019]) with some specific covariance structure $\boldsymbol{\Sigma}(\boldsymbol{X})$. A common choice, called the stochastic modified equation (SME) [Li et al., 2017, Mandt et al., 2016], matches the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{X})$ of the Gaussian noise with the actual covariance of the stochastic gradients at $\boldsymbol{x}$ (*i.e.* $\boldsymbol{\Sigma}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^n (\nabla f(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}))(\nabla f(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x}))^T$). This covariance makes SME have correct mean behavior so the expected function values of the SME model are good approximations for the expected function values of SGD. Li et al. [2017] show that by taking the stepsize γ small, the behavior of SGD and SME align. They and Mandt et al. [2016] also give a modified SME which gives even higher order accuracy of SGD as stepsize goes to 0.

These SDEs have been used to study numerous properties of SGD including the dynamics of regularized loss functions [Kunin et al., 2020] and generalization [Jastrzebski et al., 2017, Pflug, 1986, Simsekli et al., 2019, Zhu et al., 2019]. Despite their wide use, it has been observed that there is no small stepsize limit SGD that converges to an SDE [Yaida, 2019]. Our approach, instead, looks at the large-n limit and shows, in fact, that SGD concentrates while maintaining fixed stepsize. Moreover, our Volterra equation is relatively easy to analyze. We note that the SME has the same mean behavior as SGD so when $n \to \infty$, the mean behavior of SME and our Volterra equation match. *However the SME does not capture this concentration effect and greatly overestimates the fluctuations of the sub-optimality.*

2 Dynamics of SGD: reduction to the Volterra equation

In this section, we develop the framework for the dynamics of SGD and sketch the argument of our main result (Theorem 1.1). Full proofs can be found in Appendix B.

Step 1: Change of basis. A key feature of the SGD least squares iteration (3) is that the projection of $\boldsymbol{x}_k$ onto a singular vector $\boldsymbol{v}_j$ of $\boldsymbol{A}$ with singular value σ_j decreases in expectation exponentially in the number of iterations at a rate proportionally to the squared singular value σ_j^2 [Steinerberger, 2020, Strohmer and Vershynin, 2009]. This observation suggests the following change of basis. Consider the singular value decomposition of $\frac{1}{\sqrt{n}}\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, i.e. $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{\Sigma}$ is the $n \times d$ singular value matrix with diagonal entries $\mathbf{diag}(\sigma_j), j = 1, \ldots, d$. We define the *spectral weight* vector $\widehat{\boldsymbol{\nu}}_k \overset{\text{def}}{=} \boldsymbol{V}^T(\boldsymbol{x}_k - \widetilde{\boldsymbol{x}})$, which therefore evolves like

$$\widehat{\boldsymbol{\nu}}_{k+1} = \widehat{\boldsymbol{\nu}}_k - \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_k (\boldsymbol{U}\boldsymbol{\Sigma}\widehat{\boldsymbol{\nu}}_k - \boldsymbol{\eta}). \tag{18}$$

For this point on, we consider the evolution of $\widehat{\boldsymbol{\nu}}$. We note our above observation on the singular vectors only holds on average for individual coordinates of $\boldsymbol{x}_k$, and it does not alone explain the emergence of the Volterra equation dynamics. It also guarantees nothing about the concentration of the suboptimality.

Step 2: Embedding into continuous time. We next consider an embedding of the $\widehat{\boldsymbol{\nu}}$ into continuous time. This is done to simplify the analysis, and it does not change the underlying behavior of SGD. We let N_t be a standard univariate Poisson process with rate $\frac{n}{\beta}$, so that for any $t > 0$, $\mathbb{E}(N_t) = \frac{nt}{\beta}$. We embed the spectral weights $\widehat{\boldsymbol{\nu}}$ into continuous time, by taking $\boldsymbol{\nu}_t = \widehat{\boldsymbol{\nu}}_{\tau_{N_t}}$. We note that we have scaled time (by choosing the rate of the Poisson process) so that in a single unit of time t, the algorithm has done one complete pass (in expectation) over the data set.

We then show that $f(\boldsymbol{x}_{N_t})$ is well approximated by $\psi_0(t)$. As the mean of N_t is large for any fixed $t > 0$, the Poisson process concentrates around $\frac{nt}{\beta}$, and it follows as an immediate corollary that $f(\boldsymbol{x}_{\lfloor \frac{nt}{\beta} \rfloor})$ is also well approximated by $\psi_0(t)$.

Step 3: Doob–Meyer decomposition & the approximate Volterra equation. Under this continuous-time scaling, we can write the function values at $\boldsymbol{x}_t$ in terms of $\boldsymbol{\nu}_t$ as

$$\psi_\varepsilon(t) \stackrel{\text{def}}{=} f(\boldsymbol{x}_{N_t}) = \frac{1}{2}\|\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta}\|^2 = \frac{1}{2}\sum_{j=1}^{d}\sigma_j^2\nu_{t,j}^2 - \sum_{j=1}^{n\wedge d}\sigma_j\nu_{t,j}\,(\boldsymbol{U}^T\boldsymbol{\eta})_j + \frac{1}{2}\|\boldsymbol{\eta}\|^2, \tag{19}$$

where $\nu_{t,j}$ is the j-th coordinate of the vector $\boldsymbol{\nu}_t$. Hence the dynamics of $\psi_\varepsilon(t)$ are governed by the behaviors of $\nu_{t,j}$ and $\nu_{t,j}^2$ processes. Using (18) and Doob decomposition for quasi-martingales [Protter, 2005, Thm 18, Chpt. 3], we have an expression for the $\nu_{t,j}$ and $\nu_{t,j}^2$, that is, if we let $\mathcal{F}_t$ be the σ-algebra of the information available to the process at time $t \geq 0$, we get

$$\begin{gathered}
\nu_{t,j} = \nu_{0,j} + \int_0^t \mathcal{B}_{s,j}\,\mathrm{d}s + \widetilde{M}_{t,j} \quad \text{and} \quad \nu_{t,j}^2 = \nu_{0,j}^2 + \int_0^t \mathcal{A}_{s,j}\,\mathrm{d}s + M_{t,j}, \\
\text{where} \quad \mathcal{B}_{t,j} \stackrel{\text{def}}{=} \partial_t\,\mathbb{E}[\nu_{t,j} \mid \mathcal{F}_t] = -\gamma\sigma_j^2\nu_{t,j} + \gamma\sigma_j(\boldsymbol{U}^T\boldsymbol{\eta})_j, \\
\mathcal{A}_{t,j} \stackrel{\text{def}}{=} \partial_t\,\mathbb{E}[\nu_{t,j}^2 \mid \mathcal{F}_t] = 2\nu_{t,j}\mathcal{B}_{t,j} + \tfrac{\beta-1}{n-1}\big(\mathcal{B}_{t,j}\big)^2 + \gamma^2\sigma_j^2\big(1 - \tfrac{\beta-1}{n-1}\big)\sum_{i=1}^{n}\big(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\big)^2\big(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})\big)^2,
\end{gathered} \tag{20}$$

and $(M_{t,j}, \widetilde{M}_{t,j} : t \geq 0)$ are $\mathcal{F}_t$–adapted martingales. The last identities for $\mathcal{B}_{t,j}$ and $\mathcal{A}_{t,j}$ are derived in Lemma B.1 in Appendix B).

We will now see how the terms $\mathcal{A}_{t,j}$ and $\mathcal{B}_{t,j}$ can be simplified in the large-n limit. In this regime, sums of spectral quantities converge to integrals against the limiting spectral measure μ as a direct consequence of Assumptions 1.1 and 1.2. Since we are working in the regime where $\beta = o(n)$, the terms with $\frac{\beta-1}{n-1}$ vanish in the large-n limit, disappearing entirely when $\beta = 1$, and explaining why $\beta = o(n)$ does not affect the limiting dynamics of SGD. Our *key lemma*, which explains the Volterra dynamics of the mean of $f(\boldsymbol{x}_{N_T})$ (Lemma B.5, App. B.6.2), is that $\big(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\big)^2$ self averages to $\frac{1}{n}$, and this is the point where we leverage orthogonal invariance of $\boldsymbol{A}$ most heavily.

These simplifications can be summarized as

$$\mathcal{A}_{t,j} \approx \widehat{\mathcal{A}}_{t,j} \stackrel{\text{def}}{=} -\gamma 2\sigma_j^2\nu_{t,j}^2 + \gamma 2\sigma_j\nu_{t,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j + \gamma^2\tfrac{2\sigma_j^2\psi_\varepsilon(t)}{n}. \tag{21}$$

The expression $\widehat{\mathcal{A}}_{t,j}$ explains the limiting Volterra dynamics for $\psi_\epsilon(t)$, and why the mean "gradient flow" term does not correctly describe the dynamics of SGD. Due to the gradient flow term, the squared spectral weights $\nu_{t,j}^2$ tend to decay linearly with rate $2\gamma\sigma_j^2$. On the other hand, coordinates can not decay too quickly, as there is a mass redistribution term, which explains the rate at which mass from other spectral weights is added to $\nu_{t,j}^2$ and which is due to SGD updates being noisy analogues for gradient flow. Finally, there is a noise term which in principle depends on $\nu_{t,j}$ which would greatly complicate the limiting dynamics. However, when averaged in j the independence of the noise $\boldsymbol{\eta}$ leads to a concentration effect, due to which only the mean behavior of $\nu_{t,j}$ survives. As this mean is just gradient flow, this leads to a simple deterministic forcing term in the Volterra equation.

Plugging (21) into (19) and (20), we can produce a perturbed Volterra equation for $\psi_\varepsilon(t)$. For any $t > 0$ we have

$$\psi_\varepsilon(t) = \frac{Rh_1(t)}{2} + \frac{\widetilde{R}(rh_0(t) + (1-r))}{2} + \varepsilon_1^{(n)}(t) + \int_0^t \big(\gamma^2 r h_2(s) + \varepsilon_2^{(n)}(s)\big)\psi_\varepsilon(t-s)\,\mathrm{d}s, \tag{22}$$

for error terms $\varepsilon_i^{(n)}$ (see Appendix B.6 for a precise definition of the errors). The $h_k(t)$ are defined in Theorem 1.1 as the Laplace transforms of the measure μ, and arise naturally due to the presence of the gradient flow generator.

Step 4: Control of the errors and stability of the Volterra equation. The expression (22) is a Volterra equation of convolution type — a well-studied equation, see *e.g.*, Gripenberg et al. [1990], with established stability and existence/uniqueness theorems. In particular, we can summarily conclude that (see Proposition B.1 in Appendix B.5)

$$\max_{i=1,2}\sup_{0\leq t\leq T}|\varepsilon_i^{(n)}(t)| \xrightarrow[n\to\infty]{\Pr} 0 \quad \Longrightarrow \quad \sup_{0\leq t\leq T}|\psi_\varepsilon(t) - \psi_0(t)| \xrightarrow[n\to\infty]{\Pr} 0.$$

Thus, Theorem 1.1, the dynamics for SGD immediately follows provided control of the errors in (22).

Beyond controlling the error of $\mathcal{A}_{t,j} - \widehat{\mathcal{A}}_{t,j}$, we must separately control the fluctuations of the martingale terms in (20), which represent the randomness of SGD. A central challenge here is to show in a suitable sense that the entries of $\sqrt{n}\boldsymbol{\nu}_t$ remain bounded on compact sets of time (see the discussion in Appendix B.6 for a detailed overview), which in turn can be seen as a consequence of the updates of SGD being very nearly orthogonal to any fixed row of $\boldsymbol{U}$. Here again we use the orthogonal invariance of $\boldsymbol{A}$, but in a weaker way, in that we only need that the maximum of the entries of $\boldsymbol{U}$ are in control. Such results are well-developed for other random matrix ensembles.

3 Explicit formulas for isotropic features

We solve the Volterra equation and derive exact expressions for the average-case analysis, the critical stepsize γ_*, and the rate λ^* (Thm 1.2) under the isotropic features model. In this case, the empirical spectral measure converges to the Marchenko-Pastur measure μ_{MP} (8). Volterra equations of convolution type can be solved using Laplace transforms, which conveniently, for Marchenko-Pastur, are explicit due to a connection with the Stieltjes transform. This leads us to our next main result.

Theorem 3.1 (Dynamics of SGD in noiseless setting)**.** *Suppose $\widetilde{R} = 0$ and the stepsize $\gamma < \frac{2}{r}$. Define the constants ϱ and ω and critical stepsize*

$$\varrho = \frac{1+r}{2}\left(1 - \frac{r\gamma}{2}\right), \ \omega = \frac{1}{4}\left(1 - \frac{r\gamma}{2}\right)^2 \left(\frac{8}{\gamma} - (1+r)^2\right), \text{ and } \gamma_* = \frac{2}{\sqrt{r}(r - \sqrt{r} + 1)}. \tag{23}$$

The iterates of SGD satisfy if $\gamma \leq \gamma_*$,

$$f(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}) \xrightarrow[n\to\infty]{\Pr} R \cdot \frac{1}{\gamma}\left(1 - \frac{r\gamma}{2}\right) \int_0^\infty \frac{x e^{-2\gamma x t}}{(x-\varrho)^2 + \omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x)$$

and if $\gamma > \gamma_*$, for some explicit constant $c(\gamma, r)$, the iterates of SGD follow

$$f(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}) \xrightarrow[n\to\infty]{\Pr} R \cdot \frac{1}{\gamma}\left(1 - \frac{r\gamma}{2}\right) \int_0^\infty \frac{x e^{-2\gamma x t}}{(x-\varrho)^2 + \omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x) + R \cdot c(\gamma, r) \cdot e^{-2\gamma(\varrho + \sqrt{|\omega|})t}.$$

We only record the dynamics for SGD in the noiseless regime and refer the reader to the Appendix E.2, Theorem E.5 for the noisy setting. We first observe the freezing transition as predicted by renewal theory – a jamming term appears for $\gamma > \gamma_*$ that slows convergence. We note that when the ratio of features to samples r does not equal 1, the least squares problem in (2) is (almost surely) strongly convex as $\mathrm{d}\mu_{\mathrm{MP}}$ has a gap between the first non-zero eigenvalue and zero (see Figure 2). As r approaches 1, the smallest non-zero eigenvalue become arbitrarily close to 0. This phenomenon suggests different convergence rates in the regimes $r = 1$ and $r \neq 1$. Moreover, we see the explicit value of λ^*, $\varrho + \sqrt{|\omega|}$ which vanishes when $r = 1$. We present our average-case rates in Table 1.

4 Numerical simulations

We compare models of SGD's dynamics on two data distributions for moderately-sized problems ($n = 1000$): the isotropic features model (see Section 1.1) and one-hidden layer network with random weights. In the latter model, the entries of $\boldsymbol{A}$ are the result of a matrix multiplication composed with an activation function $g : \mathbb{R} \to \mathbb{R}$:

$$A_{ij} \overset{\text{def}}{=} g\big(\tfrac{[\boldsymbol{W}\boldsymbol{Y}]_{ij}}{\sqrt{m}}\big), \quad \text{where } \boldsymbol{W} \in \mathbb{R}^{n\times m}, \boldsymbol{Y} \in \mathbb{R}^{m\times d} \text{ are random matrices.} \tag{24}$$

For the simulations, we took this activation function g to be a shifted ReLU function; the shift makes $\mathbb{E}[\boldsymbol{A}] = 0$ (see Appendix A for details). This model encompasses two-layer neural networks with a squared loss, where the first layer has random weights and the second layer's weights are given by the regression coefficients $\boldsymbol{x}$. Note that while the isotropic features model satisfies our assumptions, the one-hidden layer model does not. For all these approaches, we compute the objective suboptimality as a function of the number of passes over the dataset (epochs) for the models: (1). SDE (*i.e.*, $\boldsymbol{\Sigma}(\boldsymbol{X}) = 0.01\boldsymbol{I}$ in (17)), (2). SME (*i.e.*, $\boldsymbol{\Sigma}(\boldsymbol{X})$ matches covariance of the stochastic gradients), (3). streaming (regenerate $\boldsymbol{a}_i$ at each step), and (4). our Volterra equation. See Appendix F for full details on the setup as well as experiments with other values of r. The outcome is displayed in Figure 4 and discussed in the caption.

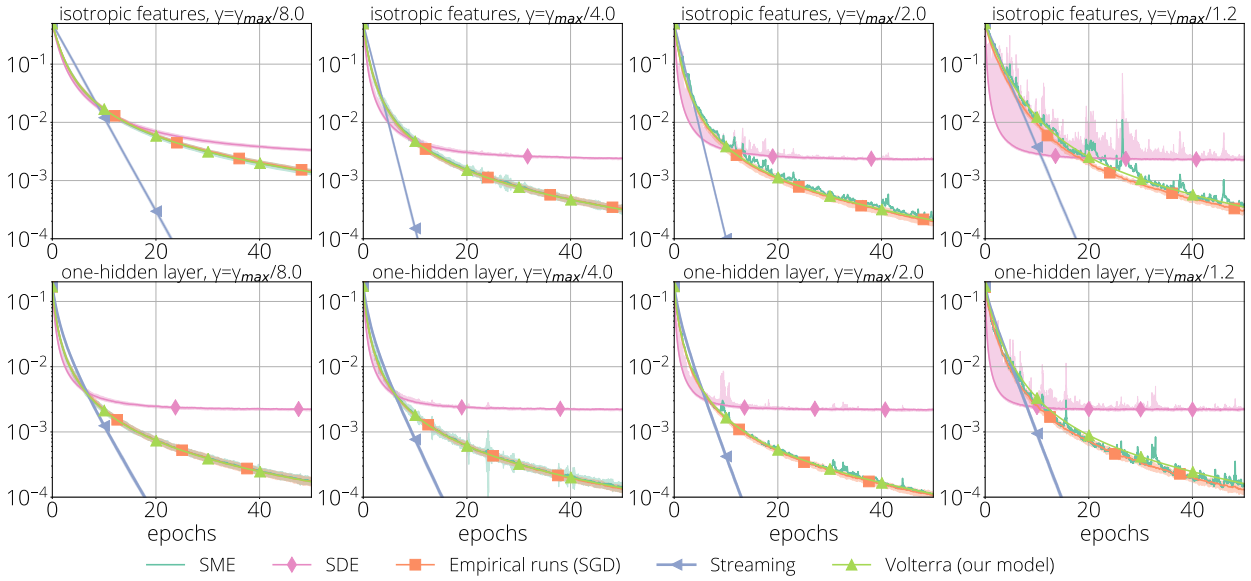

Figure 4: **Comparison of different SGD models**: isotropic features (top) and one-hidden layer network (bottom); $r = 1.2$. Across all stepsizes the Volterra overlaps the objective suboptimality of the empirical SGD runs (orange). The SME (teal) fits SGD for small stepsizes, whereas streaming (blue) and SDE (pink) have noticeable divergences from SGD for all stepsizes. Stochastic methods were averaged across 10 runs, with filled area representing the standard deviation. The parameter $\gamma_{\max}$ is the largest stepsize which still yields convergence of SGD, $\gamma_{\max} = \frac{2}{r}(\frac{1}{d}\mathrm{tr}(\boldsymbol{H}))^{-1}$ from Theorem 1.2.

The fit of the Volterra equation to SGD is extremely accurate across different stepsizes and data distributions (some not covered by our assumptions) and even for medium-sized problems ($n = 1000$). We also note that while SME is often a good approximation, obtaining convergence rates from it is an open problem. On the other hand, the proposed Volterra equation can be analyzed through its link with renewal theory.

Conclusion and future work We have shown that the SGD method on least squares objectives admits a tight analysis in the large n and d limit. We described the dynamics of this algorithm through a Volterra integral equation and characterize its average-case convergence rate as well as its stepsize regimes. Although our results only hold in the large n-limit, the Volterra equation is remarkably accurate for relatively small dimensions (see *e.g.* Figure 4).

While our theoretical results focus on problems with isotropic data matrix $\boldsymbol{A}$, Figure 4 shows that the Volterra equation also predicts remarkably well the dynamics on data generated from a one-hidden layer network model. This suggests that the Volterra prediction might hold in even greater generality, a conjecture that is left for future work. Another direction of future work consists in extending to include other algorithms and problems. We believe the framework presented here should apply to methods like SGD momentum, RMSprop or ADAM and problems such as PCA.

Acknowledgements

The authors would like to thank our colleagues Nicolas Le Roux, Bart van Merriënboer, Zaid Harchaoui, Manuela Girotti, Gauthier Gidel, and Dmitriy Drusvyatskiy for their feedback on this manuscript.

References

N. Alexeev, F. Götze, and A. Tikhomirov. Asymptotic distribution of singular values of powers of random matrices. *Lith. Math. J.*, 50(2):121–132, 2010.

J. An, J. Lu, and L. Ying. Stochastic modified equations for the asynchronous stochastic gradient descent. *arXiv preprint arXiv:1805.08244*, 2018.

S. Asmussen. *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

S. Asmussen, S. Foss, and D. Korshunov. Asymptotics for sums of random variables with local subexponential behaviour. *J. Theoret. Probab.*, 16(2):489–518, 2003.

Z. Bai and J. Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

G. Behrooz, S. Krishnan, and Y. Xiao. An Investigation into Neural Net Optimization via Hessian Eigenvalue Density. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

L. Benigni and S. Péché. Eigenvalue distribution of nonlinear models of random matrices. *arXiv preprint arXiv:1904.03090*, 2019.

D. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM J. Optim.*, 7(4):913–926, 1997.

D. Bertsekas and J. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM J. Optim.*, 10(3):627–642, 2000.

R. Bollapragada, R. Byrd, and J. Nocedal. Adaptive sampling strategies for stochastic optimization. *SIAM J. Optim.*, 28(4):3312–3343, 2018.

K. Borgwardt. *A Probabilistic Analysis of the Simplex Method*. Springer-Verlag, Berlin, Heidelberg, 1986.

L. Bottou, F.E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2): 223–311, 2018.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *International Conference on Learning Representations (ICLR)*, 2018.

G. Cipolloni, L. Erdős, and D. Schröder. Eigenstate Thermalization Hypothesis for Wigner Matrices. *arXiv e-prints*, art. arXiv:2012.13215, 2020.

Y.N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

P.A. Deift and T. Trogdon. The conjugate gradient algorithm on well-conditioned Wishart matrices is almost deterministic. *Quarterly of Applied Mathematics*, 2020.

B Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, 24:2613–2626, Sep 1981.

M. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM J. Sci. Comput.*, 34 (3):A1380–A1405, 2012.

Y. Fyodorov and J. Bouchaud. Freezing and extreme-value statistics in a random energy model with logarithmically correlated potential. *J. Phys. A*, 41(37):372001, 12, 2008.

C. Gerbelot, A. Abbara, and F. Krzakala. Asymptotic Errors for High-Dimensional Convex Penalized Linear Regression beyond Gaussian Matrices. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 1682–1713. PMLR, 2020.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013.

R. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning (ICML)*. PMLR, 2019.

G. Gripenberg, S.O. Londen, and O. Staffans. *Volterra Integral and Functional Equations*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1990.

M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The Heavy-Tail Phenomenon in SGD. *arXiv preprint arXiv:2006.04740*, 2020.

T. Hastie, A. Montanari, S. Rosset, and R.J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

C.A.R. Hoare. Quicksort. *The Computer Journal*, 5(1):10–16, 01 1962.

W. Hu, C. Junchi Li, L. Li, and J. Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

D. Huang, J. Niles-Weed, J. Tropp, and R. Ward. Matrix Concentration for Products. *arXiv preprints arXiv:2003.05437*, 2020.

P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent for Least Squares Regression. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, volume 75, pages 545–604, 2018.

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three Factors Influencing Minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

N. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv preprint arXiv:1609.04836*, 2016.

J. F. C. Kingman. *Poisson Processes*. Clarendon Press·Oxford, 1993.

B. Klar. Bounds on Tail Probabilities of Discrete Distributions. *Probability in the Engineering and Informational Sciences*, 14:161–171, 2000.

A. Knowles and J. Yin. Anisotropic local laws for random matrices. *Probab. Theory Related Fields*, 169(1-2):257–352, 2017.

D. Kunin, J. Sagastuy-Brena, S. Ganguli, D.L. K. Yamins, and H. Tanaka. Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics. *arXiv preprint arXiv:2012.04728*, 2020.

H. Kushner and G.G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

J. Lacotte and M. Pilanci. Optimal Randomized First-Order Methods for Least-Squares Problems. *arXiv preprint arXiv:2002.09488*, 2020.

D. Lépingle. Sur le comportement asymptotique des martingales locales. In *Séminaire de Probabilités, XII (Univ. Strasbourg, Strasbourg, 1976/1977)*, volume 649 of *Lecture Notes in Math.*, pages 148–161. Springer, Berlin, Heidelberg, 1978.

Q. Li, C. Tai, and W. E. Stochastic Modified Equations and Adaptive Stochastic Gradient Algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICLR)*, volume 70, pages 2101–2110, 2017.

D. Liu, C. Song, and Z. Wang. On explicit probability densities associated with Fuss-Catalan numbers. *Proc. Amer. Math. Soc.*, 139(10):3735–3738, 2011.

L. Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, AC-22(4):551–575, 1977.

M. Mahsereci and P. Hennig. Probabilistic Line Searches for Stochastic Optimization. *Journal of Machine Learning Research*, 18(119):1–59, 2017.

S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International conference on machine learning (ICML)*, 2016.

V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1967.

C.H. Martin and M.W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*, 2018.

E. Meckes. *The random matrix theory of the classical compact groups*, volume 218 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2019.

S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

S. Mei, A. Montanari, and P. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018.

S. Mei, T. Misiakiewicz, and A. Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory (COLT)*, volume 99, pages 2388–2464, 2019.

A. Nemirovski. Information-based complexity of convex programming. *Lecture Notes*, 1995.

Y. Nesterov. *Introductory lectures on convex optimization*. Springer, 2004.

T. Nguyen, U. Simsekli, M. Gurbuzbalaban, and G. Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

V. Papyan. The full spectrum of deepnet hessians at scale: Dynamics with SGD Training and Sample Size. *arXiv preprint arXiv:1811.07062*, 2018.

C. Paquette, B. van Merriënboer, and F. Pedregosa. Halting Time is Predictable for Large Models: A Universality Property and Average-case Analysis. *arXiv preprint arXiv:2006.04299*, 2020.

E. Paquette and T. Trogdon. Universality for the conjugate gradient and MINRES algorithms on sample covariance matrices. *arXiv preprint arXiv:2007.00640*, 2020.

F. Pedregosa and D. Scieur. Average-case Acceleration Through Spectral Density Estimation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

J. Pennington and P. Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

G. C. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM J. Control Optim.*, 24(4): 655–666, 1986.

P.E. Protter. *Stochastic integration and differential equations*, volume 21 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2005.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2008.

S. Resnick. *Adventures in stochastic processes*. Birkhäuser Boston, Inc., Boston, MA, 1992.

H. Robbins and S. Monro. A Stochastic Approximation Method. *Ann. Math. Statist.*, 1951.

L. Sagun, L. Bottou, and Y. LeCun. Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond. *arXiv preprint arXiv:1611.07476*, 2016.

L. Sagun, T. Trogdon, and Y. LeCun. Universal halting times in optimization and machine learning. *Quarterly of Applied Mathematics*, 76:1, 09 2017.

A. Saxe, J. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

T. Schaul, S. Zhang, and Y. LeCun. No more pesky learning rates. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *Proceedings of Machine Learning Research*, pages 343–351, 2013.

G. Shorack and J. Wellner. *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, 1986.

U. Simsekli, L. Sagun, and M. Gurbuzbalaban. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 5827–5837, 2019.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: a law of large numbers. *SIAM J. Appl. Math.*, 80(2):725–752, 2020.

S. Smale. On the average number of steps of the simplex method of linear programming. *Mathematical Programming*, 27(3):241–262, 1983.

D. Spielman and S. Teng. Smoothed Analysis of Algorithms: Why the Simplex Algorithm Usually Takes Polynomial Time. *J. ACM*, 51(3):385–463, 2004.

S. Steinerberger. Randomized Kaczmarz converges along small singular vectors. *arXiv preprint arXiv:2006.16978*, 2020.

T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.

W. Su, S. Boyd, and E.J. Candès. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28, pages 1139–1147, 2013.

T. Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien. Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 3732–3745, 2019.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

S. Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations(ICLR)*, 2019.

M. Yor. On optional stochastic integrals and a remarkable series of exponential formulas. *Strasbourg probability seminar*, 10:481–500, 1976.

Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 7654–7663, 2019.

SGD in the Large:
Average-case Analysis, Asymptotics, and Stepsize Criticality

Supplementary material

The appendix is organized into six sections as follows:

1. Appendix A expands upon the data examples in Section 1.1.
2. Appendix B derives the Volterra equation and proves the main concentration for the dynamics of SGD (Theorem 1.1).
3. We show in Appendix C that the error terms associated with concentration of measure on the high-dimensional orthogonal group disappear in the large-n limit. This includes the *key lemma*, Proposition B.5.
4. Appendix D shows the error terms which vanish due to martingale concentration results.
5. Appendix E derives the average-case complexity results from Section 3 and provides a proof of the Malthusian exponent (Theorem 1.2).
6. Appendix F contains details on the simulations.

Unless otherwise stated, all the results hold under Assumptions 1.1 and 1.2. We include all statements from the previous sections for clarity.

Notation. All stochastic quantities defined hereafter live on a probability space denoted by $(\Pr, \Omega, \mathcal{F})$ with probability measure $\Pr$ and the σ-algebra $\mathcal{F}$ containing subsets of Ω. A random variable (vector) is a measurable map from Ω to $\mathbb{R}$ ($\mathbb{R}^d$) respectively. Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ be a random variable mapping into the borel σ-algebra $\mathcal{B}$ and the set $B \in \mathcal{B}$. We use the standard shorthand for the event $\{X \in B\} = \{\omega : X(\omega) \in B\}$. We denote the minimum of a and b by $\min\{a, b\} = a \wedge b$. An event E that occurs *with high probability* (or shortened to w.h.p.) is one whose probability depends on n, related to matrix dimension in our paper, and the probability of its complementary event goes to 0 as $n \to \infty$. Whereas event E is said to occur *with overwhelming probability* (or w.o.p.) if the probability of its complementary event goes to 0 faster than any polynomial order of n as $n \to \infty$, i.e. for any $k > 0$,

$$\lim_{n\to\infty} \Pr(E^c) n^k = 0.$$

Throughout the paper, $\beta = \beta(n)$, which denotes the size of B_k, a uniformly random subset of $\{1, \cdots, n\}$ at the k-th iteration of SGD, is assumed to satisfy $\beta \leq n^{1/5-\delta}$ for some $\delta > 0$.

A Data distributions

A.1 Elaboration on isotropy

We recall that in Assumption 1.2, we have assumed that the data matrix $\boldsymbol{A}$ is *orthogonally invariant*. This is a strong from of isotropy, under which the matrix looks the same in any orthogonal basis. On a technical level, we work in this setting as it leads to a singular value decomposition with especially simple statistics.

To state this property, we recall that the set of $n \times n$ orthogonal matrices form a group $O(n)$ under multiplication, and that this group naturally admits a Lie group structure. In particular, there is a probability measure on this group, the *Haar measure*, which is invariant by left and right multiplication by fixed orthogonal matrices. To refer to a random matrix whose law is Haar measure, we will simply say a Haar-distributed orthogonal matrix. While it may appear unwieldy, there are many exceptionally nice tools that exist for working with this measure. We will elaborate on many of them in Section C. We also refer to Meckes [2019] for a rich exposition on the intrinsic properties of this group.

The main feature that we will need is the following:

Lemma A.1. *Suppose that $\boldsymbol{A}$ is an $n \times d$ orthogonally invariant random matrix in that $\boldsymbol{A}\boldsymbol{O}_d \overset{law}{=} \boldsymbol{A}$ and $\boldsymbol{O}_n\boldsymbol{A} \overset{law}{=} \boldsymbol{A}$ for any orthogonal matrices $\boldsymbol{O}_n \in O(n)$ and $\boldsymbol{O}_d \in O(d)$. Then there is a singular value decomposition*

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$$

with $\boldsymbol{\Sigma}$ an $n \times d$ random matrix having

$$\boldsymbol{\Sigma}_{11} \geq \boldsymbol{\Sigma}_{22} \geq \boldsymbol{\Sigma}_{33} \geq \cdots \geq \boldsymbol{\Sigma}_{mm} \quad \textit{where} \quad m = \min\{n, d\},$$

so that $(\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V})$ are independent, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are Haar orthogonally distributed.

Proof. The key observation is that if we introduce a new, independent Haar distributed random matrix $\boldsymbol{U} \in O(n)$, then $\boldsymbol{U}^T\boldsymbol{A}$ has the same law as $\boldsymbol{A}$ and moreover $(\boldsymbol{U}, \boldsymbol{U}^T\boldsymbol{A})$ are independent. To see that $\boldsymbol{U}^T\boldsymbol{A}$ has the same law as $\boldsymbol{A}$ we just observe that conditionally on $\boldsymbol{U}^T$, $\boldsymbol{U}^T\boldsymbol{A} \overset{\text{law}}{=} \boldsymbol{A}$ by assumption. As the conditional law does not depend on $\boldsymbol{U}$, it follows that $\boldsymbol{U}^T\boldsymbol{A} \overset{\text{law}}{=} \boldsymbol{A}$ *and* $\boldsymbol{U}$ is independent of $\boldsymbol{U}^T\boldsymbol{A}$. Extending this, if we introduce two new independent Haar distributed random matrices $\boldsymbol{U} \in O(n)$ and $\boldsymbol{V}$ in $O(d)$, it follows that $(\boldsymbol{U}, \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V}^T, \boldsymbol{V})$ is a triple on independent random matrices. Let

$$\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V}^T = \widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{V}}$$

be the singular value decomposition with $\boldsymbol{\Sigma}$ having the properties stated in the lemma. Then

$$\boldsymbol{A} = \boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V})\boldsymbol{V}^T = (\boldsymbol{U}\widetilde{\boldsymbol{U}})\boldsymbol{\Sigma}(\widetilde{\boldsymbol{V}}\boldsymbol{V}),$$

with $\boldsymbol{U}, \boldsymbol{V}$ and $(\widetilde{\boldsymbol{U}}\boldsymbol{\Sigma}\widetilde{\boldsymbol{V}})$ independent. By invariance of Haar measure, the triple $(\boldsymbol{U}\widetilde{\boldsymbol{U}}, \boldsymbol{V}\widetilde{\boldsymbol{V}}, \boldsymbol{\Sigma})$ remain independent and uniformly distributed on $O(n)$ and $O(d)$ respectively. □

As a consequence, we will frequently condition on the singular values $\boldsymbol{\Sigma}$ of $\boldsymbol{A}$, and most estimates we need are estimates that hold conditionally on $\boldsymbol{\Sigma}$.

A.2 Isotropic features and Random features

In this section, we expand upon Assumptions 1.1 and 1.2 in the main text of the paper. We discuss in detail two examples: isotropic features and one-hidden layer networks.

A.2.1 Isotropic features

In their seminal work, Marčenko and Pastur [1967] show that the spectrum of $\boldsymbol{H} = \frac{1}{n}\boldsymbol{A}^T\boldsymbol{A}$ under the isotropic features model converged to a deterministic measure. Subsequent work then characterized the convergence of the largest eigenvalue of $\boldsymbol{H}$. We summarize these results below.

Lemma A.2 (Isotropic features)**.** ***(Bai and Silverstein [2010, Theorem 5.8])*** *Suppose the matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is generated using the isotropic features model. Then the empirical spectral measure (EMS) $\mu_{\boldsymbol{H}}$ converges weakly almost surely to the Marchenko-Pastur measure μ_{MP} and the largest eigenvalue of $\boldsymbol{H}$, $\lambda^+_{\boldsymbol{H}}$, converges in probability to λ^+ where $\lambda^+ = (1 + \sqrt{r})^2$ is the top edge of the support of the Marchenko-Pastur measure.*

The results stated so far did not require that the entries of $\boldsymbol{A}$ are normally distributed, and hold equally well for any i.i.d. matrices with mean 0, entry variance 1 and bounded fourth-moment. Under the additional assumption that the entries of $\boldsymbol{A}$ are normally distributed, it is easily checked by a covariance computation that for fixed orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ the entries $\boldsymbol{U}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{V}$ remain independent, mean 0 and variance 1. We summarize this claim below.

Lemma A.3. *For an $n \times d$ matrix $\boldsymbol{A}$ of i.i.d. standard normal random variables, $\boldsymbol{U}\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{V}$ are again $n \times d$ matrices of independent standard normals for fixed orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$.*

We emphasize that while this is not true for matrices $\boldsymbol{A}$ with entries that are independent of mean 0 and variance 1, there are many senses in which this is approximately true (see Knowles and Yin [2017]).

A.2.2 One-hidden layer networks

One-hidden layer network with random weights. In this model, the entries of $\boldsymbol{A}$ are the result of a matrix multiplication composed with a (potentially non-linear) activation function $g : \mathbb{R} \to \mathbb{R}$:

$$A_{ij} \stackrel{\text{def}}{=} g\big(\tfrac{[\boldsymbol{WY}]_{ij}}{\sqrt{m}}\big), \quad \text{where } \boldsymbol{W} \in \mathbb{R}^{n\times m}, \boldsymbol{Y} \in \mathbb{R}^{m\times d} \text{ are random matrices.} \tag{25}$$

The entries of $\boldsymbol{W}$ and $\boldsymbol{Y}$ are i.i.d. with zero mean, isotropic variances $\mathbb{E}\,[W_{ij}^2] = \sigma_w^2$ and $\mathbb{E}\,[Y_{ij}^2] = \sigma_y^2$, and light tails (see App. A.2 for details). As in the previous case to study the large dimensional limit, we assume that the different dimensions grow at comparable rates given by $\frac{m}{n} \to r_1 \in (0,\infty)$ and $\frac{m}{d} \to r_2 \in (0,\infty)$. This model encompasses two-layer neural networks with a squared loss, where the first layer has random weights and the second layer's weights are given by the regression coefficients $\boldsymbol{x}$. Particularly, the optimization problem in (2) becomes

$$\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) = \frac{1}{2n} \|g\big(\tfrac{1}{\sqrt{m}}\boldsymbol{WY}\big)\boldsymbol{x} - \boldsymbol{b}\|_2^2 \right\}. \tag{26}$$

The model was introduced by [Rahimi and Recht, 2008] as a randomized approach for scaling kernel methods to large datasets, and has seen a surge in interest in recent years as a way to study the generalization properties of neural networks [Hastie et al., 2019, Mei and Montanari, 2019, Pennington and Worah, 2017].

The difference between this and the isotropic features model is the activation function, g. We assume g to be entire with a growth condition and have zero Gaussian-mean (App. A.2). These assumptions hold for common activation functions such as sigmoid $g(z) = (1+e^{-z})^{-1}$ and softplus $g(z) = \log(1+e^z)$, a smoothed variant of ReLU.

Benigni and Péché [2019] recently showed that the empirical spectral measure and largest eigenvalue of $\boldsymbol{H}$ converge to a deterministic measure and largest element in the support, respectively. This implies that this model verifies Assumption 1.2. However, contrary to the isotropic features model, the limiting measure does not have an explicit expression, except for some specific instances of g in which it is known to coincide with the Marchenko-Pastur distribution.

For the one-hidden layer network, following [Benigni and Péché, 2019], we assume that the activation function g is an entire function with a growth condition satisfying the following zero Gaussian mean property:

$$\text{(Gaussian mean)} \qquad \int g(\sigma_w\sigma_y z)\tfrac{e^{-z^2/2}}{\sqrt{2\pi}}\,\mathrm{d}z = 0\,. \tag{27}$$

The additional growth condition on the function g is precisely given as there exists positive constants $C_g, c_g, A_0 > 0$ such that for any $A \geq A_0$ and any $n \in \mathbb{N}$

$$\sup_{z\in[-A,A]} |g^{(n)}(z)| \leq C_g A^{c_g n}\,. \tag{28}$$

This growth condition is verified for polynomials which can approximate to arbitrary precision common activation functions such as the sigmoid $g(z) = (1+e^{-z})^{-1}$ and the softplus $g(z) = \log(1+e^z)$, a smoothed approximation to the ReLU. The Gaussian mean assumption (27) can always be satisfied by incorporating a translation into the activation function.

In addition to the i.i.d., mean zero, and isotropic entries, we also require an assumption on the tails of $\boldsymbol{W}$ and $\boldsymbol{Y}$, that is, there exists constants $\theta_w, \theta_y > 0$ and $\alpha > 0$ such that for any $t > 0$

$$\Pr(|W_{11}| > t) \leq \exp(-\theta_w t^\alpha) \quad \text{and} \quad \Pr(|Y_{11}| > t) \leq \exp(-\theta_y t^\alpha). \tag{29}$$

Although stronger than bounded fourth moments, this assumption holds for any sub-Gaussian random variables (*e.g.*, Gaussian, Bernoulli, etc). Under these hypotheses, Assumption 1.2 is verified.

Lemma A.4 (One-hidden layer network)**.** ***(Benigni and Péché [2019, Theorems 2.2 and 5.1])*** *Suppose the matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ is generated using the random features model. Then there exists a deterministic compactly supported measure μ such that $\mu_{\boldsymbol{H}} \underset{d\to\infty}{\longrightarrow} \mu$ weakly almost surely. Moreover $\lambda_{\boldsymbol{H}}^+ \underset{d\to\infty}{\xrightarrow{\Pr}} \lambda^+$ where λ^+ is the top edge of the support of μ.*

B Derivation of the dynamics of SGD

In this section, we derive the Volterra equation from (12), that is,

$$\begin{aligned}\psi_0(t) &= \tfrac{R}{2}h_1(t) + \tfrac{\widetilde{R}}{2}\big(rh_0(t) + (1-r)\big) + \int_0^t \gamma^2 r h_2(t-s)\psi_0(s)\,ds,\\ \text{and} \quad h_k(t) &= \int_0^\infty x^k e^{-2\gamma t x}\,d\mu(x).\end{aligned} \tag{30}$$

and prove Theorem 1.1:

$$\sup_{0\le t\le T}\left|f\big(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t\rfloor}\big) - \psi_0(t)\right| \xrightarrow[n\to\infty]{\Pr} 0, \tag{31}$$

provided that error terms go to zero. We begin by setting up the tools to derive an approximate Volterra equation.

B.1 Change of basis

Recall the iterates of SGD satisfy

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\gamma}{n}\sum_{i\in B_k}\nabla f_i(\boldsymbol{x}_k) = \boldsymbol{x}_k - \frac{\gamma}{n}\boldsymbol{A}^T\boldsymbol{P}_k(\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}), \quad \text{where} \quad \boldsymbol{P}_k \overset{\text{def}}{=} \sum_{i\in B_k}\mathbb{e}_i\mathbb{e}_i^T. \tag{32}$$

Here $\boldsymbol{P}_k$ is a random orthogonal projection matrix with $\mathbb{e}_i$ the i-th standard basis vector, $\beta \in \mathbb{N}$ is a batch-size parameter, which we will allow to depend on n, $\gamma > 0$ is a stepsize parameter, and the function f_i is the i-th element of the sum in (2).

We recall that $\boldsymbol{b}$ has the representation $\boldsymbol{b} = \boldsymbol{A}\widetilde{\boldsymbol{x}} + \sqrt{n}\,\boldsymbol{\eta}$, and both $\widetilde{\boldsymbol{x}}$ and $\boldsymbol{\eta}$ have norms bounded independently of n. Hence we can represent the updates of SGD (3) equation in matrix form as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{\gamma}{n}\boldsymbol{A}^T\boldsymbol{P}_k(\boldsymbol{A}(\boldsymbol{x}_k - \widetilde{\boldsymbol{x}}) + \sqrt{n}\boldsymbol{\eta}).$$

We will consider a singular value decomposition guaranteed by Lemma A.1 of $\frac{1}{\sqrt{n}}\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are Haar distributed orthogonal matrices, i.e. $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{\Sigma}$ is the $n \times d$ singular value matrix with diagonal entries $\mathbf{diag}(\sigma_i), i = 1,\dots,n$. Our analysis will use a different choice of variables. So we define the vector $\widehat{\boldsymbol{\nu}}_k \overset{\text{def}}{=} \boldsymbol{V}^T(\boldsymbol{x}_k - \widetilde{\boldsymbol{x}})$, which therefore evolves like

$$\widehat{\boldsymbol{\nu}}_{k+1} = \widehat{\boldsymbol{\nu}}_k - \gamma\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_k(\boldsymbol{U}\boldsymbol{\Sigma}\widehat{\nu}_k - \boldsymbol{\eta}). \tag{33}$$

B.2 Embedding into continuous time

We next consider an embedding of the process $\widehat{\boldsymbol{\nu}}_k$ into continuous time. This is done to simplify the analysis, and it does not change the underlying behavior of SGD. Let $\mathbb{N}_0 = \mathbb{N}\cup\{0\}$. We define an infinite random sequence $\{\tau_k : k \in \mathbb{N}_0\} \subset [0,\infty)$ with $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$, which will record the time at which the k-th update of SGD occurs. The distribution of these $\{\tau_k : k \in \mathbb{N}_0\}$ will follow a standard rate-$(\frac{n}{\beta})$ Poisson process. This means that the family of interarrival times $\{\tau_k - \tau_{k-1} : k \in \mathbb{N}\}$ are i.i.d. Exp$(\frac{n}{\beta})$ random variables, *i.e.*, those with mean $\frac{\beta}{n}$, and we note that this randomization is independent of both the SGD, $\boldsymbol{A}$ and $\boldsymbol{b}$. The function N_t will count the number of arrivals of this Poisson process before time t, that is

$$N_t = \sup\{k \in \mathbb{N}_0 \,:\, \tau_k \le t\}.$$

Then for any $t > 0$, N_t has the distribution of Poisson$(\frac{n}{\beta}t)$.

We embed the process $\widehat{\boldsymbol{\nu}}$ into continuous time, by taking $\boldsymbol{\nu}_t = \widehat{\boldsymbol{\nu}}_{\tau_{N_t}}$. We note that we have scaled time (by choosing the rate of the Poisson process) so that in a single unit of time t, the algorithm has done one complete pass (in expectation) over the data set, i.e. SGD has completed one epoch.

B.3 Doob-Meyer decomposition

We compute the Doob decomposition for quasi-martingales [Protter, 2005, Thm. 18, Chpt. 3] of $\boldsymbol{\nu}_t$ and of $\nu_{t,j}^2$, where $\nu_{t,j}$ is the j-th coordinate of $\boldsymbol{\nu}_t$ where j ranges from $j = 1, \dots, d$. Here we let $\mathcal{F}_t$ be the σ-algebra of information available to the process at time $t \geq 0$. So we take, for any $j \in [d]$,

$$\begin{aligned}\boldsymbol{\mathcal{B}}_t &\stackrel{\text{def}}{=} \partial_t \, \mathbb{E}[\boldsymbol{\nu}_t \mid \mathcal{F}_t] = \lim_{\epsilon \downarrow 0} \epsilon^{-1} \, \mathbb{E}[\boldsymbol{\nu}_{t+\epsilon} - \boldsymbol{\nu}_t \mid \mathcal{F}_t] \\ \mathcal{A}_{t,j} &\stackrel{\text{def}}{=} \partial_t \, \mathbb{E}[\nu_{t,j}^2 \mid \mathcal{F}_t] = \lim_{\epsilon \downarrow 0} \epsilon^{-1} \, \mathbb{E}[\nu_{t+\epsilon,j}^2 - \nu_{t,j}^2 \mid \mathcal{F}_t].\end{aligned}$$

In terms of this random variable, we have a decomposition

$$\begin{aligned}\boldsymbol{\nu}_t &= \boldsymbol{\nu}_0 + \int_0^t \boldsymbol{\mathcal{B}}_s \, \mathrm{d}s + \widetilde{\boldsymbol{M}}_t, \\ \nu_{t,j}^2 &= \nu_{0,j}^2 + \int_0^t \mathcal{A}_{s,j} \, \mathrm{d}s + M_{t,j},\end{aligned} \tag{34}$$

where $(M_{t,j}, \widetilde{\boldsymbol{M}_t} : t \geq 0)$ are $\mathcal{F}_t$–adapted martingales.

For the computation of $\mathcal{A}_{t,j}$ we observe that as $\epsilon \to 0$, $\mathcal{A}_{t,j}$ is dominated by the contribution of a single Poisson point arrival; as in time ϵ, the probability of having multiple Poisson point arrivals is $O(\beta^{-2} n^2 \epsilon^2)$, whereas the probability of having a single arrival is $1 - e^{-\beta^{-1} n \epsilon} \sim \beta^{-1} n \epsilon$ as $\epsilon \to 0$. For notational simplicity, we let the projection matrix $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ be an i.i.d. copy of $\boldsymbol{P}_1$, which is independent of all the randomness so far and we let B be the corresponding random subset of $\{1, 2, \dots, n\}$ that defines $\boldsymbol{P}$. It follows

$$\begin{aligned}\boldsymbol{\mathcal{B}}_t &= \frac{n}{\beta} \, \mathbb{E}\left[\boldsymbol{\nu}_t - \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P} (\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{\nu}_t - \boldsymbol{\eta}) - \boldsymbol{\nu}_t \;\middle|\; \mathcal{F}_t \right], \\ \mathcal{A}_{t,j} &= \frac{n}{\beta} \, \mathbb{E}\left[\left(\nu_{t,j} - \gamma \mathbb{e}_j^T \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P} (\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{\nu}_t - \boldsymbol{\eta}) \right)^2 - \nu_{t,j}^2 \;\middle|\; \mathcal{F}_t \right].\end{aligned}$$

The mean term of $\boldsymbol{\nu}_t$ simplifies significantly, and by no accident: by construction, the SGD update rule has a conditional expectation which is proportional to the gradient of the objective function. Observe that since

$$\mathbb{E}[\boldsymbol{P}] = \frac{\beta}{n} \sum_{i=1}^n \mathbb{e}_i \mathbb{e}_i^T = \frac{\beta}{n} \boldsymbol{I},$$

the previous equation simplifies to:

$$\boldsymbol{\mathcal{B}}_t = -\gamma \boldsymbol{\Sigma}^T (\boldsymbol{\Sigma} \boldsymbol{\nu}_t - \boldsymbol{U}^T \boldsymbol{\eta}) \quad \text{and} \quad \mathcal{B}_{t,j} = -\gamma \sigma_j^2 \nu_{t,j} + \gamma \sigma_j (\boldsymbol{U}^T \boldsymbol{\eta})_j. \tag{35}$$

We now turn to the evaluation of $\mathcal{A}_{t,j}$. If we let $\boldsymbol{P} = \sum_{i \in B} \mathbb{e}_i \mathbb{e}_i^T$,

$$\begin{aligned}\mathcal{A}_{t,j} &= -2 \nu_{t,j} \gamma \mathbb{e}_j^T \boldsymbol{\Sigma}^T \boldsymbol{U}^T (\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{\nu}_t - \boldsymbol{\eta}) + \gamma^2 \frac{n}{\beta} \, \mathbb{E}\left(\mathbb{e}_j^T \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P} (\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{\nu}_t - \boldsymbol{\eta}) \;\middle|\; \mathcal{F}_t \right)^2 \\ &= 2 \nu_{t,j} \mathcal{B}_{t,j} + \gamma^2 \frac{n \sigma_j^2}{\beta} \, \mathbb{E}\left(\sum_{i \in B} \left(\mathbb{e}_j^T \boldsymbol{U}^T \mathbb{e}_i \right) \left(\mathbb{e}_i^T (\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{\nu}_t - \boldsymbol{\eta}) \right) \;\middle|\; \mathcal{F}_t \right)^2.\end{aligned} \tag{36}$$

To compute this conditional expectation, we record the following lemma.

Lemma B.1. *Suppose that $\boldsymbol{u}$ and $\boldsymbol{v}$ are fixed vectors in $\mathbb{R}^n$. Then*

$$\mathbb{E}\left(\sum_{i \in B} u_i v_i \right)^2 = \frac{\beta(\beta - 1)}{n(n-1)} (\boldsymbol{u}^T \boldsymbol{v})^2 + \left(\frac{\beta}{n} - \frac{\beta(\beta-1)}{n(n-1)} \right) \sum_{i=1}^n (u_i v_i)^2.$$

Proof. This reduces to the two probabilities:

$$\Pr(i \in B) = \frac{\beta}{n}, \quad \text{and} \quad \Pr(i, \ell \in B) = \frac{\beta(\beta-1)}{n(n-1)},$$

where $i \neq \ell$ are any fixed numbers in $\{1, 2, \ldots, n\}$. The proof now follows by expanding both sides. □

Using Lemma B.1, we can therefore simplify (36) by writing

$$\mathcal{A}_{t,j} = 2\nu_{t,j}\mathcal{B}_{t,j} + \frac{\beta-1}{n-1}\big(\mathcal{B}_{t,j}\big)^2 + \gamma^2\sigma_j^2\left(1 - \frac{\beta-1}{n-1}\right)\sum_{i=1}^{n}\big(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\big)^2\big(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})\big)^2. \tag{37}$$

B.4 Constructing the approximate Volterra equation

In this section, we derive an approximate Volterra equation. First, we can write the function values at the iterates $\boldsymbol{x}_t$ in terms of $\boldsymbol{\nu}_t$,

$$\begin{aligned}\psi_\varepsilon(t) \stackrel{\text{def}}{=} f(\boldsymbol{x}_{N_t}) &= \frac{1}{2n}\|\boldsymbol{A}(\boldsymbol{x}_{N_t} - \widetilde{\boldsymbol{x}}) - \sqrt{n}\,\boldsymbol{\eta}\|^2 = \frac{1}{2}\|\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta}\|^2. \\ &= \frac{1}{2}\sum_{j=1}^{d}\sigma_j^2\nu_{t,j}^2 - \sum_{j=1}^{n\wedge d}\sigma_j\nu_{t,j}\,(\boldsymbol{U}^T\boldsymbol{\eta})_j + \frac{1}{2}\|\boldsymbol{\eta}\|^2.\end{aligned} \tag{38}$$

Hence the dynamics of $\psi_\varepsilon(t)$ are governed by the behaviors of $\nu_{t,j}$ and $\nu_{t,j}^2$ processes. We now return to $\mathcal{A}_{t,j}$. The *key lemma* to simplifying (37) is that the $\big(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\big)^2$ expression in (37) self-averages to $\frac{1}{n}$. Furthermore, we are working in the regime when $\beta = o(n)$, and hence the terms with $\frac{\beta}{n}$ will vanish in the large-n limit. Thus we define

$$\widehat{\mathcal{A}}_{t,j} \stackrel{\text{def}}{=} 2\nu_{t,j}\mathcal{B}_{t,j} + \gamma^2\sigma_j^2\sum_{i=1}^{n}\frac{1}{n}\Big(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})\Big)^2 = -\gamma 2\sigma_j^2\nu_{t,j}^2 + \gamma 2\sigma_j\nu_{t,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j + \gamma^2\frac{2\sigma_j^2\psi_\varepsilon(t)}{n}. \tag{39}$$

We will show that $\widehat{\mathcal{A}}_{t,j}$ is a good approximation for $\mathcal{A}_{t,j}$ in a suitably strong sense so that we can derive a deterministic Volterra equation description for $\psi_\varepsilon(t)$ in the large-n limit. For the moment, let's group these error terms together. Define the càdlàg process

$$\mathcal{E}_{t,j} \stackrel{\text{def}}{=} \nu_{t,j}^2 - \nu_{0,j}^2 - \int_0^t \widehat{\mathcal{A}}_{s,j}\,\mathrm{d}s. \tag{40}$$

In the following lemma, we get an expression for $\nu_{t,j}$.

Lemma B.2. *For any $t \geq 0$ and for any $1 \leq j \leq d$,*

$$\begin{aligned}\nu_{t,j}^2 &= e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 + \int_0^t e^{-2(t-s)\gamma\sigma_j^2}\left(\big(\gamma 2\sigma_j\nu_{s,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j + \tfrac{\gamma^2 2\sigma_j^2\psi_\varepsilon(s)}{n}\big)\,\mathrm{d}s + \mathrm{d}\mathcal{E}_{s,j}\right) \\ \textit{and} \qquad \nu_{t,j} &= e^{-\gamma\sigma_j^2 t}\nu_{0,j} + \int_0^t e^{-\gamma\sigma_j^2(t-s)}\big(\gamma\sigma_j(\boldsymbol{U}^T\boldsymbol{\eta})_j\,\mathrm{d}s + \mathrm{d}\widetilde{M}_{s,j}\big).\end{aligned}$$

Proof. We show the first equation. The second follows by a similar argument. Using the definition of $\mathcal{E}_{t,j}$, the following holds

$$\nu_{t,j}^2 = \nu_{0,j}^2 + \int_0^t\left(-\gamma 2\sigma_j^2\nu_{s,j}^2 + \gamma 2\sigma_j\nu_{s,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j + \gamma^2\frac{2\sigma_j^2\psi_\varepsilon(s)}{n}\right)\mathrm{d}s + \mathcal{E}_{t,j}.$$

Using càdlàg differentiation, we get that

$$\mathrm{d}(e^{2t\gamma\sigma_j^2}\nu_{t,j}^2) = e^{2t\gamma\sigma_j^2}\gamma 2\sigma_j\nu_{t,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j + e^{2t\gamma\sigma_j^2}\gamma^2\frac{2\sigma_j^2\psi_\varepsilon(t)}{n} + e^{2t\gamma\sigma_j^2}\,\mathrm{d}\mathcal{E}_{t,j}.$$

Hence integrating both sides, one obtains

$$\nu_{t,j}^2 = e^{-2t\gamma\sigma_j^2}\left(\nu_{0,j}^2 + \int_0^t e^{2s\gamma\sigma_j^2}\left(\gamma 2\sigma_j\nu_{s,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j + \gamma^2\frac{2\sigma_j^2\psi_\varepsilon(s)}{n}\right)\mathrm{d}s + \int_0^t e^{2s\gamma\sigma_j^2}\,\mathrm{d}\mathcal{E}_{s,j}\right),$$

which completes the proof. □

We then apply Lemma B.2 to (38) and we derive the *approximate Volterra equation*

$$\begin{aligned}\psi_\varepsilon(t) = &\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2\left(e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 + \int_0^t e^{-2(t-s)\gamma\sigma_j^2}\gamma^2\frac{2\sigma_j^2\psi_\varepsilon(s)}{n}\,\mathrm{d}s\right)\\ &+\frac{1}{2}\sum_{j=1}^{d}\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\gamma 2\sigma_j^3\nu_{s,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j\,\mathrm{d}s + \frac{1}{2}\|\boldsymbol{\eta}\|^2 - \sum_{j=1}^{n\wedge d}\sigma_j\nu_{t,j}\,(\boldsymbol{U}^T\boldsymbol{\eta})_j\\ &+\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\,\mathrm{d}\mathcal{E}_{s,j}.\end{aligned} \tag{41}$$

In this expression, we have gathered the terms on each line that have different limit behaviors. On the first line, we have the terms, that due to the convergence of the empirical measure of singular values (Assumption 1.2), will have continuum limits. The second line are those terms that survive in the limit due to the effect of noise $\boldsymbol{\eta}$. The third line are error terms that vanish in the limit. We will make explicit the convergence in the first two lines in the following lemmata.

B.5 Stability of the Volterra equation

We begin by defining some Laplace transforms of the limiting spectral measures, for $k \in \mathbb{N}_0$,

$$h_k(t) \stackrel{\text{def}}{=} \int_0^\infty x^k e^{-2\gamma t x}\,\mathrm{d}\mu(x). \tag{42}$$

We begin by showing that the terms on the first line of (41) converge to some finite limit. Under our Assumption 1.2,

Lemma B.3. *Locally uniformly on compact sets of time,*

$$\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 \xrightarrow[n\to\infty]{\Pr} \frac{Rh_1(t)}{2} \quad \textit{and} \tag{43}$$

$$\sum_{j=1}^{d}\gamma^2\frac{\sigma_j^4}{n}e^{-2t\gamma\sigma_j^2} \xrightarrow[n\to\infty]{\Pr} \gamma^2 r h_2(t). \tag{44}$$

Proof. We begin by showing that each term in (43) and (44) converge pointwise in probability. Note that the convergence of (44) is trivial because of Assumption 1.2. Hence, it only remains to show pointwise convergence of (43).

Under Assumption 1.1 and using uniform distribution of $\boldsymbol{V}$ we have that $\boldsymbol{\nu}_0 = \boldsymbol{V}(\boldsymbol{x}_0 - \widetilde{\boldsymbol{x}})$ is a uniformly distributed vector on the sphere of norm $\sqrt{R}$. Observe first that conditioned on $\boldsymbol{\Sigma}$, the conditional expectation of the LHS of (43) is given by

$$\mathbb{E}\left[\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2\,\middle|\,\boldsymbol{\Sigma}\right] = \frac{R}{2}\frac{1}{d}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}.$$

The vector $\boldsymbol{\nu}_0^2$ follows the Dirichlet distribution, which is negatively associated. In particular, $\mathbb{E}(\nu_{0,j}^2\nu_{0,j}^2) \le \mathbb{E}(\nu_{0,j}^2)\mathbb{E}(\nu_{0,j}^2)$. Further $\mathbb{E}(\nu_{0,j}^4) \le 3R^2d^{-2}$, as the moments are strictly bounded by the normal moments. Hence, the variance is

bounded by

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2\middle|\mathbf{\Sigma}\right) &= \mathbb{E}\left[\left(\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 - \frac{R}{2}\frac{1}{d}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\right)^2\middle|\mathbf{\Sigma}\right] \\ &= \mathbb{E}\left[\left(\frac{1}{2d}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}(d\nu_{0,j}^2 - R)\right)^2\middle|\mathbf{\Sigma}\right] \\ &\leq \frac{1}{4d^2}\mathbb{E}\left[\sum_{j=1}^{d}(\sigma_j^2 e^{-2t\gamma\sigma_j^2})^2(d\nu_{0,j}^2 - R)^2\middle|\mathbf{\Sigma}\right] \\ &= \frac{1}{4d}\left[\frac{1}{d}\sum_{j=1}^{d}\sigma_j^4 e^{-2\cdot 2t\gamma\sigma_j^2}\right](d^2\mathbb{E}[\nu_{0,j}^4] - R^2) = \mathcal{O}\left(\frac{1}{d}\right).\end{aligned}$$

Therefore, for $\epsilon > 0$, conditional Chebyshev inequality gives

$$\Pr\left(\left|\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 - \frac{R}{2}\frac{1}{d}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\right| > \epsilon\middle|\mathbf{\Sigma}\right) \leq \frac{1}{\epsilon^2}\,\mathrm{Var}\left(\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2\middle|\mathbf{\Sigma}\right) \xrightarrow[n\to\infty]{} 0.$$

Applying the law of total probability to this and combining it with the weak convergence of ESM in probability (Assumption 1.2) gives

$$\begin{aligned}&\Pr\left(\left|\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 - \frac{Rh_1(t)}{2}\right| > \epsilon\right) \\ &\qquad\leq \Pr\left(\left|\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 - \frac{R}{2}\frac{1}{d}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\right| > \frac{\epsilon}{2}\right) \\ &\qquad+ \Pr\left(\left|\frac{R}{2}\frac{1}{d}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2} - \frac{Rh_1(t)}{2}\right| > \frac{\epsilon}{2}\right) \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Now we show the uniform convergence of (43) on time interval $[0, T]$ for a fixed time $T > 0$ (the same argument applies to (44)). Considering mesh points on $[0, T]$ with spacing, let us say $\lambda > 0$, we can say that the pointwise convergence holds on those mesh points and so does the supremum convergence on them. For an arbitrary time $t \in [0, T]$, there exists a mesh point t_0 such that $|t - t_0| \leq \lambda$. Then, since $e^{-2t\gamma\sigma_j^2}$ is a Lipschitz function on $[0, T]$ with some Lipschitz constant $C > 0$, we have

$$\sup_{0\leq t\leq T}\left|\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2(e^{-2t\gamma\sigma_j^2} - e^{-2t_0\gamma\sigma_j^2})\nu_{0,j}^2\right| \leq \frac{C\lambda}{2}\sum_{j=1}^{d}\sigma_j^2\nu_{0,j}^2.$$

Note that $\sum_{j=1}^{d}\sigma_j^2\nu_0^2 \xrightarrow[n\to\infty]{\Pr} R\int_0^\infty x^2 d\mu(x) < \infty$ using a similar idea by conditioning on $\mathbf{\Sigma}$ and applying Assumption 1.2. Then observe, applying triangle inequality and taking supremum on $t \in [0, T]$ gives

$$\begin{aligned}\sup_{0\leq t\leq T}\left|\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 - \frac{Rh_1(t)}{2}\right| &\leq \sup_{t_0\in[0,T]}\left|\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 e^{-2t_0\gamma\sigma_j^2}\nu_{0,j}^2 - \frac{Rh_1(t_0)}{2}\right| \\ &+ \sup_{0\leq t\leq T}\left|\frac{1}{2}\sum_{j=1}^{d}\sigma_j^2(e^{-2t\gamma\sigma_j^2} - e^{-2t_0\gamma\sigma_j^2})\nu_{0,j}^2\right| \\ &+ \sup_{t,t_0\in[0,T]}\frac{R}{2}\left|h_1(t_0) - h_1(t)\right|.\end{aligned}$$

Given that μ has a finite support, we have $\sup_{t,t_0\in[0,T]}|h_1(t) - h_1(t_0)| \leq C'\lambda$ for some $C' > 0$. Now the claim follows as λ can be chosen as small as possible. □

We can now recast (41) as an approximate Volterra type integral equation, where

$$\psi_\varepsilon(t) = \frac{Rh_1(t)}{2} + \widetilde{R} \cdot \frac{rh_0(t) + (1-r)}{2} + \varepsilon_1^{(n)}(t) + \int_0^t \left(\gamma^2 r h_2(t-s) + \varepsilon_2^{(n)}(t-s)\right)\psi_\varepsilon(s)\ \mathrm{d}s, \tag{45}$$

and where $\varepsilon_i^{(n)}$ are defined implicitly by comparison with (41). In particular, $\varepsilon_2^{(n)}(t)$ is given by the difference

$$\varepsilon_2^{(n)}(t) = \sum_{j=1}^d \gamma^2 \frac{\sigma_j^4}{n} e^{-2t\gamma\sigma_j^2} - \gamma^2 r h_2(t),$$

which is therefore guaranteed to converge to 0 by Lemma B.3. The other error $\varepsilon_1^{(n)}$ is substantially more complicated; we discuss it fully in (54), Section B.6.

However, all we need to show is that this error tends to 0, as Volterra equations are stable:

Proposition B.1 (Stability of the Volterra equation)**.** *Fix a constant $T > 0$ and suppose ψ_ε solves (45) with bounded error terms,*

$$\sup_{0\le t\le T} |\varepsilon_1^{(n)}(t)| \xrightarrow[n\to\infty]{\Pr} 0 \quad \textit{and} \quad \sup_{0\le t\le T} |\varepsilon_2^{(n)}(t)| \xrightarrow[n\to\infty]{\Pr} 0.$$

Suppose that ψ_0 solves (45) with $\varepsilon_1^{(n)} = \varepsilon_2^{(n)} = 0$. Then for any fixed length of time T, the perturbed solution of the Volterra equation ψ_ε converges uniformly, in probability, to the unperturbed solution of the Volterra equation, $\psi_0(t)$:

$$\sup_{0\le t\le T} |\psi_\varepsilon(t) - \psi_0(t)| \xrightarrow[n\to\infty]{\Pr} 0.$$

Proof. Throughout this proof, we set $\mathbb{R}^+ = \{x \in \mathbb{R} : x \ge 0\}$ and we use the notation $L^1_{\text{loc}}(\mathbb{R}^+)$ to be the locally integrable functions on $\mathbb{R}^+$. We will supress the n dependency in the error terms $\varepsilon_1^{(n)}$ and $\varepsilon_2^{(n)}$. We begin by defining some notation for solving Volterra equations, namely the kernel and forcing function respectively by

$$k_\varepsilon(t) \overset{\text{def}}{=} -\big(\gamma^2 r h_2(t) + \varepsilon_2(t)\big) \qquad \text{and} \qquad f_\varepsilon(t) \overset{\text{def}}{=} \frac{Rh_2(t)}{2} + \widetilde{R} \cdot \frac{rh_0(t) + 1 - r}{2} + \varepsilon_1(t). \tag{46}$$

Here we use the convention that k_0 and f_0 correspond to where $\varepsilon_1(t) = \varepsilon_2(t) = 0$. Under this notation, the Volterra equation in (45) becomes

$$\psi_\varepsilon(t) + \int_0^t k_\varepsilon(t-s)\psi_\varepsilon(s)\ \mathrm{d}s = f_\varepsilon(t). \tag{47}$$

Now we check that $k_\varepsilon(t) \in L^1_{\text{loc}}(\mathbb{R}^+)$ with high probability. To see this we only need that $h_2(t) \in L^1_{\text{loc}}(\mathbb{R}^+)$ as the supremum condition on $\varepsilon_2(t)$ guarantees that the error term in k_ε is bounded with high probability and therefore $\varepsilon_2(t)$ is in $L^1_{\text{loc}}(\mathbb{R}^+)$ with high probability. Since $h_2(t) \ge 0$, we can apply Tonelli's theorem

$$\int_0^\infty h_2(t)\ \mathrm{d}t = \int_0^\infty \int_0^\infty x^2 e^{-2\gamma t x}\ \mathrm{d}t\ \mathrm{d}\mu(x) = \int_0^\infty \frac{x}{2\gamma}\,\mathrm{d}\mu(x) < \infty.$$

Here we used that $\mu(x)$ is compactly supported to conclude the last integral. Hence it follows that $h_2(t) \in L^1(\mathbb{R}^+)$ which shows that $k_\varepsilon(t) \in L^1_{\text{loc}}(\mathbb{R}^+)$. To prove the conclusion of the proposition, we will use a stability theorem together with the existence and uniqueness for Volterra equations of convolution type kernels. The solutions of convolution kernel Volterra equations rely on a function defined through the kernel k called the *resolvent of the kernel k*. We define this resolvent as the function $r_\varepsilon : \mathbb{R}^+ \to \mathbb{R}$ such that

$$r_\varepsilon(t) = \sum_{j=1}^\infty (-1)^{j-1} k_\varepsilon^{*j}(t), \tag{48}$$

where the function $k_\varepsilon^{*j}(t)$, $j \ge 1$ is the $(j-1)$-fold convolution of the kernel k_ε with itself. We want to show that a perturbed kernel k_ε results in a perturbation of the resolvent. Since $k_\varepsilon \in L^1_{\text{loc}}(\mathbb{R}^+)$, the stability theorem for kernels,

Theorem 3.1 in Gripenberg et al. [1990], says that the resolvent $r_\varepsilon \in L^1_{\text{loc}}(\mathbb{R}^+)$ is unique and depends continuously on k_ε in the $L^1_{\text{loc}}(\mathbb{R}^+)$ topology. As $\sup_{0\le t\le T} |\varepsilon_2(t)| \xrightarrow[n\to\infty]{\Pr} 0$, we have that

$$\int_0^T |k_\varepsilon(t) - k_0(t)| \, \mathrm{d}t = \int_0^T |\varepsilon_2(t)| \, \mathrm{d}t \xrightarrow[n\to\infty]{\Pr} 0,$$

so by continuity in L^1_{loc}, we get that

$$\int_0^T |r_\varepsilon(t) - r_0(t)| \, \mathrm{d}t \xrightarrow[n\to\infty]{\Pr} 0.$$

Using this resolvent, the unique solution to the Volterra equation in (47) [Gripenberg et al., 1990, Theorem 3.5] is give by

$$\psi_\varepsilon(t) = f_\varepsilon(t) - (r_\varepsilon * f_\varepsilon)(t). \tag{49}$$

A simple computation yields that

$$\begin{aligned}
|\psi_\varepsilon(t) - \psi_0(t)| &\le |f_\varepsilon(t) - f_0(t)| + \Big| \int_0^t (r_0 - r_\varepsilon)(t-s) f_0(s) \, \mathrm{d}s \Big| \\
&\quad + \Big| \int_0^t r_\varepsilon(t-s)[f_0(s) - f_\varepsilon(s)] \, \mathrm{d}s \Big| \\
&\le |f_\varepsilon(t) - f_0(t)| + \sup_{0\le s\le t} |f_0(t)| \int_0^t |(r_0 - r_\varepsilon)|(t-s) \, \mathrm{d}s \\
&\quad + \sup_{0\le s\le t} |f_0(s) - f_\varepsilon(s)| \int_0^t |r_\varepsilon(t-s)| \, \mathrm{d}s.
\end{aligned}$$

Since $h_k(t)$ is bounded, we clearly have that $\sup_{0\le t\le T} |f_0(t)|$ is bounded. Working on the event that $r_\varepsilon(t) \in L^1_{\text{loc}}(\mathbb{R}^+)$ (*i.e.* $\int_0^T |r_\varepsilon(t)| \, dt$ is bounded), $\int_0^T |r_\varepsilon(t) - r_0(t)| \, dt$ is small, and $\sup_{0\le t\le T} |\varepsilon_1(t)|$ is small, it follows that

$$\begin{aligned}
\sup_{0\le t\le T} |\psi_\varepsilon(t) - \psi_0(t)| &\le \sup_{0\le t\le T} |\varepsilon_1(t)| + \sup_{0\le t\le T} |f_0(t)| \int_0^T |r_0 - r_\varepsilon|(t) \, \mathrm{d}t \\
&\quad + \sup_{0\le t\le T} |\varepsilon_1(t)| \int_0^T |r_\varepsilon(t)| \, \mathrm{d}t.
\end{aligned}$$

Since every term on the RHS is small and the complement of the event on which we proved the inequality above has small probability, the result immediately follows. □

This yields one of the main theorems of this paper which we restate for clarity (Theorem 1.1):

Theorem B.1 (Concentration of SGD)**.** *Suppose $\beta \in \mathbb{N}$ is a batch-size parameter such that $0 < \beta \le n^{1/5-\delta}$ for some $\delta > 0$ and the stepsize is $\gamma < \frac{2}{r}\big(\int_0^\infty x \, d\mu(x)\big)^{-1}$. Let the constant $T > 0$. Under Assumptions 1.1 and 1.2, the function values at the iterates of SGD converge to*

$$\sup_{0\le t\le T} \big| f\big(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}\big) - \psi_0(t) \big| \xrightarrow[n\to\infty]{\Pr} 0, \tag{50}$$

where the function ψ_0 is the solution to the Volterra equation

$$\begin{aligned}
\psi_0(t) &= \tfrac{R}{2} h_1(t) + \tfrac{\widetilde{R}}{2}\big(r h_0(t) + (1-r)\big) + \int_0^t \gamma^2 r h_2(t-s) \psi_0(s) \, \mathrm{d}s, \\
&\text{and} \quad h_k(t) = \int_0^\infty x^k e^{-2\gamma t x} \, \mathrm{d}\mu(x).
\end{aligned} \tag{51}$$

Proof. By definition of N_t and $\psi_\varepsilon(t)$, we have

$$f\big(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}\big) = f\big(\boldsymbol{x}_{N_{\tau_{\lfloor n\beta/t \rfloor}}}\big) = \psi_\varepsilon(\tau_{\lfloor \frac{n}{\beta} t \rfloor}).$$

Also, Proposition B.1 gives

$$\sup_{0\le t\le T} |\psi_\varepsilon(t) - \psi_0(t)| \xrightarrow[n\to\infty]{\Pr} 0.$$

Therefore, triangle inequality gives

$$\sup_{0\le t\le T} \big|f\big(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}\big) - \psi_0(t)\big| \le \sup_{0\le t\le T} \big|\psi_\varepsilon(\tau_{\lfloor \frac{n}{\beta} t \rfloor}) - \psi_0(\tau_{\lfloor \frac{n}{\beta} t \rfloor})\big| + \sup_{0\le t\le T} \big|\psi_0(\tau_{\lfloor \frac{n}{\beta} t \rfloor}) - \psi_0(t)\big|,$$

and by the continuity of $\psi_0(t)$, it would suffice to show

$$\sup_{0\le t\le T} \big|\tau_{\lfloor \frac{n}{\beta} t \rfloor} - t\big| \xrightarrow[n\to\infty]{\Pr} 0. \tag{52}$$

First, note that

$$\sup_{0\le s\le T} \big|N_s \frac{\beta}{n} - s\big| \xrightarrow[n\to\infty]{\Pr} 0 \tag{53}$$

holds. For a fixed time $s \in [0,T]$, this comes from the strong law of large numbers, see [Kingman, 1993, (4.18)]. And the result for the supremum on $[0,T]$ follows using monotonicity of N_t on $[0,T]$ and the meshing arguments as used in proving Lemma B.3.

Now for $t > 0$, let $s > 0$ be such that $\lfloor \frac{n}{\beta} t \rfloor = N_s$, or $t = N_s \frac{\beta}{n} + \frac{\beta}{n} r$ for $0 \le r < 1$. Therefore, observe

$$\big|\tau_{\lfloor \frac{n}{\beta} t \rfloor} - t\big| = \big|\tau_{N_s} - N_s \frac{\beta}{n} - \frac{\beta}{n} r\big| \le \big|\tau_{N_s} - s\big| + \big|N_s \frac{\beta}{n} - s\big| + \frac{\beta}{n} r.$$

The last term converges to 0 as $n \to \infty$, and so does the second term in probability, by (53). So, it is left to show the convergence in probability of the first term. By the definition of N_s, we have

$$s - \Delta \le \tau_{N_s} \le s,$$

where Δ denotes the largest spacing between adjacent jumps in [0,T]. Note that $N_s \le N_T \le \frac{2Tn}{\beta}$ with overwhelming probability [Klar, 2000, Prop. 1]. Recalling again that $\tau_k - \tau_{k-1}, k \in \mathbb{N}$, follow $\mathrm{Exp}(\frac{n}{\beta})$ independently on $k \in \mathbb{N}$, we have for $u > 0$,

$$\begin{aligned}
\Pr\big((\Delta > u) \cap (N_T \le \frac{2Tn}{\beta})\big) &= 1 - \Pr\big((\Delta \le u) \cap (N_T \le \frac{2Tn}{\beta})\big) \\
&\le 1 - (1 - e^{-\frac{n}{\beta} u})^{N_T} \\
&\le N_T e^{-\frac{n}{\beta} u} \le \frac{2Tn}{\beta} e^{-\frac{n}{\beta} u}.
\end{aligned}$$

This implies

$$\Pr\big(\Delta > u\big) \le \Pr\big((\Delta > u) \cap (N_T \le \frac{2Tn}{\beta})\big) + \Pr(N_T > \frac{2Tn}{\beta}) \to 0$$

as $n \to \infty$ and we obtain the claim. □

B.6 Bounding the errors $\varepsilon_1^{(n)}$

In this section, we give a high–level overview of the errors and how they converge to 0. We will have the following error pieces:

$$\varepsilon_1^{(n)}(t) \overset{\text{def}}{=} \varepsilon_{\mathrm{IC}}^{(n)}(t) + \varepsilon_{\mathrm{KL}}^{(n)}(t) + \varepsilon_{\mathrm{M}}^{(n)}(t) + \varepsilon_{\mathrm{beta}}^{(n)}(t) + \varepsilon_{\mathrm{eta}}^{(n)}(t). \tag{54}$$

We define these terms momentarily and we will verify that $\varepsilon_1^{(n)}$ is indeed equal to these pieces in Lemma B.6. We remark that before controlling the errors, we will need to make an *a priori* estimate that (effectively) shows the function values remain bounded. Thus, we define the stopping time, for any fixed $\theta > 0$, by

$$\vartheta \overset{\text{def}}{=} \inf\left\{t \ge 0 : \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}\| > n^\theta\right\}. \tag{55}$$

We then show:

Lemma B.4. *For any $\theta > 0$, and for any $T > 0$, $\vartheta > T$ with high probability.*

This is achieved by a simple martingale-type estimate, which is similar to the standard convergence arguments for SGD. The proof is given in Section B.7. We will need it in what follows. *We will also condition on $\boldsymbol{\Sigma}$ going forward.*

B.6.1 Errors from the convergence of the initial conditions

The error $\varepsilon_{\mathrm{IC}}^{(n)}(t)$ arises due to convergence errors in the signal and initialization. It was already essentially discussed in Lemma B.3. We define it by

$$\varepsilon_{\mathrm{IC}}^{(n)}(t) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 e^{-2t\gamma\sigma_j^2}\nu_{0,j}^2 - \frac{Rh_1(t)}{2}. \tag{56}$$

It accounts for the convergence of the initialization in the large d limit and relies on the convergence of the empirical spectral distribution. Due to Lemma B.3, we have already shown it converges to 0.

B.6.2 Errors which vanish due to the key lemma

The vanishing of the error $\varepsilon_{\mathrm{KL}}^{(n)}(t)$ is the *key lemma*. To explain why we call it this: let us specialize to the case of $\boldsymbol{\eta} = 0$ and $\beta = 1$. If we were content to evaluate the *expected* function values, when averaging over the randomness inherent in the SGD algorithm, then this would be the only error that we would need to control. Thus in some sense, it can be viewed as the minimal estimate that needs to be shown to prove the Volterra equation holds. This error is given by

$$\varepsilon_{\mathrm{KL}}^{(n)}(t) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 \int_0^t e^{-2(t-s)\gamma\sigma_j^2}\left(\sum_{i=1}^{n}\left(\left(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\right)^2 - \frac{1}{n}\right)\left(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_s - \boldsymbol{\eta})\right)^2\right)\,\mathrm{d}s. \tag{57}$$

After interchanging the order of summation, it suffices to show:

Lemma B.5 (Key lemma)**.** *For any $T > 0$ and for any $\epsilon > 0$, with overwhelming probability*

$$\max_{1\le i\le n}\max_{0\le t\le T}\left|\sum_{j=1}^{d}\sigma_j^2 e^{-2t\gamma\sigma_j^2}\left(\left(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\right)^2 - \frac{1}{n}\right)\right| \le n^{\epsilon-1/2}.$$

This we show in Section C.2. Note that by combining this with (57) and Lemma B.4, we conclude that for any $\epsilon > 0$ and any T, with high probability

$$\max_{0\le t\le T}|\varepsilon_{\mathrm{KL}}^{(n)}(t)| \le n^{2\epsilon-1/2}\int_0^T \frac{1}{2}\,\mathrm{d}s \to 0.$$

B.6.3 Martingale errors

The martingale errors are due to the randomness in the algorithm itself. They in part are small because the singular vector matrix $\boldsymbol{U}$ is delocalized, in that its offdiagonal entries in any fixed orthogonal basis are $n^{\epsilon-1/2}$ with overwhelming probability. The martingale errors are given by

$$\varepsilon_{\mathrm{M}}^{(n)}(t) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{j=1}^{d}\sigma_j^2\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\,\mathrm{d}M_{s,j}. \tag{58}$$

Estimating this error requires a substantial build-up. The most important technical input, which we will use in multiple places, is that the function values do not concentrate too heavily in any coordinate direction. As an input, we will use Lemma B.4, and so we work with the stopped process defined for any $t \ge 0$ by $\boldsymbol{\nu}_t^{\vartheta} \stackrel{\text{def}}{=} \boldsymbol{\nu}_{t\wedge\vartheta}$. In some sense, this is the most challenging and important technical statement that we prove:

Proposition B.2. *For any $T > 0$, any $\epsilon > 0$, there is a sufficiently small $\theta > 0$ so that*

$$\sup_{0\le t\le T}\sup_{1\le i\le n}\left(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t^{\vartheta} - \boldsymbol{\eta})\right)^2 \le \beta n^{\epsilon-1}$$

with overwhelming probability.

We expect that the upper bound on β in Theorem B.1 is a limitation of our method, and that similar statements should hold for larger β. This proposition is proven in Section D.2.

With Proposition B.2 in hand, we can then bound the martingale errors.

Proposition B.3. *For any $T > 0$, with overwhelming probability,*

$$\sup_{0 \le t \le T} |\varepsilon_{\mathrm{M}}^{(n)}(t \wedge \vartheta)| \xrightarrow{\Pr} 0.$$

This is proven in Section D.2. Having done Proposition B.2, the proof of Proposition B.3 reduces to standard martingale techniques.

B.6.4 Errors due to minibatching

The Volterra equation that we prove (12) importantly does not depend on the minibatching size. Naturally, the dynamics do depend on β, and so there are error terms which must be controlled and which are in part small due to the minibatching parameter β satisfying $\beta/n \to 0$. These errors are given by

$$\varepsilon_{\mathrm{beta}}^{(n)}(t) \stackrel{\mathrm{def}}{=} \frac{1}{2} \sum_{j=1}^{d} \sigma_j^2 \int_0^t e^{-2(t-s)\gamma\sigma_j^2} \left(\frac{\beta - 1}{n-1} \mathcal{B}_{s,j}^2 - \frac{\beta - 1}{n-1} \sum_{i=1}^{n} \left(\mathbb{e}_j^T \boldsymbol{U}^T \mathbb{e}_i \right)^2 \left(\mathbb{e}_i^T (\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{\nu}_s - \boldsymbol{\eta}) \right)^2 \right) \mathrm{d}s. \tag{59}$$

Note in particular that when $\beta = 1$ this vanishes identically.

Much of this error term is controlled using delocalization of $\boldsymbol{U}$ and Lemma B.4. However, there is one error $\beta(\mathcal{B}_{s,j})^2$ which requires extra work. This we would like to tend to 0 in a sufficiently strong sense. On consideration of (35), we see that this in turn requires that $\boldsymbol{\nu}_s$ itself be delocalized in the sense that $|\boldsymbol{\nu}_{s,j}| \le n^{\epsilon - 1/2}$ with overwhelming probability.

Proposition B.4. *For any $\epsilon > 0$ and $T > 0$, with overwhelming probability,*

$$\sup_{0 \le t \le T} \max_{1 \le j \le d} |\nu_{t,j}^{\vartheta}| \le \beta n^{\epsilon - 1/2}.$$

The dependence on β is only through Proposition B.2, on which this relies. The proof is found in Section D.2. Now Proposition B.4 with eigenvector delocalization gives the following proposition.

Proposition B.5. *For any $\epsilon > 0$ and any $T > 0$, with overwhelming probability,*

$$\sup_{0 \le t \le T} |\varepsilon_{\mathrm{beta}}^{(n)}(t)| \le n^{\epsilon - 1/2}.$$

This is proven in Section C.3.

B.6.5 Errors due to the model noise

Finally, there are errors that arise due to the noise $\boldsymbol{\eta}$. The model noise $\boldsymbol{\eta}$ in fact induces a change in the dynamics of the algorithm. This change is reflected in an additional forcing term that appears in the Volterra equation. This forcing term is controlled (in some sense) by the mean behavior of $\nu_{t,j}$. The model noise error is defined by

$$\varepsilon_{\mathrm{eta}}^{(n)}(t) \stackrel{\mathrm{def}}{=} \sum_{j=1}^{d} \int_0^t e^{-2(t-s)\gamma\sigma_j^2} \gamma \sigma_j^3 \nu_{s,j} (\boldsymbol{U}^T \boldsymbol{\eta})_j \, \mathrm{d}s + \frac{1}{2} \|\boldsymbol{\eta}\|^2 - \sum_{j=1}^{n \wedge d} \sigma_j \nu_{t,j} (\boldsymbol{U}^T \boldsymbol{\eta})_j - \widetilde{R} \cdot \frac{r h_0(t) + (1 - r)}{2}. \tag{60}$$

The fundamental identity that needs to be shown here is that averages of $\nu_{s,j}(\boldsymbol{U}^T \boldsymbol{\eta})_j$ converge. Within $\varepsilon_{\mathrm{eta}}^{(n)}(t)$ there are many such averages, and so we formulate a general claim to this effect.

Proposition B.6. *Let $\{c_j\}_1^n$ be a deterministic sequence with $|c_j| \le 1$ for all j and define ϑ as in Lemma B.4. Then for any $t > 0$ and any $\epsilon > 0$,*

$$\left| \sum_{j=1}^{n} c_j \sigma_j \nu_{t,j}^{\vartheta} (\boldsymbol{U}^T \boldsymbol{\eta})_j - \frac{\|\boldsymbol{\eta}\|^2}{n} \sum_{j=1}^{n} c_j (1 - e^{-(t \wedge \vartheta)\gamma\sigma_j^2}) \right| \le \|\boldsymbol{\eta}\|^2 \sqrt{\beta} n^{\epsilon - 1/2},$$

with overwhelming probability.

This is proven in Section D.2.

Using a mesh argument, and appealing to the convergence of the empirical spectrum, we can then show that $\varepsilon_{\text{eta}}^{(n)}(t)$ tends to 0.

Proposition B.7. *For any $T > 0$,*

$$\max_{0 \le t \le T} |\varepsilon_{\text{eta}}^{(n)}(t)| \xrightarrow[n\to\infty]{\Pr} 0.$$

This is proven in Section C.4.

B.6.6 Verification of (54)

The combination of Propositions B.5, B.3, B.5, and B.7 together show all remaining errors are small in (54). Before proceeding, we connect (54) to the previous sections to demonstrate this is truly the sum of errors that must be controlled.

Lemma B.6. *Equation (54) holds.*

Proof. We recall the approximate Volterra equation in (41). The error $\varepsilon_1^{(n)}(t)$ is defined implicitly in (45). By using (56) and (60), we conclude that

$$\varepsilon_1^{(n)}(t) - \varepsilon_{\text{IC}}^{(n)}(t) - \varepsilon_{\text{eta}}^{(n)}(t) = \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 \int_0^t e^{-2(t-s)\gamma\sigma_j^2}\, \mathrm{d}\mathcal{E}_{s,j}. \tag{61}$$

Recall from (40) and (34)

$$\mathcal{E}_{t,j} = \int_0^t \mathcal{A}_{s,j}\, \mathrm{d}s + M_{t,j} - \int_0^t \widehat{\mathcal{A}}_{s,j}\, \mathrm{d}s.$$

Using (37) and (39), we can express

$$\begin{aligned}
\mathrm{d}\mathcal{E}_{t,j} &= \mathrm{d}M_{t,j} \\
&\quad + \left(\frac{\beta-1}{n-1} \big(\mathcal{B}_{t,j}\big)^2 - \frac{\beta-1}{n-1} \sum_{i=1}^{n} \big(\mathbb{e}_j^T \boldsymbol{U}^T \mathbb{e}_i\big)^2 \big(\mathbb{e}_i^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})\big)^2 \right) \mathrm{d}t \\
&\quad + \left(\sum_{i=1}^{n} \left(\big(\mathbb{e}_j^T \boldsymbol{U}^T \mathbb{e}_i\big)^2 - \frac{1}{n} \right) \big(\mathbb{e}_i^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})\big)^2 \right) \mathrm{d}t.
\end{aligned}$$

Each of these three lines, on substituting into (61), produces $\varepsilon_{\text{M}}^{(n)}(t)$, $\varepsilon_{\text{beta}}^{(n)}(t)$, and $\varepsilon_{\text{KL}}^{(n)}(t)$, respectively. □

B.6.7 Proof organization

We organize the remainder of the proof as follows. We begin by proving Lemma B.4 in Section B.7 using standard martingale techniques. Arguments along this line are well–known in the context of analysis of SGD, and this argument is similar (and in fact easier) than convergence arguments for SGD.

In Section C, we introduce standard machinery for concentration of Lipschitz functions on the orthogonal group. In this section, we then make the error estimates that follow from this type of estimate. In particular we prove that the key lemma, Lemma B.5 holds. We also show Proposition B.5 and Proposition B.7 hold. Note these latter propositions depend on some estimates that require other martingale techniques.

In Section D, we give the estimates that depend heavily on martingale concentration techniques. In Section D.1, we outline the general martingale concentration techniques that we need. These extend general martingale techniques in ways that are appropriate to our setting. In Section D.2, we prove the remaining propositions, beginning with the main technical proposition Proposition B.2. We then give the bounds that prove Propositions B.3, B.4, and B.6.

B.7 An *a priori* bound for the objective function values

Here we combine some of the estimates already developed to give a simple starting bound for the function values $\psi_\varepsilon(t)$ in the proof of Lemma B.4. We will need this starting bound in many of our future estimates. We will do this by constructing an appropriate supermartingale, which we then use to control the evolution of ψ_ε.

Proof of Lemma B.4. For convenience, we will set $\psi_\varepsilon = \psi$. We recall from (38) that

$$\psi(t) = \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 \nu_{t,j}^2 - \sum_{j=1}^{n\wedge d} \sigma_j \nu_{t,j} \left(\boldsymbol{U}^T \boldsymbol{\eta}\right)_j + \frac{1}{2}\|\boldsymbol{\eta}\|^2.$$

Hence using (36) and (35),

$$\begin{aligned}
\lim_{\epsilon\downarrow 0} \epsilon^{-1}\mathbb{E}[\psi(t+\epsilon) - \psi(t) \mid \mathcal{F}_t] &= \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 \mathcal{A}_{t,j} - \sum_{j=1}^{n\wedge d} \sigma_j \mathcal{B}_{t,j} \left(\boldsymbol{U}^T \boldsymbol{\eta}\right)_j \\
&= \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 \left(-2\nu_{t,j}\gamma \mathbb{e}_j^T \boldsymbol{\Sigma}^T \boldsymbol{U}^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}) + \gamma^2 \frac{n}{\beta} \mathbb{E}\left(\mathbb{e}_j^T \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}) \,\middle|\, \mathcal{F}_t \right)^2 \right) \\
&+ \gamma \sum_{j=1}^{n} (\boldsymbol{\eta}^T \boldsymbol{U}\boldsymbol{\Sigma}\mathbb{e}_j) \mathbb{e}_j^T \boldsymbol{\Sigma}^T (\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta}) \\
&= -\gamma (\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta})^T \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T (\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta}) + \gamma^2 \frac{n}{2\beta} \sum_{j=1}^{n} \mathbb{E}\left(\mathbb{e}_j^T \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}) \,\middle|\, \mathcal{F}_t \right)^2.
\end{aligned}$$

Using Lemma B.1, we can give the expression

$$\begin{aligned}
\lim_{\epsilon\downarrow 0} \epsilon^{-1}\mathbb{E}[\psi(t+\epsilon) - \psi(t) \mid \mathcal{F}_t] &= -\gamma (\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta})^T \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T (\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{U}^T\boldsymbol{\eta}) \\
&+ \gamma^2 \frac{\beta - 1}{2(n-1)} \sum_{j=1}^{d} \left(\mathbb{e}_j^T \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T \boldsymbol{U}^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}) \right)^2 \\
&+ \gamma^2 \frac{1}{2}\left(1 - \frac{\beta - 1}{n-1}\right) \sum_{j=1}^{d}\sum_{i=1}^{n} \left(\mathbb{e}_j^T \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T \boldsymbol{U}^T \mathbb{e}_i \right)^2 \left(\mathbb{e}_i^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}) \right)^2.
\end{aligned}$$

All three terms have the interpretation as a quadratic form $\boldsymbol{x}^T \widehat{\boldsymbol{A}} \boldsymbol{x}$ for some matrix $\widehat{\boldsymbol{A}}$ and the vector $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}$. Specifically, we have

$$\widehat{\boldsymbol{A}} = -\gamma \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T + \frac{\gamma^2(\beta-1)}{2(n-1)} \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T + \frac{\gamma^2}{2}\left(1 - \frac{\beta-1}{n-1}\right) \sum_{i=1}^{n} \mathbb{e}_i \mathbb{e}_i^T \|\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\mathbb{e}_i\|^2.$$

As $\boldsymbol{\Sigma}$ is bounded, we can let $\rho_\star$ be the largest eigenvalue of $\widehat{\boldsymbol{A}}$, which is symmetric, and which can be bounded solely in terms of the norm of $\boldsymbol{\Sigma}$. Then we conclude that

$$\lim_{\epsilon\downarrow 0} \epsilon^{-1}\mathbb{E}[\psi(t+\epsilon) - \psi(t) \mid \mathcal{F}_t] \le 2\rho_\star \psi(t).$$

It follows immediately that

$$X_t \stackrel{\text{def}}{=} e^{-2\rho_\star t}\psi(t)$$

is a positive supermartingale. Hence by optional stopping, for any $T > 0$

$$\Pr[\sup_{0\le t\le T} X_t \ge \lambda \mid \mathcal{F}_0] \le \frac{\psi(0)}{\lambda}.$$

Hence,

$$\Pr[\sup_{0\le t\le T} \psi(t) \ge \lambda \mid \mathcal{F}_0] \le \frac{\psi(0)e^{2\rho_\star T}}{\lambda}.$$

Taking $\lambda = n^{2\theta}/2$ completes the proof. □

C Estimates based on concentration of measure on the high–dimensional orthogonal group

C.1 Generalities

We recall a few properties of Haar measure on the orthogonal group. We endow the orthogonal group $O(n)$ with the metric given by the Frobenius norm, so that $d(\boldsymbol{O}, \boldsymbol{U}) = \|\boldsymbol{O} - \boldsymbol{U}\|_F$. Say that a function $F : O(n) \to \mathbb{R}$ is Lipschitz with constant L if

$$|F(\boldsymbol{O}) - F(\boldsymbol{U})| \leq L\|\boldsymbol{O} - \boldsymbol{U}\|_F.$$

Recall that the orthogonal group can be partitioned as two disconnected copies of the special orthogonal group $\mathrm{SO}(n)$, which we endow with the same metric. These are given as the preimages of $\{\pm 1\}$ under the determinant map. Haar measure on the special orthogonal group enjoys a strong concentration of measure property.

Theorem C.1. *Suppose that $F : \mathrm{SO}(n) \to \mathbb{R}$ is Lipschitz with constant L. Then for all $t \geq 0$,*

$$\Pr[|F(\boldsymbol{U}) - \mathbb{E}[F(\boldsymbol{U})]| > t] \leq 2e^{-cnt^2/L^2},$$

where $c > 0$ is an absolute constant.

See [Vershynin, 2018, Theorem 5.2.7] or [Meckes, 2019, Theorem 5.17] for precise constants. We can derive concentration for even functions of the orthogonal group automatically:

Corollary C.1. *Suppose that $F : O(n) \to \mathbb{R}$ is Lipschitz with constant L and suppose that*

$$\mathbb{E}[F(\boldsymbol{U}) \det \boldsymbol{U}] = \mathbb{E}[F(\boldsymbol{U})] = 0.$$

Then for all $t \geq 0$,

$$\Pr[|F(\boldsymbol{U}) - \mathbb{E}[F(\boldsymbol{U})]| > t] \leq 2e^{-cnt^2/L^2},$$

where $c > 0$ is an absolute constant.

Proof. Under the assumption, the mean of F conditioning on either $\det \boldsymbol{U} = 1$ or $\det \boldsymbol{U} = -1$ is 0. Hence, by conditioning, we achieve the desired concentration around 0, which is the mean $\mathbb{E}[F(\boldsymbol{U})]$. □

As a useful illustration, an entry of $\boldsymbol{U}$, which is a Haar–distributed random matrix on $O(n)$ is concentrated.

Corollary C.2. *For any $n > 1$, there is an absolute constant $c > 0$ so that for all $t \geq 0$ and all i, j in $\{1, 2, ..., n\}$*

$$\Pr[|U_{ij}| > t] \leq 2e^{-cnt^2}.$$

The same statement holds for any generalized entry $\boldsymbol{x}^T \boldsymbol{U} \boldsymbol{y}$ for fixed unit vectors $\boldsymbol{x}, \boldsymbol{y}$, i.e.

$$\Pr[|\boldsymbol{x}^T \boldsymbol{U} \boldsymbol{y}| > t] \leq 2e^{-cnt^2}.$$

Proof. The entry map $\boldsymbol{U} \mapsto U_{ij}$ is 1–Lipschitz. Moreover, it has mean 0, restricted to either component, as

$$\mathbb{E}[U_{ij} \det(\boldsymbol{U})] = 0 = \mathbb{E}[U_{ij}].$$

Note that by distributional invariance, negating row i of $\boldsymbol{U}$ leaves the distribution of $\boldsymbol{U}$ invariant. Doing so shows the second equality. Negating any row except for i (which exists as $n > 1$) shows the first equality.

For the generalized entry, by the linearity of the expectation,

$$\mathbb{E}[\boldsymbol{x}^T \boldsymbol{U} \boldsymbol{y} \det(\boldsymbol{U})] = \sum_{i,j} x_i y_j \, \mathbb{E}[U_{ij} \det(\boldsymbol{U})] = 0 = \sum_{i,j} x_i y_j \, \mathbb{E}[U_{ij}] = \mathbb{E}[\boldsymbol{x}^T \boldsymbol{U} \boldsymbol{y}].$$

Using that $\boldsymbol{U} \mapsto \boldsymbol{x}^T \boldsymbol{U} \boldsymbol{y}$ is 1–Lipschitz, the proof follows. □

C.2 Applications to the Volterra equation errors

More to the point, we need concentration of random combinations of functions weighted by entries of $\boldsymbol{U}$.

Lemma C.1. *Let $T > 0$ be fixed, and suppose that $\{g_i\}$ are functions from $[0,T] \to \mathbb{R}$ which are bounded by $L > 0$ and have Lipschitz constant 1. Then, for any $\epsilon > 0$, and any fixed unit vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^n$, with overwhelming probability*

$$\max_{0\le t\le T}\left|\sum_{j=1}^{n} g_j(t)\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big)\right| \le n^{\epsilon-1/2}.$$

Proof. We first prove the claim for a fixed $t \in [0,T]$ and generalize the result for any $t \in [0,T]$ later using a mesh points argument. In proving this, we can take advantage of Corollary C.1. For $t \in [0,T]$, let $F_t : O(n) \to \mathbb{R}$ be

$$F_t(\boldsymbol{U}) \stackrel{\text{def}}{=} \sum_{j=1}^{n} g_j(t)\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big). \tag{62}$$

We can show that F_t is a Lipschitz function on $O(n)$. Indeed, for $\boldsymbol{U}, \boldsymbol{V} \in O(n)$,

$$\mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j] = \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{V})_j(\boldsymbol{b}^T\boldsymbol{V})_j] = \frac{1}{n}\langle \boldsymbol{a}, \boldsymbol{b}\rangle$$

and

$$\begin{aligned}
&|F_t(\boldsymbol{U}) - F_t(\boldsymbol{V})| \\
&\quad= \big|\sum_{j=1}^{n} g_j(t)\big((\boldsymbol{a}^T\boldsymbol{U})_j - (\boldsymbol{a}^T\boldsymbol{V})_j\big)(\boldsymbol{b}^T\boldsymbol{U})_j + \sum_{j=1}^{n} g_j(t)(\boldsymbol{a}^T\boldsymbol{V})_j\big((\boldsymbol{b}^T\boldsymbol{U})_j - (\boldsymbol{b}^T\boldsymbol{V})_j\big)\big| \\
&\quad\le \sqrt{\sum_{j=1}^{n}(\boldsymbol{a}^T(\boldsymbol{U}-\boldsymbol{V}))^2}\sqrt{\sum_{j=1}^{n} g_j^2(t)(\boldsymbol{b}^T\boldsymbol{U})_j^2} + \sqrt{\sum_{j=1}^{n}(\boldsymbol{b}^T(\boldsymbol{U}-\boldsymbol{V}))^2}\sqrt{\sum_{j=1}^{n} g_j^2(t)(\boldsymbol{a}^T\boldsymbol{V})_j^2} \\
&\quad\le L\|\boldsymbol{a}^T(\boldsymbol{U}-\boldsymbol{V})\|_2\|\boldsymbol{b}^T\boldsymbol{U}\|_2 + L\|\boldsymbol{b}^T(\boldsymbol{U}-\boldsymbol{V})\|_2\|\boldsymbol{a}^T\boldsymbol{V}\|_2 \\
&\quad\le 2L\|\boldsymbol{U}-\boldsymbol{V}\|_F.
\end{aligned}$$

Therefore, we conclude that F_t is a Lipschitz function of Lipschitz constant $2L$. For $j \in \{1, \cdots, n\}$, let

$$f_j \stackrel{\text{def}}{=} (\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j].$$

Then conditioning on $\det \boldsymbol{U} = 1$ and $\det \boldsymbol{U} = -1$, negating any column of $\boldsymbol{U}$ leaves the distribution of f_j invariant, which gives

$$\mathbb{E}[f_j(\boldsymbol{U})\det \boldsymbol{U}] = \mathbb{E}[f_j(\boldsymbol{U})] = 0,$$

and thus, using linearity,

$$\mathbb{E}[F_t(\boldsymbol{U})\det \boldsymbol{U}] = \mathbb{E}[F_t(\boldsymbol{U})] = 0.$$

Now Corollary C.1 gives, for $s \ge 0$,

$$\Pr[|F_t(\boldsymbol{U}) - \mathbb{E}[F_t(\boldsymbol{U})]| > s] \le 2e^{-cns^2},$$

where $c > 0$ is an absolute constant. Or, replacing $s = n^{\epsilon-1/2}$ gives the claim for a fixed time $t \in [0,T]$.

Now we generalize the result to any time in $[0,T]$. Assume that the claim is attained for mesh points on $[0,T]$ with arbitrarily small spacing, say λ. Then for any $t \in [0,T]$, there exists a mesh point $t_0 \in [0,T]$ such that $|t - t_0| \le \lambda$. The assumption that $\{g_j\}$ are Lipschitz functions with Lipschitz constant 1 implies that $|g_j(t) - g_j(t_0)| \le |t - t_0| \le \lambda$. Then we see

$$\begin{aligned}
&\big|\sum_{j=1}^{n}(g_j(t) - g_j(t_0))\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big)\big| \\
&\quad\le \sum_{j=1}^{n}|g_j(t) - g_j(t_0)|\big|\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big)\big| \le 2\lambda n.
\end{aligned}$$

Note that λ can be arbitrarily small. Thus we have

$$\begin{aligned}
&\left|\sum_{j=1}^{n} g_j(t)\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big)\right| \\
&\qquad\qquad \le \left|\sum_{j=1}^{n} g_j(t_0)\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big)\right| \\
&\qquad\qquad + \left|\sum_{j=1}^{n} (g_j(t) - g_j(t_0))\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big)\right| \\
&\qquad\qquad \le n^{\epsilon-1/2} + 2n\lambda < n^{\epsilon-1/2}
\end{aligned}$$

with overwhelming probability and with small enough ϵ in the last part. All in all, we have with overwhelming probability

$$\max_{0\le t\le T}\left|\sum_{j=1}^{n} g_j(t)\big((\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j - \mathbb{E}[(\boldsymbol{a}^T\boldsymbol{U})_j(\boldsymbol{b}^T\boldsymbol{U})_j]\big)\right| \le n^{\epsilon-1/2}.$$

□

As $\epsilon > 0$ is arbitrary, Lemma B.5 follows immediately.

Proof of Lemma B.5. We just need to apply Lemma C.1 with

$$g_j(t) = \sigma_j^2 e^{-2\gamma\sigma_j^2 t} \qquad \text{and} \qquad \boldsymbol{a} = \boldsymbol{b} = \mathbb{e}_i.$$

Note that we are conditioning on $\boldsymbol{\Sigma}$ so that the expectation in the statement of Lemma C.1 is only taken over $\boldsymbol{U}$. By the boundedness of σ_j, by dividing by a sufficiently large constant depending on $\boldsymbol{\Sigma}$, the Lemma applies. □

C.3 Control of the beta errors

Next, we prove Proposition B.5 provided that Proposition B.4 holds.

Proof of Proposition B.5. For $t \in [0, T]$, by equation (35), we have

$$(\mathcal{B}_{s,j})^2 = (-\gamma\sigma_j^2\nu_{t,j}^{\vartheta} + \gamma\sigma_j(\boldsymbol{U}^T\boldsymbol{\eta})_j)^2 \le 2(\gamma^2\sigma_j^4\beta + \gamma^2\sigma_j^2)n^{2\epsilon-1},$$

with overwhelming probability. Here Proposition B.4 was used to bound $\nu_{t,j}^{\vartheta}$, and Assumption 1.1 and Corollary C.2 imply $|(\boldsymbol{U}^T\boldsymbol{\eta})_j| \le n^{\epsilon-1/2}$ w.o.p. by observing

$$\big((\boldsymbol{U}^T\boldsymbol{\eta})_j : 1 \le j \le n\big) \overset{\text{law}}{=} \|\boldsymbol{\eta}\|\big(U_{1,j} : 1 \le j \le n\big).$$

We recall the definition of ϑ from Lemma B.4 as

$$\vartheta = \inf\big\{t \ge 0 \,:\, \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta}\| > n^{\theta}\big\},$$

where $\theta < \epsilon/2$. By applying Corollary C.2 and the definition of ϑ, we get

$$\sum_{i=1}^{n}\big(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\big)^2\big(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t^{\vartheta} - \boldsymbol{\eta})\big)^2 \le n^{\epsilon-1}\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t^{\vartheta} - \boldsymbol{\eta}\|^2 \le n^{2\epsilon-1},$$

with overwhelming probability. Therefore,

$$
\begin{aligned}
&|\varepsilon^{(n)}_{\mathrm{beta}}(t)| \\
&= \frac{1}{2}\Big|\sum_{j=1}^{d}\sigma_j^2\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\left(\frac{\beta-1}{n-1}\big(\mathcal{B}_{s,j}\big)^2-\frac{\beta-1}{n-1}\sum_{i=1}^{n}\big(\mathbb{e}_j^T\boldsymbol{U}^T\mathbb{e}_i\big)^2\big(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_s-\boldsymbol{\eta})\big)^2\right)\mathrm{d}s\Big| \\
&\le \frac{1}{2}\sum_{j=1}^{d}\sigma_j^2\left(\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\,\mathrm{d}s\right)\left(\frac{\beta}{n}\cdot 2\big(\gamma^2\sigma_j^4+\gamma^2\sigma_j^2\big)n^{2\epsilon-1}+\frac{\beta}{n}\cdot n^{2\epsilon-1}\right) \\
&= \frac{1}{2}\sum_{j=1}^{d}\frac{1}{2\gamma}\cdot\frac{\beta}{n}\cdot n^{2\epsilon-1}(2\gamma^2\sigma_j^4+2\gamma^2\sigma_j^2+1) \\
&\le n^{\epsilon-1/2},
\end{aligned}
$$

with small enough ϵ in the last line, given our assumption on $\beta\le n^{1/5-\epsilon}$. □

C.4 Control of the eta errors

We now give the proof of Proposition B.7 provided that Proposition B.6 holds.

Proof of Proposition B.7. Note that the proof of Proposition B.6 is based on conditioning on $\boldsymbol{\Sigma}$. Therefore, Proposition B.6 by substituting $c_j=\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\gamma\sigma_j^2\,\mathrm{d}s$ and $c_j=1$, respectively, implies with overwhelming probability

$$
\left|\sum_{j=1}^{d}\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\gamma\sigma_j^3\nu^{\vartheta}_{s,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j-\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{d}\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\gamma\sigma_j^2(1-e^{-(s\wedge\vartheta)\gamma\sigma_j^2})\,\mathrm{d}s\right|\le\|\boldsymbol{\eta}\|^2\sqrt{\beta}n^{\epsilon-1/2}
$$

and

$$
\left|\sum_{j=1}^{n\wedge d}\sigma_j\nu^{\vartheta}_{t,j}(\boldsymbol{U}^T\boldsymbol{\eta})_j-\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{n\wedge d}(1-e^{-(t\wedge\vartheta)\gamma\sigma_j^2})\right|\le\|\boldsymbol{\eta}\|^2\sqrt{\beta}n^{\epsilon-1/2}.
$$

Therefore, it suffices to show

$$
\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{d}\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\gamma\sigma_j^2(1-e^{-s\gamma\sigma_j^2})\,\mathrm{d}s+\frac{1}{2}\|\boldsymbol{\eta}\|^2-\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{n\wedge d}(1-e^{-t\gamma\sigma_j^2})-\widetilde{R}\cdot\frac{rh_0(t)+(1-r)}{2}\tag{63}
$$

converges to 0 in probability as $n\to\infty$.

Observe

$$
\begin{aligned}
&\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{d}\int_0^t e^{-2(t-s)\gamma\sigma_j^2}\gamma\sigma_j^2(1-e^{-s\gamma\sigma_j^2})\,\mathrm{d}s \\
&\qquad=\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{d}\gamma\sigma_j^2e^{-2t\gamma\sigma_j^2}\int_0^t e^{2s\gamma\sigma_j^2}(1-e^{-s\gamma\sigma_j^2})\,\mathrm{d}s \\
&\qquad=\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{d}e^{-2t\gamma\sigma_j^2}\left[\frac{1}{2}(e^{2t\gamma\sigma_j^2}-1)-(e^{t\gamma\sigma_j^2}-1)\right] \\
&\qquad=\frac{\|\boldsymbol{\eta}\|^2}{n}\left[\frac{d}{2}+\frac{1}{2}\sum_{j=1}^{d}e^{-2\gamma\sigma_j^2t}-\sum_{j=1}^{d}e^{-t\gamma\sigma_j^2}\right].
\end{aligned}
$$

Note that $\frac{1}{2}\|\boldsymbol{\eta}\|^2\xrightarrow[n\to\infty]{\Pr}\frac{\widetilde{R}}{2}$, $\frac{d\|\boldsymbol{\eta}\|^2}{2n}\xrightarrow[n\to\infty]{\Pr}\frac{r\widetilde{R}}{2}$ by Assumptions 1.1 and $\frac{\|\boldsymbol{\eta}\|^2}{2n}\sum_{j=1}^{d}e^{-2\gamma\sigma_j^2t}\xrightarrow[n\to\infty]{\Pr}\frac{\widetilde{R}r}{2}h_0(t)$ by Assumptions 1.2. Moreover, $\sum_{j=1}^{n\wedge d}(1-e^{-t\gamma\sigma_j^2})=\sum_{j=1}^{d}(1-e^{-t\gamma\sigma_j^2})$ always holds because $\sigma_j=0$ for $j>n\wedge d$ and this cancels out $\frac{\|\boldsymbol{\eta}\|^2}{n}\sum_{j=1}^{d}e^{-t\gamma\sigma_j^2}$ in (63).

□

D Estimates based on martingale concentration

D.1 General techniques

We recall that the martingales $M_{t,j}$ and $\widetilde{M}_{t,j}$ are defined in (34). Both of these martingales need to be controlled, but only after summing them in a specific way. First, we do not need these martingales directly, but only certain integrals against these martingales. These are defined for all $t \geq 0$ and all $j \in \{1, 2, \ldots, d\}$,

$$\begin{aligned} \widetilde{X}_{t,j} &= \int_0^t e^{s\gamma\sigma_j^2} \, \mathrm{d}\widetilde{M}_{s,j} \\ X_{t,j} &= \int_0^t e^{2s\gamma\sigma_j^2} \, \mathrm{d}M_{s,j}, \end{aligned} \tag{64}$$

which are again càdlàg, finite variation martingales. We will need to show concentration for sums of these martingales such as $\sum_{j=1}^d c_j \widetilde{X}_{t,j}$ and $\sum_{j=1}^d c_j X_{t,j}$ for bounded coefficients $\{c_j\}$.

We formulate some general concentration lemmas for càdlàg, finite variation martingales Y_t with jumps given exactly by $\{\tau_k : k \geq 0\}$. For such a process, the jumps entirely determine its fluctuations. We will define for any càdlàg process Y,

$$\Delta Y_t \stackrel{\text{def}}{=} Y_t - Y_{t-},$$

which is 0 for all t except $\{\tau_k : k \geq 0\}$. For concreteness and for reference, we record that the jumps of $\widetilde{X}$ and X are given by

$$\begin{aligned} \Delta\widetilde{X}_{\tau_k,j} &= e^{\gamma\sigma_j^2\tau_k}\big(\mathfrak{e}_j^T\gamma\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_{k-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-} - \boldsymbol{\eta})\big), \\ \Delta X_{\tau_k,j} &= e^{2\gamma\sigma_j^2\tau_k}\big(\nu_{\tau_k,j} + \nu_{\tau_k-,j}\big)\big(\mathfrak{e}_j^T\gamma\boldsymbol{\Sigma}^T\boldsymbol{U}^T\boldsymbol{P}_{k-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-} - \boldsymbol{\eta})\big), \end{aligned} \tag{65}$$

To control the fluctuations of these martingales, we need to control their quadratic variations. The *quadratic variation* $[Y_t]$ is the sum of squares of all jumps of the process, and hence

$$[Y_t] \stackrel{\text{def}}{=} \sum_{k=1}^{N_t} (\Delta Y_{\tau_k})^2.$$

Likewise the *predictable quadratic variation* $\langle Y_t \rangle$ is

$$\langle Y_t \rangle \stackrel{\text{def}}{=} \sum_{k=1}^{N_t} \mathbb{E}\big[(\Delta Y_{\tau_k})^2 \, | \mathcal{F}_{\tau_k-}\big].$$

Moreover, for some of the martingales we consider here, it is possible to find good events on which the quadratic variation or the predictable quadratic variations are in control. Then it is a relatively standard fact that the fluctuations of these processes are in control:

Lemma D.1. *Suppose that $(Y_t : t \geq 0)$ is a càdlàg finite variation martingale. Suppose there is an event $\mathcal{G}$ which is measurable with respect to $\mathcal{F}_0$ that holds with overwhelming probability, and so that for some $T > 0$*

$$(i) \quad [Y_T]\mathbb{1}_{\mathcal{G}} \leq \tfrac{\beta}{nT} N_T; \qquad \textit{or} \qquad (ii) \quad \langle Y_T \rangle \mathbb{1}_{\mathcal{G}} \leq \tfrac{\beta}{nT} N_T \quad \textit{and} \quad \max_{0 \leq t \leq T} |Y_t - Y_{t-}| \mathbb{1}_{\mathcal{G}} \leq 1.$$

Then for any $\epsilon > 0$ with overwhelming probability

$$\sup_{0 \leq t \leq T} |Y_t| \leq n^{\epsilon}.$$

Proof. We begin with the proof of *(i)*. Using the Burkholder–Davis–Gundy inequalities (see [Protter, 2005, Theorem IV.49]), for any $p > 1$ there is a constant C_p so that

$$\mathbb{E}\bigg(\sup_{0 \leq t \leq T} |Y_t| \mathbb{1}_{\mathcal{G}}\bigg)^p \leq C_p \, \mathbb{E}\big[[Y_T]^p \mathbb{1}_{\mathcal{G}}\big] \leq C_p \big(\tfrac{\beta}{nT}\big)^p \, \mathbb{E}[N_T^p].$$

There is an absolute constant $C > 0$ so that

$$\mathbb{E}(N_T^p) \leq Cp!(\mathbb{E}\, N_T)^p,$$

and so we conclude that

$$\mathbb{E}\left(\sup_{0\leq t\leq T} |Y_t|\mathbb{1}_{\mathcal{G}}\right)^p \leq CC_p.$$

Using Markov's inequality, we conclude that

$$\Pr(\{\sup_{0\leq t\leq T} |Y_t| \geq n^\epsilon\}) \leq \Pr(\mathcal{G}^c) + CC_p n^{-\epsilon p}.$$

Hence letting p tend slowly to infinity with n, this concludes the proof of *(i)*.

We turn to the proof of *(ii)*. We need a tail bound for martingales (see [Shorack and Wellner, 1986, Appendix B.6 Inequality 1]), which states that

$$\Pr(\{\sup_{0\leq t\leq T} |Y_t| > s\} \cap \{\langle Y_T\rangle \leq r\} \cap \mathcal{G}) \leq 2\exp\left(-\frac{s^2}{\frac{2s}{3} + 2r}\right).$$

Taking $s = r = n^\epsilon$, this vanishes faster than any power of n. The probability that $N_T > n^\epsilon(\mathbb{E}[N_T])$ additionally decays faster than any power of n, so that we conclude that on $\mathcal{G}$, $\sup_{0\leq t\leq T} |Y_t| \leq n^\epsilon$ with overwhelming probability. □

We will need an extension of this standard type of concentration, which allows for exceptional jumps. Suppose we can decompose the jumps $\{\tau_k\}$ of $(Y_t : t \geq 0)$ into two types $\{\tau_{k,1}, \tau_{k,2}\}$. In our application, we shall pick the jumps of the second type to be those for which a fixed coordinate $1 \in B_k$ and the first type to be all that remains. Thus by properties of the Poisson process, the two processes $\{\tau_{k,1}, \tau_{k,2}\}$ are independent Poisson processes.

Lemma D.2. *Suppose that $(Y_t : t \geq 0)$ is a càdlàg finite variation martingale with jumps given by $\{\tau_k\}$. Suppose these jumps are divided into two groups $\{\tau_{k,1}, \tau_{k,2}\}$ by a rule depending only on $(k, \boldsymbol{P}_k)$. Let $N_{t,1}$ and $N_{t,2}$ be the counting functions of the number of jumps from either type. Suppose that the jumps of Y_t of type 1 (the typical ones) satisfy*

$$\mathbb{E}[\left(\Delta Y_{\tau_{k,1}}\right)^2 \mid \mathcal{F}_{\tau_{k,1}-}] \leq \frac{\beta}{n} \quad \textit{and} \quad |\Delta Y_{\tau_{k,1}}| \leq 1.$$

For the jumps of the second type, suppose that for some $T > 0$ there is a constant $C > 1$ so that $\mathbb{E}[N_{T,2}] \leq C$ and a constant $\delta \in (0, 1)$ so that

$$(i). \quad |\Delta Y_{\tau_{k,2}}| \leq \delta |Y_{\tau_{k,2}-}| + 1 \qquad \textit{or} \qquad (ii). \quad |Y_{\tau_{k,2}}| \leq \delta |Y_{\tau_{k,2}-}| + 1.$$

Then for any $\epsilon > 0$ with overwhelming probability

$$\sup_{0\leq t\leq T} |Y_t| \leq n^\epsilon.$$

Proof. Let $n_{t,1}$ and $n_{t,2}$ be the Lévy measures for the jumps of Y of types 1 and 2; i.e. the measures so that for any bounded continuous function f and $\ell \in \{1, 2\}$,

$$\sum_{k=1}^{N_{t,\ell}} f\left(\Delta Y_{\tau_{k,\ell}}\right) - \int f(x) n_{t,\ell}(\mathrm{d}x)$$

is a martingale. We decompose the martingale $(Y_t : t \geq 0)$ into pieces. Define

$$Y_{t,\ell} = \sum_{k=1}^{N_{t,\ell}} \Delta Y_{\tau_{k,\ell}} - \int x n_{t,\ell}(\mathrm{d}x).$$

Then $Y_t = Y_{t,1} + Y_{t,2}$ for all $t \geq 0$.

We use two different versions of the exponential martingale. The first, which we believe originates with Yor [1976] (c.f. [Lépingle, 1978, Lemme 2]) is

$$\widehat{Z}_{t,1} \stackrel{\text{def}}{=} \exp\Big(\lambda Y_{t,1} - \int \big(e^{\lambda x} - 1 - \lambda x\big) n_{t,1}(\mathrm{d}x)\Big),$$

which is a martingale. The second is the Doléans exponential, which is the more commonly cited ([Protter, 2005, II. Theorem 37], Yor [1976]), and which shows

$$\widehat{Z}_{t,2} \stackrel{\text{def}}{=} \exp\big(\lambda Y_{t,2}\big) \prod_{k=1}^{N_{t,2}} f\big(\lambda(\Delta Y_{\tau_{k,2}})\big) \quad \text{where} \quad f(x) = (1+x)e^{-x}.$$

As both processes are finite variation and have no common jumps, their product remains a martingale. Thus

$$\widehat{Z}_t \stackrel{\text{def}}{=} \exp\Big(\lambda Y_{t,1} + \lambda Y_{t,2} - \int \big(e^{\lambda x} - 1 - \lambda x\big) n_{t,1}(\mathrm{d}x)\Big) \prod_{k=1}^{N_{t,2}} f\big(\lambda \Delta Y_{\tau_{k,2}}\big)$$

is a martingale. Note that the two martingales combine to form Y_t. By the assumption, the jumps of $Y_{t,1}$ are less than or equal to 1. Hence the measure $n_{t,1}(\mathrm{d}x)$ is supported on $[-1, 1]$. For $|u| \le 1$,

$$e^u - 1 - u \le \tfrac{u^2}{e-2}.$$

So we define a supermartingale $(Z_t : t \ge 0)$ for any $\lambda \le 1$ by

$$\widehat{Z}_t \ge Z_t \stackrel{\text{def}}{=} \exp\Big(\lambda Y_t - \int \frac{\lambda^2 x^2}{e-2} n_{t,1}(\mathrm{d}x)\Big) \prod_{k=1}^{N_{t,2}} f\big(\lambda \Delta Y_{\tau_{k,2}}\big).$$

The integral $\int x^2 n_{t,1}(\mathrm{d}x)$ is the predictable quadratic variation

$$\langle Y_{t,1} \rangle = \sum_{k=1}^{N_{t,1}} \mathbb{E}\big[(\Delta Y_{\tau_{k,1}})^2 \mid \mathcal{F}_{\tau_{k,1}-}\big] \le \tfrac{\beta}{n} N_{t,1}.$$

Now we fix a parameter $r > \frac{1}{1-\delta}$ and let

$$\vartheta = \inf\{t \ge 0 \ : \ |Y_t| \ge r\}.$$

By optional stopping for any bounded stopping time $\rho \ge 0$,

$$\mathbb{E}\big[Z_{\vartheta\wedge\rho} \mathbb{1}_{\vartheta \le \rho}\big] \le \mathbb{E}[Z_{\vartheta\wedge\rho}] \le \mathbb{E}[Z_0] = 1. \tag{66}$$

So, for $\lambda \in (0, 1)$,

$$\begin{aligned} Z_{\vartheta\wedge\rho} \mathbb{1}_{\vartheta\le\rho} \ge & \exp\big(\lambda r - \tfrac{\lambda^2}{e-2}\tfrac{\beta}{n} N_{\rho,1}\big) \prod_{k=1}^{N_{\rho,2}} f\big(\lambda \Delta Y_{\tau_{k,2}}\big)\big) \mathbb{1}_{\{Y_{\vartheta\wedge\rho} \ge r\}} \\ & - \exp\big(-\lambda r - \tfrac{\lambda^2}{e-2}\tfrac{\beta}{n} N_{\rho,1}\big) \prod_{k=1}^{N_{\rho,2}} |f\big(\lambda \Delta Y_{\tau_{k,2}}\big)| \mathbb{1}_{\{Y_{\vartheta\wedge\rho} \le -r\}}. \end{aligned}$$

We produce a similar bound on taking $-\lambda \in (0, 1)$ although with the roles reversed.

The product may in principle be negative or 0. So we consider taking $\rho = T \wedge \tau_{1,2}$, for some fixed $T > 0$. Then if $\vartheta < \tau_{1,2}$, we have an empty product. Otherwise, we have $\vartheta = \tau_{1,2}$, in which case the product contains a single term.

If assumption (ii) is in force, then either the jump decreases the absolute value of $Y_{\tau_{1,2}}$ as it is opposite sign from $Y_{\tau_{1,2}-}$ and does not cross 0 or the second condition is in force. In that case, since $|Y_{\tau_{1,2}-}| \le r$, and since

$$r \le |Y_{\tau_{1,2}}| \le \delta r + 1,$$

we would have $r \le \frac{1}{1-\delta}$. However, we have chosen r large enough that this is not the case. So, we conclude that when assumption (ii) is in force, we could not have had $\vartheta = \tau_{1,2}$. We conclude in the case of assumption (ii) that

$$Z_{\vartheta\wedge\rho}\mathbb{1}_{\{\vartheta\le\rho\}} \ge \exp\big(\lambda r - \tfrac{\lambda^2}{e-2}\tfrac{\beta}{n}N_{\rho,1}\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\ge r\}} - \exp\big(-\lambda r - \tfrac{\lambda^2}{e-2}\tfrac{\beta}{n}N_{\rho,1}\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\le -r\}}. \tag{67}$$

If assumption (i) is in force, then if $Y_\vartheta \ge r$, the jump of Y at $\tau_{1,2}$ is necessarily positive, as this is the first time the martingale jumped above some level. As assumption (i) is in force, then $Y_{\tau_{1,2}-} > 0$ as well, and so the jump of type 2 must satisfy

$$\Delta Y_{\tau_{1,2}} \le \delta Y_{\tau_{1,2}-} + 1 \le \delta r + 1.$$

We conclude that when $Y_\vartheta \ge r$ and assumption (i) holds,

$$f\big(\lambda\Delta Y_{\tau_{1,2}}\big) \ge e^{-\lambda\Delta Y_{\tau_{1,2}}} \ge e^{-\lambda(\delta r+1)}.$$

If on the other hand $Y_\vartheta \le -r$, then the jump must have been negative, and we conclude similarly that

$$|f\big(\lambda\Delta Y_{\tau_{1,2}}\big)| \le (1-\lambda\Delta Y_{\tau_{1,2}})e^{-\lambda\Delta Y_{\tau_{1,2}}} \le (1+\lambda(\delta r+1))e^{\lambda(\delta r+1)}.$$

Hence

$$\begin{aligned} Z_{\vartheta\wedge\rho}\mathbb{1}_{\{\vartheta\le\rho\}} \ge{}& \exp\big(-1+\lambda(1-\delta)r - \tfrac{\lambda^2}{e-2}\tfrac{\beta}{n}N_{\rho,1}\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\ge r\}} \\ &- (1+\lambda(\delta r+1))\exp\big(1-\lambda(1-\delta)r - \tfrac{\lambda^2}{e-2}\tfrac{\beta}{n}N_{\rho,1}\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\le -r\}}. \end{aligned} \tag{68}$$

In either case of (67) or (68), using (66) and the boundedness of

$$r \mapsto \lambda(\delta r+1)e^{-\lambda(1-\delta)r},$$

there is a constant $C_\delta > 0$ so that

$$\mathbb{E}\bigg(\exp\big(\lambda(1-\delta)r - \tfrac{\lambda^2}{e-2}\tfrac{\beta}{n}N_{\rho,1}\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\ge r\}}\bigg) \le C_\delta.$$

With overwhelming probability $\frac{\beta}{n}N_{\rho,1} \le \frac{\beta}{n}N_T \le 2T$, and hence on the event $\mathcal{E}$ that $\frac{\beta}{n}N_{\rho,1} \le 2T$,

$$\mathbb{E}\bigg(\exp\big(\lambda(1-\delta)r - \tfrac{\lambda^2}{e-2}2T\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\mathbb{1}_{\mathcal{E}}\ge r\}}\bigg) \le C_\delta. \tag{69}$$

Thus taking $\lambda = 1$ and $r = (1-\delta)^{-1}(\log n)^2$, we conclude that

$$e^{(\log n)^2}\Pr(Y_{\vartheta\wedge\rho} \ge n^\epsilon \cap \mathcal{E}) = O(1),$$

and hence $Y_{\vartheta\wedge\rho} \le n^\epsilon$ with overwhelming probability.

By applying the same argument to $-Y_t$ which is again a martingale satisfying the same assumptions, we can conclude with overwhelming probability that

$$\sup\big\{|Y_t| : 0 \le t \le (T\wedge\tau_{1,2})\big\} \le 2(1-\delta)^{-1}(\log n)^2.$$

Now we suppose that with overwhelming probability, we have shown for some $\ell \in \mathbb{N}$

$$\sup\big\{|Y_t| : 0 \le t \le (T\wedge\tau_{\ell,2})\big\} \le 2^\ell(1-\delta)^{-\ell}(\log n)^2.$$

We now apply the same bounds to

$$Z_t/Z_{t\wedge\tau_{\ell,2}} \stackrel{\text{def}}{=} \exp\bigg(\lambda(Y_t - Y_{t\wedge\tau_{\ell,2}}) - \int\frac{\lambda^2x^2}{e-2}n_{t,1}(\mathrm{d}x)\bigg)\prod_{k=\ell+1}^{N_{t,2}} f\big(\lambda\Delta Y_{\tau_{k,2}}\big).$$

In particular taking the conditional expectation, with the same ϑ and with $\rho = T\wedge\tau_{\ell+1,2}$

$$\mathbb{E}[Z_{\vartheta\wedge\rho}/Z_{\vartheta\wedge\rho\wedge\tau_{\ell,2}}\mathbb{1}_{\{\vartheta\le\rho\}} \mid \mathcal{F}_{\tau_{k,2}}] \le 1.$$

Rearranging, we conclude

$$\mathbb{E}\bigg(\exp\bigg(\lambda Y_{\vartheta\wedge\rho} - \int \frac{\lambda^2 x^2}{e-2} n_{t,1}(\mathrm{d}x)\bigg) \prod_{k=\ell+1}^{N_{\vartheta\wedge\rho,2}} f\big(\lambda \Delta Y_{\tau_{k,2}}\big) \mid \mathcal{F}_{\tau_{\ell,2}}\bigg) \le \exp\bigg(\lambda Y_{\vartheta\wedge\rho\wedge\tau_{\ell,2}}\bigg).$$

Hence following the same line of argument that leads to (69),

$$\mathbb{E}\bigg(\exp\big(\lambda(1-\delta)r - \tfrac{\lambda^2}{e-2}2T\big)\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\ge r\}}\mathbb{1}_{\mathcal{E}} \mid \mathcal{F}_{\tau_{\ell,2}}\bigg) \le C_\delta \exp\bigg(\lambda Y_{\vartheta\wedge\rho\wedge\tau_{\ell,2}}\bigg).$$

Taking $\lambda = 1$ and $r = 2^{\ell+1}(1-\delta)^{-\ell-1}$ and restricting to the event in the inductive hypothesis,

$$\mathbb{E}\bigg(\exp\big(2^{\ell+1}(1-\delta)^{-\ell}(\log n)^2\big)\mathbb{1}_{\{Y_{\vartheta\wedge\rho}\ge r\}}\mathbb{1}_{\mathcal{E}} \mid \mathcal{F}_{\tau_{\ell,2}}\bigg) \le C_\delta \exp\bigg(2^{\ell}(1-\delta)^{-\ell}(\log n)^2\bigg).$$

In particular, with overwhelming probability,

$$\sup\big\{|Y_t| : 0 \le t \le (T \wedge \tau_{\ell+1,2})\big\} \le 2^{\ell}(1-\delta)^{-\ell}(\log n)^2.$$

The number of type-2 jumps before T is $N_{T,2}$, which is Poisson with mean C. Hence with overwhelming probability, for any $\epsilon > 0$, $N_{T,2} \le \big(\log \frac{2}{1-\delta}\big)^{-1} \frac{\epsilon}{2} \log n$. Hence, we conclude that with overwhelming probability,

$$\sup\big\{|Y_t| : 0 \le t \le T\big\} \le \big(\tfrac{2}{1-\delta}\big)^{N_{T,2}} (\log n)^2 \le n^{\epsilon}(\log n)^2.$$

As $\epsilon > 0$ may be picked as small as desired, the proof follows. □

D.2 Applications to the control of errors in the Volterra equation

D.2.1 Delocalization of the function values: the proof of Proposition B.2

Proof of Proposition B.2. It suffices to prove that for a fixed i and for any $T > 0$ and any $\epsilon > 0$,

$$\sup_{0\le t\le T} \big(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t^{\vartheta} - \boldsymbol{\eta})\big)^2 \le \beta n^{\epsilon-1}$$

with overwhelming probability.

Using Lemma B.2, we have the representation

$$\nu_{t,j} = e^{-\gamma\sigma_j^2 t}\nu_{0,j} + \int_0^t e^{-\gamma\sigma_j^2(t-s)}\gamma\sigma_j(\boldsymbol{U}^T\boldsymbol{\eta})_j \,\mathrm{d}s + e^{-\gamma\sigma_j^2 t}\widetilde{X}_{t,j}.$$

Note that $\nu_{t,j}^{\vartheta}$ has the same representation by replacing $t \to t\wedge\vartheta$. Observe that each of the first two terms has the contribution of $\mathcal{O}(n^{\epsilon-1/2})$ to $|\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t^{\vartheta} - \boldsymbol{\eta})|$ with overwhelming probability. Indeed, we have

$$\begin{aligned}\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t^{\vartheta} - \boldsymbol{\eta}) - \sum_{j=1}^d U_{ij}\sigma_j e^{-\gamma t\sigma_j^2}\widetilde{X}_{t,j} &= -\eta_i + \sum_{j=1}^d U_{ij}\sigma_j e^{-\gamma\sigma_j^2 t}\nu_{0,j} \\ &\quad + \sum_{j=1}^d U_{ij}\sigma_j \int_0^t e^{-\gamma\sigma_j^2(t-s)}\gamma\sigma_j(\boldsymbol{U}^T\boldsymbol{\eta})_j \,\mathrm{d}s \\ &= \mathcal{O}(n^{\epsilon-1/2}).\end{aligned} \tag{70}$$

Here the order of first term comes from Assumption 1.1. Corollary C.1 gives the order of the second term, by defining $F(U) := \sum_{j=1}^d U_{ij}\sigma_j e^{-\gamma\sigma_j^2 t}\nu_{0,j}$ with conditioning on $\boldsymbol{\nu}_0$. The order of the last term is obtained from Lemma C.1 with setting $\boldsymbol{a} = \mathbb{e}_i, \boldsymbol{b} = \boldsymbol{\eta}$ and conditioning on $\boldsymbol{\Sigma}$. Indeed, note that $\mathbb{E}[U_{ij}(\boldsymbol{U}^T\boldsymbol{\eta})_j] = \frac{1}{n}\eta_i$ so that w.o.p.,

$$\mathbb{E}\bigg[\sum_{j=1}^d \gamma U_{ij}\sigma_j^2(\boldsymbol{U}^T\boldsymbol{\eta})_j \int_0^t e^{-(t-s)\gamma\sigma_j^2}\,\mathrm{d}s\bigg] = \bigg(\frac{1}{n}\sum_{j=1}^d \gamma\sigma_j^2 \int_0^t e^{-(t-s)\gamma\sigma_j^2}\bigg)\eta_i = \mathcal{O}(n^{\epsilon-1/2}).$$

Therefore, it would suffice to bound

$$Y_t \stackrel{\text{def}}{=} \sum_{j=1}^{d} U_{ij}\sigma_j e^{-\gamma q \sigma_j^2} \widetilde{X}_{t,j} \qquad 0 \le t \le q, \tag{71}$$

for some fixed $q \ge 0$. Note that as we did in Lemma B.3, showing the bound for a fixed $q \in [0,T]$ should be sufficient, considering mesh points on $[0,T]$ with spacing, let us say, $\lambda = \lambda(n) > 0$, which depends on n. Since the process $\boldsymbol{\nu}_t$ is constant between jumps, the only cases which cannot be covered by mesh points are having multiple jumps between two adjacent mesh points. However, as the possibility of such events is given by $\mathcal{O}(\beta^{-2}n^2\lambda)$, which can be smaller than any power of polynomial of n, we conclude that every jump can be covered by mesh points with overwhelming probability. Each jump for the process $Y_{(\cdot)}$ is given by

$$\Delta Y_{\tau_{k+1}} \stackrel{\text{def}}{=} Y_{\tau_{k+1}} - Y_{\tau_{(k+1)-}} = -\sum_{j=1}^{d} U_{ij}\sigma_j e^{-(q-\tau_{k+1})\gamma\sigma_j^2} \mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_k (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\vartheta}_{\tau_{(k+1)-}} - \boldsymbol{\eta}). \tag{72}$$

Note that there are two different types of jumps, i.e.

1. B_k does not include the index i.
2. B_k includes the index i.

Replacing $\boldsymbol{P}_k$ by $\sum_{l\in B_k} \mathbb{e}_l \mathbb{e}_l^T$, the jump $\Delta Y_{\tau_{k+1}}$ can be translated as

$$\Delta Y_{\tau_{k+1}} = -\sum_{l\in B_k} \Big[\sum_{j=1}^{d} \gamma U_{ij}U_{lj}\sigma_j^2 e^{-(q-\tau_{k+1})\gamma\sigma_j^2}\Big] \mathbb{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\vartheta}_{\tau_{(k+1)-}} - \boldsymbol{\eta}). \tag{73}$$

Let

$$\Phi_{i,l}(t) \stackrel{\text{def}}{=} \sum_{j=1}^{d} \gamma U_{ij}U_{lj}\sigma_j^2 e^{-(q-t)\gamma\sigma_j^2}, \tag{74}$$

and let $\mathcal{G} = \mathcal{G}(\theta)$ for $\theta > 0$ be the event defined as

$$\mathcal{G} \stackrel{\text{def}}{=} \left\{ \sup_{1\le l\le n, l\neq i} \max_{0\le t\le T} |\Phi_{i,l}(t)| \le n^{\theta-1/2}, \ \max_{0\le t\le T} |\Phi_{i,i}(t)| < 2 \right\}. \tag{75}$$

Note that this holds with overwhelming probability, by Lemma C.1 and condition on the stepsize γ, see Theorem 1.2. Furthermore, in order to apply the *bootstrap* argument, let us define, for $\aleph \in [-\epsilon/2, 1/2)$,

$$\hbar \stackrel{\text{def}}{=} \inf\{t \le \vartheta : \max_{1\le l\le n} |\mathbb{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})| > n^{-\aleph}\}. \tag{76}$$

Now we are ready to apply Lemma D.2 to prove the claim.

Case 1. When B_k does not include the index i: we need to control $\mathbb{E}[(\Delta Y^{\hbar}_{\tau_{k+1,1}})^2 | \mathcal{F}_{\tau_{(k+1)-}}]$ and $|\Delta Y^{\hbar}_{\tau_{k+1,1}}|$. Observe,

$$|\Delta Y^{\hbar}_{\tau_{k+1,1}}| = |\sum_{l\in B_k} \Phi_{i,l} \mathbb{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,1}-} - \boldsymbol{\eta})| \le \beta n^{\theta-1/2-\aleph}. \tag{77}$$

On the other hand,

$$\begin{aligned}
\mathbb{E}[(\Delta Y^{\hbar}_{\tau_{k+1,1}})^2 | \mathcal{F}_{\tau_{(k+1)-}}] &= \frac{\beta(\beta-1)}{(n-1)(n-2)} \Big[\sum_{l=1}^{n} \Phi_{i,l} \mathbb{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,1}-} - \boldsymbol{\eta})\Big]^2 \\
&\quad + \left(\frac{\beta}{n-1} - \frac{\beta(\beta-1)}{(n-1)(n-2)}\right) \sum_{l=1}^{n} \Phi_{i,l}^2 (\mathbb{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,1}-} - \boldsymbol{\eta}))^2 \\
&\le \frac{\beta^2}{n^2} n^{4\theta} + \frac{\beta}{n} n^{4\theta-1} \\
&\le \frac{\beta}{n} (\beta n^{4\theta-1} + n^{4\theta-1}).
\end{aligned} \tag{78}$$

Here Cauchy-Schwarz inequality as well as the definition of ϑ were used for the inequality.

Case 2. When B_k includes the index i: In this case, we want to have the following: for some $T > 0$, there is a constant $C > 1$ so that $\mathbb{E}N_{T,2} \le C$ and a constant $\delta \in (0,1)$ so that

$$(i). \quad |\Delta Y^{\hbar}_{\tau_{k+1,2}}| \le \delta |Y^{\hbar}_{\tau_{k+1,2}-}| + 1 \qquad \text{or} \qquad (ii). \quad |Y^{\hbar}_{\tau_{k+1,2}}| \le \delta |Y^{\hbar}_{\tau_{k+1,2}-}| + 1.$$

First recall that N_t has the distribution of Poisson$(\frac{n}{\beta}t)$. Since B_k contains the index i with probability $\binom{n-1}{\beta-1}/\binom{n}{\beta} = \frac{\beta}{n}$, we have

$$\mathbb{E}[N_{T,2}] = \frac{\beta}{n}\frac{nT}{\beta} = T < \infty.$$

Now observe, with $\mathfrak{t} \stackrel{\text{def}}{=} \tau_{k+1,2} \wedge \hbar$,

$$\begin{aligned}
\Delta Y^{\hbar}_{\tau_{k+1,2}} &= -\sum_{l \in B_k} \big[\sum_{j=1}^{d} \gamma U_{ij} U_{lj} \sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2}\big] \mathfrak{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,2}-} - \boldsymbol{\eta}) \\
&= -\sum_{j=1}^{d} \gamma U_{ij}^2 \sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2} \mathfrak{e}_i^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,2}-} - \boldsymbol{\eta}) \qquad (79) \\
&\quad - \sum_{l \in B_k, l \neq i} \big[\sum_{j=1}^{d} \gamma U_{ij} U_{lj} \sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2}\big] \mathfrak{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,2}-} - \boldsymbol{\eta}).
\end{aligned}$$

We will see that the first term will lead to the one including $Y^{\hbar}_{\tau_{k+1,2}-}$ with errors. From Lemma B.2, we have

$$\nu_{t,j} = e^{-\gamma\sigma_j^2 t}\nu_{0,j} + \int_0^t e^{-\gamma\sigma_j^2(t-s)} \gamma\sigma_j (\boldsymbol{U}^T\boldsymbol{\eta})_j \, \mathrm{d}s + e^{-\gamma\sigma_j^2 t}\widetilde{X}_{t,j},$$

and this gives with overwhelming probability

$$\begin{aligned}
Y^{\hbar}_{\tau_{k+1,2}-} &= \sum_{j=1}^{d} U_{ij}\sigma_j e^{-\mathfrak{t}\gamma\sigma_j^2} \widetilde{X}^{\hbar}_{(\tau_{k+1,2}-),j} \\
&= \sum_{j=1}^{d} U_{ij}\sigma_j \Big[\nu^{\hbar}_{(\tau_{k+1,2}-),j} - e^{-\mathfrak{t}\gamma\sigma_j^2}\nu_{0,j} - \int_0^{\mathfrak{t}} e^{-(\mathfrak{t}-s)\gamma\sigma_j^2}\gamma\sigma_j(\boldsymbol{U}^T\boldsymbol{\eta})_j \, \mathrm{d}s\Big] \\
&= \mathfrak{e}_i^T \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,2}-} - \sum_{j=1}^{d} U_{ij}\sigma_j e^{-\mathfrak{t}\gamma\sigma_j^2}\nu_{0,j} - \sum_{j=1}^{d}\gamma U_{ij}\sigma_j^2 (\boldsymbol{U}^T\boldsymbol{\eta})_j \int_0^{\mathfrak{t}} e^{-(\mathfrak{t}-s)\gamma\sigma_j^2}\,\mathrm{d}s \\
&= \mathfrak{e}_i^T \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_{k+1,2}-} + \mathcal{O}(n^{\theta-1/2}).
\end{aligned}$$

In the last line to get the order, we used Corollary C.1 for the second term and Lemma C.1 for the last term with setting $\boldsymbol{a} = \mathfrak{e}_i, \boldsymbol{b} = \boldsymbol{\eta}$ and conditioning on $\boldsymbol{\Sigma}$. See the arguments after (70) for detail. Thus, from (79) we have with overwhelming probability

$$\begin{aligned}
\Delta Y^{\hbar}_{\tau_{k+1,2}} &= -\sum_{j=1}^{d} \gamma U_{ij}^2 \sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2} (Y^{\hbar}_{\tau_{k+1,2}-} + \mathcal{O}(n^{\theta-1/2})) \\
&\quad - \sum_{l \in B_k, l \neq i} \big[\sum_{j=1}^{d} \gamma U_{ij} U_{lj} \sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2}\big] \mathfrak{e}_l^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_k} - \eta) \qquad (80) \\
&= -\sum_{j=1}^{d} \gamma U_{ij}^2 \sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2} Y^{\hbar}_{\tau_{k+1,2}-} + \mathcal{O}(n^{\theta-1/2}) + \mathcal{O}(\beta n^{\theta-1/2-\aleph}),
\end{aligned}$$

where Lemma C.1 was used again in the last line; when $l \neq i$,

$$\mathbb{E}\big[\sum_{j=1}^{d} \gamma U_{ij}U_{lj}\sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2}\big] = 0.$$

Note that condition (ii) is satisfied from (80) after some appropriate scaling of $Y_t^\hbar$, since

$$\begin{aligned} Y^\hbar_{\tau_{k+1,2}} &= Y^\hbar_{\tau_{k+1,2}-} + \Delta Y^\hbar_{\tau_{k+1,2}} \\ &= \Big(1 - \sum_{j=1}^{d} \gamma U_{ij}^2\sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2}\Big) Y^\hbar_{\tau_{k+1,2}-} + \mathcal{O}(n^{\theta-1/2}) + \mathcal{O}(\beta n^{\theta-1/2-\aleph}), \end{aligned} \tag{81}$$

and $\Big|1 - \sum_{j=1}^{d} \gamma U_{ij}^2\sigma_j^2 e^{-(q-\mathfrak{t})\gamma\sigma_j^2}\Big| < 1$ on $\mathcal{G}$.

Now, in view of (77), (78) and (81), scaling $Y_t^\hbar$ by $\max\{\sqrt{\beta}n^{2\theta-1/2}, \beta n^{\theta-1/2-\aleph}, n^{\theta-1/2}\}$ makes every condition for cases 1 and 2 valid, so Lemma D.2 gives

$$\sup_{0\le t\le T} |Y_t^\hbar| \le n^{2\theta-1/2}\max\{\sqrt{\beta}n^\theta, \beta n^{-\aleph}, 1\}. \tag{82}$$

We summarize the following conclusion: if we let, for any $\epsilon > 0$, $\psi_i^{(T)} \stackrel{\text{def}}{=} \max_{0\le t\le T} |\mathfrak{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})|$,

$$\psi_i^{(T)} \le n^{-\aleph} \text{ w.o.p.} \implies \psi_i^{(T)} \le n^{2\theta-1/2}\max\{\sqrt{\beta}n^\theta, \beta n^{-\aleph}, 1\} \text{ w.o.p.}$$

Thus under the assumption that $\beta \le n^{1/5-\delta} \le n^{1/2-\delta}$, and picking $\theta < \delta/4$ we conclude

$$\psi_i^{(T)} \le n^{-\aleph} \text{ w.o.p.} \implies \psi_i^{(T)} \le \max\{\sqrt{\beta}n^{3\theta-1/2}, n^{-\aleph-\delta/2}, n^{2\theta-1/2}\} \text{ w.o.p.}$$

Hence by iterating this inequality finitely many times, $\max\{\sqrt{\beta}n^{3\theta-1/2}, n^{-\aleph-\delta/2}, n^{2\theta-1/2}\}$ becomes $\sqrt{\beta}n^{3\theta-1/2}$ and the conclusion follows with the choice of $\theta < \min\{\delta/4, \epsilon/6\}$. The only thing left to check is to bound $\psi_i^{(T)}$ with the initial condition $\aleph = -\epsilon/2$, *i.e.*,

$$\max_{0\le t\le T} |\mathfrak{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t^\vartheta - \boldsymbol{\eta})| \le n^{\epsilon/2}$$

with overwhelming probability. But this was already given by Lemma B.4. □

D.2.2 Delocalization of the spectral weights: the proof of Proposition B.4

Proof of Proposition B.4. It is sufficient to prove the same claim for a fixed j. To take advantage of the main technical assumption Proposition B.2, we will introduce a stopping time $\hbar$, defined as (for some $\alpha \in (0, \frac{1}{2})$)

$$\hbar \stackrel{\text{def}}{=} \inf\left\{t \le \vartheta : \max_{1\le i\le n} \big(\mathfrak{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})\big)^2 > \beta n^{-2\alpha}\right\}. \tag{83}$$

As with overwhelming probability this does not occur, it suffices to show a bound for the stopped processes $\nu_{t,j}^\hbar \stackrel{\text{def}}{=} \nu_{t\wedge\hbar,j}$.

Using Lemma B.2, we have the representation

$$\nu_{t,j} = e^{-\gamma\sigma_j^2 t}\nu_{0,j} + \int_0^t e^{-\gamma\sigma_j^2(t-s)}\gamma\sigma_j(\boldsymbol{U}^T\boldsymbol{\eta})_j\,\mathrm{d}s + e^{-\gamma\sigma_j^2 t}\widetilde{X}_{t,j}.$$

By replacing $t \to t\wedge\hbar$, we have the same representation for $\nu_{t,j}^\hbar$. We let $\mathcal{G}$ be the event that

$$|(\boldsymbol{U}^T\boldsymbol{\eta})_j| \le n^{\epsilon/2-1/2} \quad \text{and} \quad \max_i \big|\mathfrak{e}_j^T\gamma\boldsymbol{\Sigma}^T\boldsymbol{U}^T\mathfrak{e}_i\big| \le n^{\epsilon/2-1/2}. \tag{84}$$

By Corollary C.2 this holds with overwhelming probability. The first two terms is $n^{\epsilon-1/2}$ with overwhelming probability, using Assumption 1.1. Hence it suffices to show that for any $\epsilon > 0$,

$$\sup_{0\le t\le T} |\widetilde{X}^{\hbar}_{t,j}| \le \beta n^{\epsilon-1/2}$$

with overwhelming probability.

The quadratic variation is, from (65)

$$[\widetilde{X}^{\hbar}_{t,j}] = \sum_{k=1}^{N_{t\wedge\hbar}} e^{2\tau_{k+1}\gamma\sigma_j^2} \big(\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1}(\boldsymbol{U}\boldsymbol{\Sigma}\nu^{\hbar}_{\tau_k} - \boldsymbol{\eta})\big)^2.$$

We observe that for $\tau_k \le \hbar$ on $\mathcal{G}$,

$$\big(\mathbb{e}_j^T \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1}(\boldsymbol{U}\boldsymbol{\Sigma}\nu_{\tau_k} - \boldsymbol{\eta})\big)^2 \le \beta^3 \max_i \big|\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \mathbb{e}_i\big|^2 n^{-2\alpha} \le C(T,\boldsymbol{\Sigma})\beta^3 n^{\epsilon-1-2\alpha}.$$

Using part *(i)* of Lemma D.1, we have that $\max_{0\le t\le T} |\nu^{\hbar}_{t,j}| \le \beta n^{\epsilon-\alpha}$ with overwhelming probability. □

D.2.3 Concentration of the function values: the proof of Proposition B.3

Proof of Proposition B.3. Recall that for fixed $q \in [0,T]$

$$\varepsilon^{(n)}_{\mathrm{M}}(q) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 \int_0^q e^{-2(q-s)\gamma\sigma_j^2}\, \mathrm{d}M_{s,j}.$$

Hence we can write this as

$$\varepsilon^{(n)}_{\mathrm{M}}(q) = \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 e^{-2q\gamma\sigma_j^2} X_{q,j}.$$

We consider the martingale

$$Y_t \stackrel{\text{def}}{=} \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 e^{-2q\gamma\sigma_j^2} X_{t,j},$$

and we show concentration for Y_t, $0 \le t \le q$. As in the proof of Proposition B.2, it would suffice to bound Y_q for fixed $q \in [0,T]$ using the mesh arguments because the probability of having multiple jumps between two adjacent mesh points converges to zero faster than any polynomial order. Also, Proposition B.2 allows us to adopt a stopping time $\hbar$, defined as (for some $\alpha \in (0, \frac{1}{2})$)

$$\hbar \stackrel{\text{def}}{=} \inf\left\{ t \le \vartheta : \max_{1\le i\le n} \big(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_t - \boldsymbol{\eta})\big)^2 > \beta n^{-2\alpha} \right\}. \tag{85}$$

As with overwhelming probability this does not occur, it suffices to show a bound for the stopped processes $\nu^{\hbar}_{t,j} \stackrel{\text{def}}{=} \nu_{t\wedge\hbar,j}$.

The jumps of this martingale are given by (see (65))

$$\Delta Y^{\hbar}_{\tau_k} = \frac{1}{2}\sum_{j=1}^{d} \sigma_j^2 e^{-2\gamma(q-\tau_k\wedge\hbar)\sigma_j^2} \big(\nu^{\hbar}_{\tau_k,j} + \nu^{\hbar}_{\tau_k-,j}\big)\big(\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}^{\hbar}_{\tau_k-} - \boldsymbol{\eta})\big).$$

Therefore, the quadratic variation is

$$
\begin{aligned}
[Y_q^\hbar] &= \sum_{k=1}^{N_q} (\Delta Y_{\tau_k}^\hbar)^2 \\
&= \sum_{k=1}^{N_q} \Big[\frac{1}{2} \sum_{j=1}^{d} \sigma_j^2 e^{-2\gamma(q-\tau_k\wedge\hbar)\sigma_j^2} \big(\nu_{\tau_k,j}^\hbar + \nu_{\tau_k-,j}^\hbar \big) \big(\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1} (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-}^\hbar - \boldsymbol{\eta}) \big) \Big]^2 \\
&= \frac{1}{4} \sum_{k=1}^{N_q} \Big[\sum_{j=1}^{d} \sigma_j^2 e^{-2\gamma(q-\tau_k\wedge\hbar)\sigma_j^2} \big(2\nu_{\tau_k-,j}^\hbar + \mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1} (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-}^\hbar - \boldsymbol{\eta}) \big) \\
&\qquad\qquad \cdot \big(\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1} (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-}^\hbar - \boldsymbol{\eta}) \big) \Big]^2 \\
&\le \frac{1}{2} \sum_{k=1}^{N_q} \Big[\sum_{i\in B_{k-1}} \Big(\sum_{j=1}^{d} \sigma_j^2 e^{-2\gamma(q-\tau_k\wedge\hbar)\sigma_j^2} 2\nu_{\tau_k-,j}^\hbar (\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \mathbb{e}_i) \Big) \big(\mathbb{e}_i^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-}^\hbar - \boldsymbol{\eta}) \big) \Big]^2 \\
&\quad + \frac{1}{2} \sum_{k=1}^{N_q} \Big[\sum_{j=1}^{d} \sigma_j^2 e^{-2\gamma(q-\tau_k\wedge\hbar)\sigma_j^2} \big(\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1} (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-}^\hbar - \boldsymbol{\eta}) \big)^2 \Big]^2 .
\end{aligned}
$$

Note that the second term is bounded as

$$
\begin{aligned}
&\frac{1}{2} \sum_{k=1}^{N_q} \Big[\sum_{j=1}^{d} \sigma_j^2 e^{-2\gamma(q-\tau_k\wedge\hbar)\sigma_j^2} \big(\sum_{i\in B_{k-1}} \big(\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \mathbb{e}_i \big) \big(\mathbb{e}_i^T (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\nu}_{\tau_k-}^\hbar - \boldsymbol{\eta}) \big)^2 \Big]^2 \\
&\quad \le \frac{N_q}{2} \big[n \cdot (\beta n^{\epsilon-1/2} \sqrt{\beta} n^{-\alpha})^2 \big]^2 = \frac{N_q \beta}{2n} \beta^5 n^{4\epsilon-4\alpha+1}.
\end{aligned}
$$

Note that Corollary C.2 was used to bound $\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \mathbb{e}_i$. As for the first term, define

$$
W_{q,v,s,i} \stackrel{\text{def}}{=} \sum_{j=1}^{d} \nu_{s,j}^\hbar e^{-2\gamma(q-v)\sigma_j^2} \mathbb{e}_j^T \boldsymbol{\Sigma}^T \boldsymbol{U}^T \mathbb{e}_i, \tag{86}
$$

for $0 \le v \le q, 0 \le s \le v$. Note that it suffices to bound $\max_{0\le s\le v} W_{q,v,s,i}$ for a fixed v, because we can apply the union bound to $0 \le v \le q$ using the same meshing arguments as the ones used after (71).

Observe,

$$
\begin{aligned}
W_{q,v,s,i} &= \sum_{j=1}^{d} e^{-2\gamma(q-v)\sigma_j^2} \sigma_j U_{ij} \big(e^{-\gamma\sigma_j^2 s} \nu_{0,j} + \int_0^s e^{-\gamma\sigma_j^2(s-u)} \gamma\sigma_j (\boldsymbol{U}^T \boldsymbol{\eta})_j \, \mathrm{d}u + e^{-\gamma\sigma_j^2 s} \widetilde{X}_{s,j}^\hbar \big) \\
&= \sum_{j=1}^{d} e^{-\gamma(2q-2v+s)\sigma_j^2} \sigma_j U_{ij} \nu_{0,j} + \sum_{j=1}^{d} \int_0^s e^{-\gamma(2q-2v+s-u)\sigma_j^2} \gamma\sigma_j^2 U_{ij} (\boldsymbol{U}^T \boldsymbol{\eta})_j \, \mathrm{d}u \\
&\quad + \sum_{j=1}^{d} e^{-\gamma 2(q-v)\sigma_j^2} e^{-\gamma s\sigma^2} \sigma_j U_{ij} \widetilde{X}_{s,j}^\hbar .
\end{aligned}
$$

Note that the first and second terms are of $\mathcal{O}(n^{\epsilon-1/2})$ with overwhelming probability by the arguments after (70). Also, the last term is bounded by $\sqrt{\beta} n^{\epsilon-1/2}$ w.o.p. by the same arguments used in showing the bound for Y_t defined in (71). It is crucial that the additional coefficients $e^{-\gamma 2(q-v)\sigma_j^2}$ are less than 1 so that the same arguments from (72) to (82) work.

Then the quadratic variation is bounded by

$$
[Y_q^\hbar] \le C N_q \big[\beta \sqrt{\beta} n^{\epsilon-1/2} \sqrt{\beta} n^{-\alpha} \big]^2 + \frac{N_q \beta}{2n} \beta^5 n^{4\epsilon-4\alpha+1} \le \frac{C N_q \beta}{nT} \beta^5 n^{\epsilon-4\alpha+1},
$$

with some $C > 0$ and small enough $\epsilon > 0$ in the last part, which will be an enough bound to apply Lemma D.1. Hence we conclude, with the same meshing arguments used in the proof of Proposition B.2 and assumption on $\beta \leq n^{1/5-\delta}$,

$$\sup_{0\leq t\leq T} |Y_t^{\hbar}| \leq \beta^{5/2} n^{\epsilon-2\alpha+1/2} \leq n^{\epsilon-5\delta/2+(1-2\alpha)},$$

and the claim follows by choosing $\alpha < 1/2$ sufficiently close to $1/2$. □

D.2.4 Concentration of cross-variation of the model noise with the spectral weights: the proof of Proposition B.6

Proof of Proposition B.6. We recall from the assumptions of the Proposition that we let $\{c_j\}_1^n$ be a deterministic sequence with $|c_j| \leq 1$ for all j. We should show that for any $t > 0$ and for some $\epsilon > 0$

$$\left| \frac{1}{\|\boldsymbol{\eta}\|^2} \sum_{j=1}^n c_j \sigma_j \nu_{t,j}^{\vartheta} (\boldsymbol{U}^T \boldsymbol{\eta})_j - \frac{1}{n} \sum_{j=1}^n c_j \big(1 - e^{-\gamma \sigma_j^2 (t\wedge\vartheta)}\big) \right| \leq \sqrt{\beta} n^{\epsilon - 1/2}$$

with overwhelming probability.

We begin again by using Lemma B.2, due to which we have the representation

$$\nu_{t,j} = e^{-\gamma\sigma_j^2 t} \nu_{0,j} + \int_0^t e^{-\gamma\sigma_j^2 (t-s)} \gamma \sigma_j (\boldsymbol{U}^T \boldsymbol{\eta})_j \, \mathrm{d}s + e^{-\gamma\sigma_j^2 t} \widetilde{X}_{t,j}.$$

We replace this expression into the sum we wish to control, and observe that

$$\sum_{j=1}^n c_j \sigma_j \nu_{t,j}^{\vartheta} (\boldsymbol{U}^T \boldsymbol{\eta})_j = \sum_{j=1}^n c_j \sigma_j (\boldsymbol{U}^T \boldsymbol{\eta})_j \bigg(e^{-\gamma\sigma_j^2 t} \nu_{0,j} + \int_0^t e^{-\gamma\sigma_j^2 (t-s)} \gamma \sigma_j (\boldsymbol{U}^T \boldsymbol{\eta})_j \, \mathrm{d}s + e^{-\gamma\sigma_j^2 t} \widetilde{X}_{t,j} \bigg).$$

Under Assumption 1.1 and Corollary C.1, the first sum vanishes with overwhelming probability. By independence of $\boldsymbol{\eta}$ from $\boldsymbol{U}$, we have that

$$\big((\boldsymbol{U}^T \boldsymbol{\eta})_j : 1 \leq j \leq n\big) \overset{\text{law}}{=} \|\boldsymbol{\eta}\| \big(U_{1,j} : 1 \leq j \leq n\big).$$

Hence, by Lemma C.1, for any $\epsilon > 0$, with overwhelming probability,

$$\left| \sum_{j=1}^n c_j \frac{(\boldsymbol{U}^T \boldsymbol{\eta})_j^2}{\|\boldsymbol{\eta}\|^2} \int_0^t e^{-\gamma\sigma_j^2 (t-s)} \gamma \sigma_j^2 ds - \sum_{j=1}^n \frac{c_j}{n} \int_0^t e^{-\gamma\sigma_j^2 (t-s)} \gamma \sigma_j^2 ds \right| \leq n^{\epsilon - 1/2}.$$

Using that $\int_0^t e^{-\gamma\sigma_j^2 (t-s)} \gamma \sigma_j^2 ds = (1 - e^{-\gamma\sigma_j^2 t})$, we have reduced the problem to showing that for any $\epsilon > 0$ with overwhelming probability

$$\big|Y_t\big| \leq \sqrt{\beta} n^{\epsilon - 1/2} \quad \text{where} \quad Y_s \overset{\text{def}}{=} \frac{1}{\|\boldsymbol{\eta}\|^2} \sum_{j=1}^n c_j \sigma_j (\boldsymbol{U}^T \boldsymbol{\eta})_j e^{-\gamma\sigma_j^2 t} \widetilde{X}_{s,j} \quad \text{for all } s \leq t.$$

To leverage Proposition B.2, we again use the stopping time $\hbar$ (83). We will again apply Lemma D.1. The jumps of $Y_t^{\hbar}$ are given by, for any $\tau_k \leq s$,

$$\Delta Y_{\tau_k}^{\hbar} = -\frac{1}{\|\boldsymbol{\eta}\|^2} \sum_{j=1}^n c_j \sigma_j (\boldsymbol{U}^T \boldsymbol{\eta})_j e^{-\gamma\sigma_j^2 t} \big(\mathbb{e}_j^T \gamma \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1} (\boldsymbol{U} \boldsymbol{\Sigma} \nu_{\tau_k -}^{\hbar} - \boldsymbol{\eta}) \big).$$

If we let $\boldsymbol{D}$ be the diagonal matrix with entries $-\gamma c_j e^{-\gamma\sigma_j^2 t}$, then we have the representation

$$\Delta Y_{\tau_k}^{\hbar} = \frac{1}{\|\boldsymbol{\eta}\|^2} \boldsymbol{\eta}^T \boldsymbol{U} \boldsymbol{D} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^T \boldsymbol{U}^T \boldsymbol{P}_{k-1} (\boldsymbol{U} \boldsymbol{\Sigma} \nu_{\tau_k -}^{\hbar} - \boldsymbol{\eta}).$$

Using that $\frac{1}{\|\boldsymbol{\eta}\|^2} \boldsymbol{\eta}^T \boldsymbol{U} \boldsymbol{D} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^T \boldsymbol{U}^T$ has a norm bounded only in terms of $t, \boldsymbol{\Sigma}$ and γ, it follows that

$$|\Delta Y_{\tau_k}^{\hbar}| \leq C(t, \boldsymbol{\Sigma}, \gamma) \beta n^{-\alpha}. \tag{87}$$

We turn to bounding the predictable quadratic variation. Using Lemma B.1,

$$\mathbb{E}((\Delta Y^{\hbar}_{\tau_k})^2 \mid \mathcal{F}_{\tau_k -}) = \frac{\beta(\beta-1)}{n(n-1)}\left(\frac{1}{\|\boldsymbol{\eta}\|^2}\boldsymbol{\eta}^T \boldsymbol{U}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T(\boldsymbol{U}\boldsymbol{\Sigma}\nu^{\hbar}_{\tau_k -} - \boldsymbol{\eta})\right)^2 + \left(\frac{\beta}{n} - \frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{i=1}^{n}\left(\frac{1}{\|\boldsymbol{\eta}\|^2}\boldsymbol{\eta}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T\mathbb{e}_i\right)^2\left(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\nu^{\hbar}_{\tau_k -} - \boldsymbol{\eta})\right)^2.$$

The first line we bound using that $\boldsymbol{U}\boldsymbol{\Sigma}\nu^{\hbar}_{\tau_k -} - \boldsymbol{\eta}$ is norm at most n^{ϵ} and that $\boldsymbol{U}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T$ has a norm bounded only by some $C(t, \boldsymbol{\Sigma}, \gamma)$. The second line we bound using that $(\mathbb{e}_i^T(\boldsymbol{U}\boldsymbol{\Sigma}\nu^{\hbar}_{\tau_k -} - \boldsymbol{\eta}))^2 \le \beta n^{-2\alpha}$. Together, these bounds give that

$$\mathbb{E}((\Delta Y^{\hbar}_{\tau_k})^2 \mid \mathcal{F}_{\tau_k -}) \le C(t, \boldsymbol{\Sigma}, \gamma)\left(\beta^2 n^{-2+\epsilon} + \beta^2 n^{-1-2\alpha}\right).$$

Hence the conclusion follows using Lemma D.1. □

E Analyzing the Volterra equation

E.1 General analysis of the Volterra equation

In this section, we analyze the solution of the Volterra equation (12) and give some basic properties of its solution for general limiting spectral measures μ which is a compactly supported measure on $[0, \infty)$. We recall for convenience that the Volterra equation is given by

$$\begin{aligned} \psi_0(t) &= \tfrac{R}{2}h_1(t) + \tfrac{\widetilde{R}}{2}\big(rh_0(t) + (1-r)\big) + \gamma^2 r\int_0^t h_2(t-s)\psi_0(s)\,\mathrm{d}s, \\ \text{and} \quad h_k(t) &= \int_0^{\infty} x^k e^{-2\gamma t x}\,\mathrm{d}\mu(x), \end{aligned} \tag{88}$$

where $\gamma > 0$ is a stepsize parameter. When convenient, we will simply write $z(t)$ for the forcing function

$$z(t) \stackrel{\text{def}}{=} \tfrac{R}{2}h_1(t) + \tfrac{\widetilde{R}}{2}\big(rh_0(t) + (1-r)\big). \tag{89}$$

The parameter $r \in (0, \infty)$ is fixed, but we may consider limits of the Volterra equation under various limits. The parameters R and $\widetilde{R}$ are both non-negative, but to avoid trivialities, we should assume at least one is positive. As the Volterra equation is linear, we may without loss of generality assume $R + \widetilde{R} = 1$.

The equation (88) appears frequently in the probability literature as the *renewal equation* [Resnick, 1992, (3.5.1)], [Asmussen, 2003, (2.1)]; it appears naturally in renewal theory and in the Lotka population model, amongst others, which are neatly described in the references just mentioned. It allows, for example, ψ_0 to be given the amusing interpretation as the expected size of a population which evolves in times (c.f. [Resnick, 1992, Example 3.5.2] or [Asmussen, 2003, Example 2.2]). Much of the behavior of the equation is determined by the properties of the function $\gamma^2 r h_2(t)$. We will let λ^- be the leftmost endpoint of the support of μ restricted to $(0, \infty)$, and we record the following elementary computation. For any $\alpha \in \mathbb{R}$,

$$\int_0^{\infty} e^{2\gamma\alpha t}\gamma^2 r h_2(t)\,\mathrm{d}t = \begin{cases} \frac{\gamma r}{2}\int_0^{\infty}\frac{x^2}{x-\alpha}\,\mathrm{d}\mu(x), & \text{if } \alpha \le \lambda^- \\ \infty & \text{otherwise.} \end{cases} \tag{90}$$

We begin by observing some elementary properties of the equation:

Theorem E.1. *There is a unique, positive solution to* (12) *which exists for all time. The solution is bounded if and only if* $\gamma < \frac{2}{r}\big(\int_0^{\infty} x\,\mathrm{d}\mu(x)\big)^{-1}$ *in which case*

$$\psi_0(\infty) \stackrel{\text{def}}{=} \lim_{t\to\infty}\psi_0(t) = \frac{\widetilde{R}}{2}\cdot\frac{r\mu(\{0\}) + (1-r)}{1 - \frac{\gamma r}{2}\big(\int_0^{\infty} x\,\mathrm{d}\mu(x)\big)}.$$

Proof. In standard renewal notation (c.f. [Resnick, 1992, (3.5.1)], [Asmussen, 2003, (2.1)]), we would write (88)

$$\psi_0 = z + \psi_0 * F$$

where F is the function

$$F(t) = \int_0^t \gamma^2 r h_2(s)\,\mathrm{d}s.$$

The existence and uniqueness is now standard (compare with Proposition B.1), see [Resnick, 1992, Theorem 3.5.1] or [Asmussen, 2003, Theorem 2.4]. By (90),

$$F(\infty) = \frac{\gamma r}{2}\int_0^\infty x\,\mathrm{d}\mu(x).$$

Hence when this is bigger than 1, the solution $\psi_0(t)$ tends to infinity exponentially fast [Asmussen, 2003, Theorem 7.1] or [Resnick, 1992, Proposition 3.11.1]. In the case that $\frac{\gamma r}{2}\int_0^\infty x\,\mathrm{d}\mu(x) = 1$, by Blackwell's Renewal theorem, $\psi_0(t)$ is asymptotic to a positive multiple of t (see [Resnick, 1992, Theorem 3.10.1] or [Asmussen, 2003, Theorem 4.4]) and hence still diverges. Finally, in the case that $\frac{\gamma r}{2}\int_0^\infty x\,\mathrm{d}\mu(x) < 1$ by [Resnick, 1992, Section 3.11] or [Asmussen, 2003, Proposition 7.4],

$$\lim_{t\to\infty}\psi_0(t) = \frac{\lim_{t\to\infty} z(t)}{1 - \frac{\gamma r}{2}\int_0^\infty x\,\mathrm{d}\mu(x)}, \tag{91}$$

which is the claimed result. □

Two phases

Hence, we will assume going forward that $\gamma < \gamma_0 \stackrel{\text{def}}{=} \left(\frac{r}{2}\int_0^\infty x\,\mathrm{d}\mu(x)\right)^{-1}$. We shall see that it is possible to say more about the rate of convergence in general. Define the *Malthusian exponent* λ^* as the solution of

$$\int_0^\infty e^{2\gamma\lambda^* t}\gamma^2 r h_2(t)\,\mathrm{d}t = 1, \tag{92}$$

when it exists. Note by virtue of (90), if this exponent exists, it can just as well be defined as the solution of

$$\frac{r}{2}\int_0^\infty \frac{x^2}{x-\lambda^*}\,\mathrm{d}\mu(x) = \frac{1}{\gamma}, \tag{93}$$

and we necessarily have that $\lambda^* \le \lambda^-$. Define

$$\gamma_* = \frac{1}{\frac{r}{2}\int_0^\infty \frac{x^2}{x-\lambda^-}\,\mathrm{d}\mu(x)} \tag{94}$$

which exists and is positive exactly when $\int_0^\infty \frac{x^2}{x-\lambda^-}\,\mathrm{d}\mu(x) < \infty$. Note that γ_* is strictly less than γ_0 if and only if $\lambda^- > 0$. Moreover, we can completely give the asymptotic behavior of the Volterra equation on either side of the critical point. Although, to do this for $\gamma < \gamma_*$, we will need some further assumptions on μ.

Recall that a function $f : (0,\infty) \to \mathbb{R}$ is *slowly varying* if $f(tx)/f(x) \to 1$ as $t \to \infty$ for any $x > 0$. A function $f : (0,\infty) \to \mathbb{R}$ is *regularly varying* if $f(t) = g(t)t^\alpha$ for a slowly varying function g. We will say that μ is *left-edge-regular* if there exists a regularly varying function L and $\alpha > 0$ so that

$$t \mapsto \mu((\lambda^-, \lambda^- + t]) \sim t^\alpha L(\tfrac{1}{t}), \quad \text{as } t \to \infty, \tag{95}$$

which for example is satisfied by Marchenko-Pastur (8). We show the following:

Theorem E.2. *For $\gamma \in (\gamma_*, \gamma_0)$, the Malthusian exponent λ^* exists and is the unique solution of (93). The function $\psi_0(t)$ satisfies that for some explicit constant $c(R, \widetilde{R}, \mu)$,*

$$\psi_0(t) - \psi_0(\infty) \sim \frac{c(R,\widetilde{R},\mu)}{\gamma} e^{-2\gamma(\lambda^*)t}.$$

If in addition $\gamma_* > 0$ and μ is left-edge-regular, then for $\gamma \in (0, \gamma_*)$

$$\psi_0(t) - \psi_0(\infty) \sim e^{-2\gamma(\lambda^-)t} g(t)$$

where $g(t)$ is some explicit regularly varying function.

Thus if one considers varying the stepsize γ from 0 up to γ_* the process undergoes a transition in behavior when $\gamma = \gamma_*$. For small γ, the exponential rate of change is frozen on the smallest eigenvalue of the Hessian λ^-. However as γ passes the transition point γ_*, the logarithm of the rate becomes a smooth function. This is strongly reminiscent of a *freezing transition*, which is often seen in the free energies of random energy models. See for example Fyodorov and Bouchaud [2008].

The proof is essentially an automatic consequence of established theory for Volterra equations. As an input to the case of $\gamma < \gamma_*$, we need the following asymptotics of the functions h_k, which are the main way in which left-edge-regularity:

Lemma E.1. *Suppose that μ has left-edge-regularity, meaning that there is an $\alpha \geq 0$ and slowly varying function L so that*

$$\mu((\lambda^-, \lambda^- + t]) \sim t^\alpha L(\tfrac{1}{t}) \quad \text{as} \quad t \to 0.$$

If $\lambda^- > 0$ then for any $k \geq 0$

$$h_k(t) - (\lambda^-)^k e^{-2\gamma\lambda^- t}\mu(\{\lambda^-\}) \sim e^{-2\gamma\lambda^- t} t^{-\alpha} L(t)\Gamma(\alpha+1)(\lambda^-)^k.$$

If $\lambda^- = 0$ then for any $k \geq 0$,

$$h_k(t) - \mathbb{1}_{k=0}\mu(\{0\}) \sim t^{-k-\alpha} L(t)\Gamma(k+\alpha+1).$$

This is a standard exercise, and we do not show its proof.

Proof of Theorem E.2. **The case of $\gamma \in (\gamma_*, \gamma_0)$.** We follow the notation of [Asmussen, 2003, Theorem 7.1] (see also [Resnick, 1992, Proposition 3.11.1]). Before beginning, we observe that the Malthusian exponent does exist for this region, as the function $\alpha \mapsto \int_0^\infty \frac{x^2}{x-\alpha}\,\mathrm{d}\mu(x)$, is an increasing continuous function on $(-\infty, \lambda^-)$. Hence by the definition of γ_*, the image of this function applied to $[0, 2\gamma\lambda^-)$ is all of $[\gamma_0^{-1}, \gamma_*^{-1})$. From [Asmussen, 2003, Proposition 7.6]

$$\lim_{t\to\infty} e^{2\gamma\lambda^* t}(\psi_0(t) - \psi_0(\infty)) = \frac{\int_0^\infty e^{2\gamma\lambda^* t}(z(t) - z(\infty))\,\mathrm{d}t - \frac{z(\infty)}{\beta}}{\int_0^\infty t e^{2\gamma\lambda^* t}\gamma^2 r h_2(t)\,\mathrm{d}t}.$$

We evaluate these two integrals for convenience. Using the definition of z in (89)

$$\int_0^\infty e^{2\gamma\lambda^* t}(z(t) - z(\infty))\,\mathrm{d}t - \frac{z(\infty)}{\beta} = \frac{R}{4\gamma}\int_0^\infty \frac{x\,\mathrm{d}\mu(x)}{x-\lambda^*} + \frac{\widetilde{R}r}{4\gamma}\int_0^\infty \frac{\mathrm{d}\mu(x)}{x-\lambda^*} - \frac{\widetilde{R}}{4\gamma\lambda^*}(1-r).$$

For the denominator,

$$\begin{aligned}\int_0^\infty t e^{2\gamma\lambda^* t}\gamma^2 r h_2(t)\,\mathrm{d}t &= \gamma^2 r \int_0^\infty \int_0^\infty x^2 t e^{2\gamma(\lambda^* - x)t}\,\mathrm{d}t\,\mathrm{d}\mu(x) \\ &= \frac{r}{4}\int_0^\infty \frac{x^2}{(x-\lambda^*)^2}\,\mathrm{d}\mu(x).\end{aligned}$$

The case of $\gamma \in (0, \gamma_*)$. By assumption we have that $\gamma_* > 0$, and hence

$$\int_0^\infty \frac{x^2}{x-\lambda^-}\,\mathrm{d}\mu(x) < \infty.$$

Set $F(t) = \int_0^t \gamma^2 r h_2(s)\,\mathrm{d}s$. Using that $\psi_0(\infty) = \frac{z(\infty)}{1-F(\infty)}$ (see (91)),

$$\psi_0(\infty) = z(\infty)\left(\tfrac{1-F(t)}{1-F(\infty)}\right) + F(t)\psi_0(\infty),$$

and hence the constant function ψ_0 solves the Volterra equation (88) with forcing function $z(\infty)\left(\frac{1-F(t)}{1-F(\infty)}\right)$. It follows that

$$\psi_0(t) - \psi_0(\infty) = z(t) - z(\infty)\left(\tfrac{1-F(t)}{1-F(\infty)}\right) + \gamma^2 r\int_0^t h_2(t-s)(\psi_0(s) - \psi_0(\infty))\,\mathrm{d}s. \tag{96}$$

Define

$$\widehat{Z}(t) = e^{2\gamma\lambda^- t}\big(\psi_0(t) - \psi_0(\infty)\big), \quad \widehat{z}(t) = e^{2\gamma\lambda^- t}\big(z(t) - z(\infty)\tfrac{1-F(t)}{1-F(\infty)}\big) \quad \text{and} \quad \frac{\mathrm{d}\widehat{F}(t)}{\mathrm{d}t} = \gamma^2 r e^{2\gamma\lambda^- t} h_2(t).$$

Then using (96),

$$\widehat{Z}(t) = \widehat{z}(t) + \int_0^t \frac{\mathrm{d}\widehat{F}(s)}{\mathrm{d}s}\widehat{Z}(t-s)\,\mathrm{d}s.$$

By the assumption that $\gamma < \gamma_*$, we have that

$$\theta \stackrel{\text{def}}{=} \int_0^\infty \gamma^2 r e^{2\gamma\lambda^- t} h_2(t)\,\mathrm{d}t < 1.$$

In preparation to apply [Asmussen et al., 2003, Theorem 5], we need to evaluate the ratio of the limits of the densities

$$\lim_{t\to\infty}\frac{\widehat{z}(t)}{\widehat{F}'(t)} = \lim_{t\to\infty}\frac{e^{2\gamma\lambda^- t}\big(z(t) - z(\infty)\frac{1-F(t)}{1-F(\infty)}\big)}{\gamma^2 r e^{2\gamma\lambda^- t} h_2(t)} = \lim_{t\to\infty}\left(\frac{z(t)-z(\infty)}{\gamma^2 r h_2(t)} + \frac{z(\infty)}{\gamma^2 r}\frac{F(t)-F(\infty)}{h_2(t)(1-F(\infty))}\right). \tag{97}$$

To simplify this, we observe that

$$F(\infty) - F(t) = \frac{\gamma r}{2}\int_0^\infty x e^{-2\gamma t x}\,\mathrm{d}\mu(x) = \frac{\gamma r}{2}h_1(t).$$

We also observe that

$$z(t) - z(\infty) = \frac{R}{2}h_1(t) + \frac{\widetilde{R}r}{2}h_{0^+}(t) \quad \text{where} \quad h_{0^+}(t) = \lim_{\epsilon\downarrow 0}\int_\epsilon^\infty e^{-2\gamma t x}\,\mathrm{d}\mu(x).$$

We conclude that

$$\lim_{t\to\infty}\frac{\widehat{z}(t)}{\widehat{F}'(t)} = \lim_{t\to\infty}\left(\frac{Rh_1(t) + \widetilde{R}r h_{0^+}(t)}{2\gamma^2 r h_2(t)} - \frac{\psi_0(\infty)}{2\gamma}\frac{h_1(t)}{h_2(t)}\right). \tag{98}$$

Hence we have from (98)

$$\lim_{t\to\infty}\frac{\widehat{z}(t)}{\widehat{F}'(t)} = \frac{(R - \gamma r\psi_0(\infty))\lambda^- + \widetilde{R}r}{2\gamma^2 r(\lambda^-)^2} \stackrel{\text{def}}{=} c_*. \tag{99}$$

By the assumption on α, it can be checked that $\int_0^\infty \widehat{z}(t)dt < \infty$. Moreover it follows that $\widehat{z}(t)$ is subexponential as $\alpha > 0$ (see [Asmussen et al., 2003, Section 3]), and hence by [Asmussen et al., 2003, Theorem 5 (ii)],

$$\begin{aligned}\widehat{Z}(t) &\sim \left(\frac{\int_0^\infty \widehat{z}(t)\,\mathrm{d}t}{(1-\theta)^2} + \frac{c_*}{1-\theta}\right)\widehat{F}'(t) \sim \left(\frac{\int_0^\infty \widehat{z}(t)\,\mathrm{d}t}{(1-\theta)^2} + \frac{c_*}{1-\theta}\right)\gamma^2 r e^{2\gamma\lambda^- t}h_2(t).\\ &\sim \left(\frac{\int_0^\infty \widehat{z}(t)\,\mathrm{d}t}{(1-\theta)^2} + \frac{c_*}{1-\theta}\right)\gamma^2 r\left((\lambda^-)^2\mu(\{\lambda^-\}) + t^{-\alpha}L(t)\Gamma(\alpha+1)(\lambda^-)^2\right).\end{aligned}$$

The second line follows from Lemma E.1. □

We finish by observing that when $\lambda^- = 0$, another behavior takes hold.

Theorem E.3. *Suppose that $\lambda^- = 0$, and that the measure μ has left-edge-regularity, meaning that there is an $\alpha > 0$ and slowly varying function L so that*

$$\mu((\lambda^-, \lambda^- + t]) \sim t^\alpha L(\tfrac{1}{t}) \quad \textit{as} \quad t \to 0.$$

Then

$$\psi_0(t) - \psi_0(\infty) \sim \frac{1}{1 - \frac{\gamma r}{2}\int_0^\infty x\,\mathrm{d}\mu(x)}\begin{cases}\frac{\widetilde{R}r}{2}t^{-\alpha}L(t)\Gamma(1+\alpha) & \textit{if } \widetilde{R} > 0,\\ \frac{R}{2}t^{-1-\alpha}L(t)\Gamma(2+\alpha) & \textit{if } \widetilde{R} = 0.\end{cases}$$

Proof. We again apply [Asmussen et al., 2003, Theorem 5], and so we use the same change of variables as in Theorem E.2. We once more must compute (98), which by Lemma E.1 is now equal to ∞. Since $\gamma < \gamma_0$,

$$\theta \stackrel{\text{def}}{=} \int_0^\infty \gamma^2 r h_2(t)\, \mathrm{d}t = \frac{\gamma r}{2} \int_0^\infty x \, \mathrm{d}\mu(x) < 1.$$

Hence by [Asmussen et al., 2003, Theorem 5 (iii)],

$$\widehat{Z}(t) \sim \frac{1}{1-\theta}\widehat{z}(t) \sim \frac{1}{1-\theta} \begin{cases} \frac{\widetilde{R}r}{2} \int_{0+}^\infty e^{-2\gamma t x} \, \mathrm{d}\mu(x) & \text{if } \widetilde{R} > 0, \\ \frac{R}{2} \int_0^\infty x e^{-2\gamma t x} \, \mathrm{d}\mu(x) & \text{if } \widetilde{R} = 0. \end{cases}$$

Then by Lemma E.1, the proof is complete. □

E.2 Explicit solution of the Volterra equation for Isotropic Features

In this section, we solve the Volterra equation for $\psi_0(t)$ in (12) when $\mathrm{d}\mu$ satisfies the Marchenko-Pastur law in (8). Throughout this section we use the following change of variables

$$\widehat{\psi}_0(t) \stackrel{\text{def}}{=} 2\psi_0\big(\tfrac{t}{2\gamma}\big).$$

Under this change of variables, the Volterra equation in (12) becomes

$$\begin{aligned} \widehat{\psi}_0(t) &= R \cdot h_1\big(\tfrac{t}{2\gamma}\big) + \widetilde{R}\big(r h_0\big(\tfrac{t}{2\gamma}\big) + (1-r)\big) + \frac{r\gamma}{2} \int_0^t h_2\big(\tfrac{1}{2\gamma}(t-s)\big)\widehat{\psi}_0(s)\, \mathrm{d}s \\ &= R \cdot \widehat{h}_1(t) + \widetilde{R}\big(r\widehat{h}_0(t) + (1-r)\big) + \int_0^t k(t-s)\widehat{\psi}_0(s)\, \mathrm{d}s, \end{aligned} \tag{100}$$

where we set $\widehat{h}_k$ a scaled version of h_k and the kernel k as

$$\widehat{h}_k(t) \stackrel{\text{def}}{=} h_k\big(\tfrac{t}{2\gamma}\big) \quad \text{and} \quad k(t) \stackrel{\text{def}}{=} \frac{r\gamma}{2} h_2\big(\tfrac{t}{2\gamma}\big). \tag{101}$$

Volterra equations of convolution type can. be solved trivially by using Laplace transforms which conveniently in the case of Marchenko-Pastur, we do have. Explicit formulas for the Laplace transforms of $h_k(t)$ via the Stieltjes transform of μ_{MP} exist.

We now solve for $\widehat{\psi}_0$ using Laplace transforms. We let $\Psi(p)$ and $K(p)$ be the Laplace transforms of $\widehat{\psi}_0(t)$ and $k(t)$ respectively. We can relate $\widehat{h}_1(t)$ in (101) to the function k and hence it's Laplace transform by the following

$$\partial_t \widehat{h}_1(t) = -R \int_0^\infty x^2 e^{-tx} \, \mathrm{d}\mu_{\mathrm{MP}}(x) = -\frac{2R}{r\gamma} k(t) \quad \text{and} \quad \mathcal{L}\{\widehat{h}_1(t)\} = \frac{R\left(1 - \frac{2}{r\gamma}K(p)\right)}{p}, \tag{102}$$

where we used that the first moment of μ_{MP} is 1 [Bai and Silverstein, 2010]. We now define the function $T(t)$ to be the Laplace transform of Marchenko-Pastur and the Laplace transform of T (*i.e.*, the Laplace transform of the Laplace transform of $\mu_{\mu_{\mathrm{MP}}}$), otherwise known as the Stieltjes transform, as the following

$$T(t) \stackrel{\text{def}}{=} \int_0^\infty e^{-xt} \, \mathrm{d}\mu_{\mathrm{MP}} \quad \text{and} \quad \mathcal{L}\{T(t)\}(p) = \frac{-p + r - 1 - \sqrt{(-p-r-1)^2 - 4r}}{2rp}. \tag{103}$$

It is clear that the Laplace transform for $\widehat{h}_0$ is given by using $\mathcal{L}\{T(t)\}$. From the Volterra equation (100) and the function $\widehat{h}_0(t)$ (101), we get the following expression for $\Psi(p)$:

$$\begin{aligned} \Psi(p) &= \frac{R\left(1 - \frac{2}{r\gamma}K(p)\right)}{p} + K(p)\Psi(p) + \widetilde{R}\left(r\mathcal{L}\{T(t)\}(p) + \frac{(1-r)}{p}\right) \\ \Psi(p) &= \frac{\frac{R\left(1-\frac{2}{r\gamma}K(p)\right)}{p} + \widetilde{R}\left(r\mathcal{L}\{T(t)\}(p) + \frac{(1-r)}{p}\right)}{1 - K(p)}. \end{aligned} \tag{104}$$

We now turn to giving an explicit expression for $\Psi(p)$. We begin with the following lemma relating the integral, $\int \frac{x}{x+p}\,\mathrm{d}\mu_{\mathrm{MP}}$, to the Stieltjes transform of the semi-circle law. We let m be the Stieltjes transform for semi-circle law

$$m(z) \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_{-2}^{2}\frac{\sqrt{4-y^2}}{y-z}\,\mathrm{d}y, \tag{105}$$

and we record a few well-known properties of this Stieltjes transform:

Lemma E.2. *The Stieltjes transform m can be expressed as*

$$m(z) = \frac{-z+\sqrt{z^2-4}}{2},$$

for $z \in \mathbb{C}$ with $\Im z > 0$, and it maps to the upper half plane (in fact to the upper half-disk). This can be extended to $z \in \mathbb{R}\setminus[-2,2]$ by continuity, and to the lower half plane by conjugation symmetry. The function m is the solution of the functional equation

$$m(z) + \frac{1}{m(z)} = -z \quad \textit{for all} \quad \Im z > 0. \tag{106}$$

Hence, we define the conjugate of m as

$$\widehat{m}(z) \stackrel{\textit{def}}{=} \frac{1}{m(z)} = \frac{-z-\sqrt{z^2-4}}{2}. \tag{107}$$

Moreover the Stieltjes transform of m is related to the Marchenko-Pastur by the identity for $p \in \mathbb{C}\setminus[\lambda^-,\lambda^+]$

$$\int_0^\infty \frac{x}{x+p}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{m(q)}{\sqrt{r}}, \tag{108}$$

where we set $q \stackrel{\textit{def}}{=} -\frac{p+1+r}{\sqrt{r}}$.

Proof. The results regarding the Stieltjes transform m are well-known and we refer the reader to [Bai and Silverstein, 2010]. It remains to prove (108) relating m to the Marchenko-Pastur. First we observe that

$$\int_0^\infty \frac{x}{x+p}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{1}{2\pi r}\int_{\lambda^-}^{\lambda^+}\frac{\sqrt{(x-\lambda^-)(\lambda^+-x)}}{x+p}\,\mathrm{d}x.$$

We recenter and rescale by sending $x = \frac{\lambda^-+\lambda^+}{2} + \frac{\lambda^+-\lambda^-}{4}y$, so that the follow holds

$$\sqrt{(x-\lambda^-)(\lambda^+-x)} = \sqrt{r}\cdot\sqrt{4-y^2}.$$

Using this change of variables and noting that $\frac{\lambda^+-\lambda^-}{4} = \sqrt{r}$, $\mathrm{d}x = \sqrt{r}\,\mathrm{d}y$, and $\frac{\lambda^++\lambda^-}{2} = 1+r$, we deduce that

$$\begin{aligned}\int_0^\infty \frac{x}{x+p}\,\mathrm{d}\mu_{\mathrm{MP}}(x) &= \frac{1}{2\pi}\int_{-2}^{2}\frac{\sqrt{4-y^2}}{p+\left(\frac{\lambda^-+\lambda^+}{2}+\frac{\lambda^+-\lambda^-}{4}y\right)}\,\mathrm{d}y \\ &= \frac{1}{2\pi\sqrt{r}}\int_{-2}^{2}\frac{\sqrt{4-y^2}}{y-\left(\frac{-p-(1+r)}{\sqrt{r}}\right)}\,\mathrm{d}y.\end{aligned}$$

The result follows after noting the definition of $m(z)$ and q. □

By exploiting the relationship between Marchenko-Pastur and the Stieltjes transform m defined in (105), we can evaluate some expressions against Marchenko-Pastur. To do so, it will be important to define the following quantities

$$\varrho = \frac{1+r}{2}\left(1-\frac{r\gamma}{2}\right) \quad \text{and} \quad \omega = \frac{1}{4}\left(1-\frac{r\gamma}{2}\right)^2\left(\frac{8}{\gamma}-(1+r)^2\right). \tag{109}$$

Lemma E.3 (Marchenko-Pastur Integrals)**.** *Suppose the constants ϱ and ω are as in (109) and fix the stepsize $0 < \gamma < \frac{2}{r}$. Define a critical stepsize γ_* as*

$$\gamma_* \stackrel{def}{=} \frac{2}{\sqrt{r}(r-\sqrt{r}+1)}. \tag{110}$$

It then follows that

$$\int_0^\infty \frac{x}{x+p_*}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \begin{cases} \left(1-\frac{r\gamma}{2}\right)^{-1}\frac{\gamma(\varrho+i\sqrt{\omega})}{2}, & \text{if } p_* = -\varrho - i\sqrt{\omega} \text{ and } \gamma < \frac{2}{r} \\ \left(1-\frac{r\gamma}{2}\right)^{-1}\frac{\gamma(\varrho-i\sqrt{\omega})}{2}, & \text{if } p_* = -\varrho + i\sqrt{\omega} \text{ and } \gamma \le \gamma_* \\ \left(1-\frac{r\gamma}{2}\right)\frac{2(\varrho+i\sqrt{\omega})}{r\gamma(\varrho^2+\omega)}, & \text{if } p_* = -\varrho + i\sqrt{\omega} \text{ and } \gamma_* < \gamma < \frac{2}{r}. \end{cases} \tag{111}$$

Proof. First suppose that $\omega < 0$, then the follow holds

$$\varrho + \sqrt{|\omega|} = \frac{1}{2}\left(1-\frac{r\gamma}{2}\right)\left(1+r+\sqrt{(1+r)^2-\tfrac{8}{\gamma}}\right).$$

We wish to show exactly when this quantity is equal to $(1-\sqrt{r})^2$ as this will give us the critical γ_*. Let $x = \sqrt{(1+r)^2-\frac{8}{\gamma}}$ and observe that $1+r-x \ge 0$. Hence we have that

$$\begin{aligned} 2(\varrho+\sqrt{|\omega|}-(1-\sqrt{r})^2)(1+r-x) &= \left(\tfrac{8}{\gamma}-4r\right) - 2(1-\sqrt{r})^2(1+r-x) \\ &= \left((1+r)^2-4r-x^2\right) - 2(1-\sqrt{r})^2(1+r-x) \\ &= -(x-(1-\sqrt{r})^2)^2. \end{aligned}$$

Thus $\varrho+\sqrt{|\omega|} < (1-\sqrt{r})^2$ except at a single value of γ at which

$$(1+r)^2-\tfrac{8}{\gamma} = (1-\sqrt{r})^4 \quad\Longleftrightarrow\quad \gamma = \gamma_* \stackrel{\text{def}}{=} \frac{2}{\sqrt{r}(r-\sqrt{r}+1)}. \tag{112}$$

Now we let p_* be either of $-\varrho \pm i\sqrt{\omega}$, where we take the branch of the square root continuous in the closed upper half plane (where $\omega \in \mathbb{C}$). We use the identity for p in $\mathbb{C}\setminus[\lambda^-,\lambda^+]$,

$$\int_0^\infty \frac{x}{x+p}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{m(q)}{\sqrt{r}},$$

where we recall $-q = \frac{p+1+r}{\sqrt{r}}$. We will apply this at p_*, and we set $-q_* = \frac{p_*+1+r}{\sqrt{r}}$.

We use the identity

$$\frac{\gamma}{2}[(p+\varrho)^2+\omega] = \big(1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}m(q)\big)\big(1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}\widehat{m}(q)\big).$$

In particular, evaluating this identity at $p=p_*$ the left-hand-side is 0, and we conclude that either

$$\big(1-\frac{r\gamma}{2}+\frac{r\gamma p_*}{2\sqrt{r}}m(q_*)\big)=0 \quad \text{or} \quad \big(1-\frac{r\gamma}{2}+\frac{r\gamma p_*}{2\sqrt{r}}\frac{1}{m(q_*)}\big)=0 \tag{113}$$

When $\omega > 0$, then as $1-\frac{r\gamma}{2} > 0$, the correct choice is dictated by having either $p_* m(q_*) < 0$ or $\frac{p_*}{m(q_*)} < 0$. If $\Im p_* = -\Im q_* < 0$, we have $\Im m(q_*) > 0$, and so it must be the second of these two identities in (113). Likewise, if $\Im p_* = -\Im q_* > 0$, then $\Im m(q_*) < 0$, and it is again the second of these identities.

For $\omega < 0$, we argue by continuity. For $p_* = -\varrho - i\sqrt{\omega}$, the mapping $\frac{1}{\gamma} \mapsto p_*$ is a continuous function for $\gamma \le \frac{2}{r}$ and we have that $p_* = -\varrho + \sqrt{|\omega|}$. For $\omega \ge 0$ (i.e. $\frac{8}{\gamma} \ge (1+r)^2$), the second identity in (113) holds. If for some value of γ, there were a transition in which identity in (113) holds, then by continuity there would need to be a value of γ so that both hold, which occurs if and only if

$$\big(1-\frac{r\gamma}{2}+\frac{r\gamma p_*}{2\sqrt{r}}m(q_*)\big)=0=\big(1-\frac{r\gamma}{2}+\frac{r\gamma p_*}{2\sqrt{r}}\frac{1}{m(q_*)}\big) \quad\Longleftrightarrow\quad m(q_*)^2=1 \text{ or } p_*=0.$$

As $m(q_*)$ is positive for $\omega \leq 0$, we must therefore have $m(q_*) = 1$ which occurs if and only if $-p_* = (1-\sqrt{r})^2$. From (112) this does not occur for $p_* = -\varrho - i\sqrt{\omega}$, and in conclusion for this branch of p_* we are always in the second case of (113).

On the other hand for $p_* = -\varrho + i\sqrt{\omega}$, when $\gamma < \gamma_*$ we must still be in the second case of (113). By continuity, it suffices to check a single value of $\gamma > \gamma_*$ to determine in which case (113) we are in. The most convenient value of $\gamma = \frac{2}{r}$, but at this point, both are 0 as $p_* = 0$. If we parameterize the first equation in terms of $t = \frac{r\gamma}{2}$, then we can write the second case of (113) as

$$1 - t + \frac{t p_*(t)}{\sqrt{r} m(q_*(t))} = 0.$$

Differentiating in t, at $t = 1$, and observing $p_*(1) = 0$, we arrive at

$$-1 + \frac{1}{\sqrt{r} m(q_*(1))} \frac{dp_*(t)}{dt}\bigg|_{t=1} = -1 - \frac{\max\{r,1\}}{\sqrt{r} m(q_*(1))} = -1 - \frac{\max\{r,1\}}{\min\{r,1\}} \neq 0,$$

except when $r = 1$. Note that when $r = 1$, no $\omega < 0$ is possible.

We summarize the outcome of this argument as follows

$$\begin{aligned}
&\big(1 - \frac{r\gamma}{2} + \frac{r\gamma p_*}{2\sqrt{r}} \frac{1}{m(q_*)}\big) = 0, && \text{for } p_* = -\varrho - i\sqrt{\omega} \text{ and } \gamma < \frac{2}{r}, \\
&\big(1 - \frac{r\gamma}{2} + \frac{r\gamma p_*}{2\sqrt{r}} \frac{1}{m(q_*)}\big) = 0, && \text{for } p_* = -\varrho + i\sqrt{\omega} \text{ and } \gamma \leq \gamma_*, \\
&\big(1 - \frac{r\gamma}{2} + \frac{r\gamma p_*}{2\sqrt{r}} m(q_*)\big) = 0, && \text{for } p_* = -\varrho + i\sqrt{\omega} \text{ and } \gamma_* < \gamma < \frac{2}{r}.
\end{aligned} \tag{114}$$

Suppose we are in the first two cases of (114), in particular, we have $1 - \frac{r\gamma}{2} + \frac{r\gamma p_*}{2\sqrt{r}} \frac{1}{m(q_*)} = 0$, then the following holds

$$m(q_*) = -\left(1 - \frac{r\gamma}{2}\right)^{-1} \frac{r\gamma p_*}{2\sqrt{r}}.$$

This then implies that

$$\int_0^\infty \frac{x}{x + p_*} \, \mathrm{d}\mu_{\mathrm{MP}}(x) = -\big(1 - \frac{r\gamma}{2}\big)^{-1} \frac{\gamma p_*}{2}. \tag{115}$$

On the other hand, when we are in the second case of (114), namely $p_* = -\varrho + i\sqrt{\omega}$ and $\gamma_* < \gamma < \frac{2}{r}$, then $m(q_*) = -\frac{2\sqrt{r}}{r\gamma p_*}\left(1 - \frac{r\gamma}{2}\right)$ and consequently,

$$\int_0^\infty \frac{x}{x + p_*} \, \mathrm{d}\mu_{\mathrm{MP}}(x) = -\left(1 - \frac{r\gamma}{2}\right) \frac{2}{r\gamma p^*}. \tag{116}$$

The result follows. □

E.2.1 Noiseless setting.

With these items in place we can start deriving an expression (104) in the noiseless setting, namely when $\widetilde{R} = 0$. In order to do so, we need to compute the Laplace transform of $k(t)$:

$$\begin{aligned}
K(p) &= \int_0^\infty e^{-pt} \left(\int_0^\infty \frac{r\gamma}{2} e^{-xt} x^2 \, \mathrm{d}\mu_{\mathrm{MP}}(x) \right) \mathrm{d}t = \frac{r\gamma}{2} \int_0^\infty \left(\int_0^\infty e^{-(x+p)t} \, \mathrm{d}t \right) x^2 \, \mathrm{d}\mu_{\mathrm{MP}}(x) \\
&= \frac{r\gamma}{2} \int_0^\infty \frac{x^2}{x + p} \, \mathrm{d}\mu_{\mathrm{MP}}(x).
\end{aligned} \tag{117}$$

Using this definition for the Marchenko-Pastur measure (8), we deduce from (117) that

$$\begin{aligned}
K(p) &= \frac{\gamma}{4\pi} \int_{\lambda^-}^{\lambda^+} \frac{x + p - p}{x + p} \sqrt{(x - \lambda^-)(\lambda^+ - x)} \, \mathrm{d}x \\
&= \frac{\gamma}{4\pi} \int_{\lambda^-}^{\lambda^+} \sqrt{(x - \lambda^-)(\lambda^+ - x)} \, dx - \frac{r \cdot \gamma \cdot p}{2} \int_0^\infty \frac{x}{x + p} \, \mathrm{d}\mu_{\mathrm{MP}}(x).
\end{aligned}$$

By applying Lemma E.2, we can connect the Stieltjes transform m to K. Recalling $q = \frac{-p-(1+r)}{\sqrt{r}}$, we deduce

$$K(p) = \frac{r\gamma}{2} - \frac{\sqrt{r}\cdot\gamma\cdot p}{2}\cdot m\left(q\right) \tag{118}$$

Consequently a simple string computations give the following identity

$$\begin{aligned}
\frac{R\left(1-\frac{2}{r\gamma}K(p)\right)}{(1-K(p))p} &= R\cdot\frac{\frac{p}{\sqrt{r}}m(q)}{p(1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}m(q))} \\
&= R\cdot\frac{\frac{1}{\sqrt{r}}m(q)}{1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}m(q)}\cdot\frac{1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}\hat{m}(q)}{1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}\hat{m}(q)} \\
&= R\cdot\frac{\frac{1}{\sqrt{r}}m(q)\left(1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}\hat{m}(q)\right)}{\left(1-\frac{r\gamma}{2}\right)^2+\frac{r\gamma p}{2\sqrt{r}}\left(1-\frac{r\gamma}{2}\right)\left(\frac{p+1+r}{\sqrt{r}}\right)+\frac{1}{r}\left(\frac{r\gamma p}{2}\right)^2} \\
&= R\cdot\frac{\frac{1}{\sqrt{r}}m(q)\left(1-\frac{r\gamma}{2}+\frac{r\gamma p}{2\sqrt{r}}\hat{m}(q)\right)}{\left(1-\frac{r\gamma}{2}\right)^2+\frac{pr\gamma(1+r)}{2r}\left(1-\frac{r\gamma}{2}\right)+\frac{r\gamma p^2}{2r}} \\
&= R\cdot\frac{\frac{1}{\sqrt{r}}m(q)\left(1-\frac{r\gamma}{2}\right)+\frac{p\gamma}{2}}{\left(1-\frac{r\gamma}{2}\right)^2+\frac{p\gamma(1+r)}{2}\left(1-\frac{r\gamma}{2}\right)+\frac{p^2\gamma}{2}}.
\end{aligned} \tag{119}$$

Lemma E.4. *Fix the stepsize $0<\gamma<\frac{2}{r}$ and set the following constants*

$$\gamma_* = \frac{2}{\sqrt{r}(r-\sqrt{r}+1)}, \quad \varrho = \frac{1+r}{2}\left(1-\frac{r\gamma}{2}\right), \quad \textit{and} \quad \omega = \frac{1}{4}\left(1-\frac{r\gamma}{2}\right)^2\left(\frac{8}{\gamma}-(1+r)^2\right).$$

If $\gamma\le\gamma_*$, then the following holds

$$\mathcal{L}^{-1}\left\{\frac{\left(1-\frac{2}{r\gamma}K(p)\right)}{(1-K(p))p}\right\} = \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x),$$

and in the case that $\gamma_*<\gamma<\frac{2}{r}$, one gets that

$$\begin{aligned}
\mathcal{L}^{-1}\left\{\frac{\left(1-\frac{2}{r\gamma}K(p)\right)}{(1-K(p))p}\right\} &= \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x) \\
&\quad + \frac{2i\sqrt{\omega}}{4\omega}\cdot\left[\varrho-i\sqrt{\omega}-\left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho+i\sqrt{\omega}}{r(\varrho^2+\omega)}\right]e^{(-\varrho+i\sqrt{\omega})t}.
\end{aligned}$$

Proof. We first suppose that $\gamma\le\gamma_*$ and define the function

$$y(t) \overset{\text{def}}{=} \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x).$$

Then the Laplace transform of this function is given by the equation

$$Y(p) \overset{\text{def}}{=} \mathcal{L}\{y(t)\}(p) = \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{x}{(x+p)((x-\varrho)^2+\omega)}\,\mathrm{d}\mu_{\mathrm{MP}}(x).$$

The roots of $(x-\varrho)^2+\omega$ are precisely $\varrho\pm i\sqrt{\omega}$ so by partial fractions, we have that

$$\begin{aligned}
\frac{x}{(x+p)((x-\varrho)^2+\omega)} &= \frac{1}{(p+\varrho)^2+\omega}\cdot\frac{x}{x+p} - \frac{2i\sqrt{\omega}}{4\omega(p+\varrho+i\sqrt{\omega})}\cdot\frac{x}{x-\varrho-i\sqrt{\omega}} \\
&\quad + \frac{2i\sqrt{\omega}}{4\omega(p+\varrho-i\sqrt{\omega})}\cdot\frac{x}{x-\varrho+i\sqrt{\omega}}
\end{aligned} \tag{120}$$

provided that $\sqrt{\omega} \neq 0$. Using Lemmas E.2 and E.3, we have that

$$\begin{aligned} \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{x}{x-\varrho-i\sqrt{\omega}}\,\mathrm{d}\mu_{\mathrm{MP}}(x) &= \varrho+i\sqrt{\omega}, \\ \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{x}{x-\varrho+i\sqrt{\omega}}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \varrho-i\sqrt{\omega}, \quad &\text{and} \quad \int_0^\infty \frac{x}{x+p}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{m(q)}{\sqrt{r}}, \end{aligned} \tag{121}$$

where the function $m(q)$ and the point $q = -\frac{p+1+r}{\sqrt{r}}$ are defined in Lemma E.2. A simple calculation using (121) shows that

$$\begin{aligned} &\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \left(\frac{2i\sqrt{\omega}}{4\omega(p+\varrho-i\sqrt{\omega})}\cdot\frac{x}{x-\varrho+i\sqrt{\omega}} - \frac{2i\sqrt{\omega}}{4\omega(p+\varrho+i\sqrt{\omega})}\cdot\frac{x}{x-\varrho-i\sqrt{\omega}}\right)\mathrm{d}\mu_{\mathrm{MP}}(x) \\ &\quad= \frac{2i\sqrt{\omega}}{4\omega(p+\varrho-i\sqrt{\omega})}\cdot(\varrho-i\sqrt{\omega}) - \frac{2i\sqrt{\omega}}{4\omega(p+\varrho+i\sqrt{\omega})}\cdot(\varrho+i\sqrt{\omega}) \\ &\quad= \frac{p}{(p+\varrho)^2+\omega}. \end{aligned} \tag{122}$$

Combining the partial fractions decomposition of $Y(s)$ in (120) and (121), we deduce that

$$Y(s) = \frac{\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\frac{m(q)}{\sqrt{r}}+p}{(p+\varrho)^2+\omega}.$$

Since both $Y(s)$ and the RHS make sense when $\omega = 0$ by continuity this result holds when $\omega = 0$. After noting that

$$\frac{\gamma}{2}\left[(p+\varrho)^2+\omega\right] = \left(1-\frac{r\gamma}{2}\right)^2 + \frac{p\gamma(1+r)}{2}\left(1-\frac{r\gamma}{2}\right) + \frac{\gamma p^2}{2},$$

the result follows for $\gamma \leq \gamma_*$ from comparing with (119).

Next, we consider the setting where $\gamma > \gamma_*$. In this case, we have that $\omega < 0$. Let A_1 and A_2 be indeterminates and define

$$\begin{aligned} w(t) &\overset{\text{def}}{=} \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + A_1\frac{2i\sqrt{\omega}}{4\omega}e^{(-\varrho+i\sqrt{\omega})t} + A_2\frac{2i\sqrt{\omega}}{4\omega}e^{(-\varrho-i\sqrt{\omega})t} \\ &= y(t) + A_1\frac{2i\sqrt{\omega}}{4\omega}e^{(-\varrho+i\sqrt{\omega})t} + A_2\frac{2i\sqrt{\omega}}{4\omega}e^{(-\varrho-i\sqrt{\omega})t}. \end{aligned}$$

Particularly the Laplace transform of $w(t)$ is

$$W(p) \overset{\text{def}}{=} \mathcal{L}\{w(t)\} = Y(p) + \frac{2i\sqrt{\omega}}{4\omega}\frac{A_1}{p+\varrho-i\sqrt{\omega}} + \frac{2i\sqrt{\omega}}{4\omega}\frac{A_2}{p+\varrho+i\sqrt{\omega}}. \tag{123}$$

As in the previous case, we have the partial fraction decomposition in (120) for $Y(p) = \mathcal{L}\{y(t)\}$. However in this case, using Lemmas E.2 and E.3, we get different formulas for (121):

$$\begin{aligned} \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{x}{x-\varrho+i\sqrt{\omega}}\,\mathrm{d}\mu_{\mathrm{MP}}(x) &= \left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho+i\sqrt{\omega}}{r(\varrho^2+\omega)}, \\ \frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{x}{x-\varrho-i\sqrt{\omega}}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \varrho+i\sqrt{\omega}, \ &\text{and} \ \int_0^\infty \frac{x}{x+p}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{m(q)}{\sqrt{r}}. \end{aligned} \tag{124}$$

With these equations, we have that

$$\begin{aligned} &\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \left(\frac{2i\sqrt{\omega}}{4\omega(p+\varrho-i\sqrt{\omega})}\cdot\frac{x}{x-\varrho+i\sqrt{\omega}}\right. \\ &\qquad\left. - \frac{2i\sqrt{\omega}}{4\omega(p+\varrho+i\sqrt{\omega})}\cdot\frac{x}{x-\varrho-i\sqrt{\omega}}\right)\mathrm{d}\mu_{\mathrm{MP}}(x) \\ &\quad= \frac{2i\sqrt{\omega}}{4\omega(p+\varrho-i\sqrt{\omega})}\left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho+i\sqrt{\omega}}{r(\varrho^2+\omega)} \\ &\qquad - \frac{2i\sqrt{\omega}}{4\omega(p+\varrho+i\sqrt{\omega})}\cdot(\varrho+i\sqrt{\omega}). \end{aligned} \tag{125}$$

We observe that if $\left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho+i\sqrt{\omega}}{r(\varrho^2+\omega)}=\varrho-i\sqrt{\omega}$, then by (122) we would be done. Therefore we can use A_1 to make this term equal to $\varrho-i\sqrt{\omega}$. Hence, we should choose A_1 such that

$$A_1+\left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho+i\sqrt{\omega}}{r(\varrho^2+\omega)}=\varrho-i\sqrt{\omega}$$
$$\Rightarrow\quad A_1=\varrho-i\sqrt{\omega}-\left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho+i\sqrt{\omega}}{r(\varrho^2+\omega)}\quad\text{and}\quad A_2=0.$$

If we define

$$Z(p)\overset{\text{def}}{=}\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty\left(\frac{2i\sqrt{\omega}}{4\omega(p+\varrho-i\sqrt{\omega})}\cdot\frac{x}{x-\varrho+i\sqrt{\omega}}\right.$$
$$\left.-\frac{2i\sqrt{\omega}}{4\omega(p+\varrho+i\sqrt{\omega})}\cdot\frac{x}{x-\varrho-i\sqrt{\omega}}\right)\mathrm{d}\mu_{\mathrm{MP}}(x),$$

then by the construction of A_1 and A_2, we deduce that

$$Z(p)+\frac{2i\sqrt{\omega}}{4\omega}\frac{A_1}{p+\varrho-i\sqrt{\omega}}+\frac{2i\sqrt{\omega}}{4\omega}\frac{A_1}{p+\varrho+i\sqrt{\omega}}=\frac{p}{(p+\varrho)^2+\omega}.$$

from (122). Putting together this with (124) and the definition of $W(p)$ in (123), we get that

$$W(p)=\frac{\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)\frac{m(q)}{\sqrt{r}}+p}{(p+\varrho)^2+\omega}.$$

The result then follows. □

Theorem E.4 (Dynamics of SGD, noiseless setting)**.** *Suppose $\widetilde{R}=0$ and the batchsize satisfies $\beta(n)\le n^{1/5-\delta}$ for some $\delta>0$, and the stepsize is $0<\gamma<\frac{2}{r}$. Define the critical stepsize $\gamma_*\in\mathbb{R}$ and constants $\varrho>0$ and $\omega\in\mathbb{C}$,*

$$\gamma_*=\frac{2}{\sqrt{r}(r-\sqrt{r}+1)},\ \varrho=\frac{1+r}{2}\left(1-\frac{r\gamma}{2}\right),\ \text{and}\ \omega=\frac{1}{4}\left(1-\frac{r\gamma}{2}\right)^2\left(\frac{8}{\gamma}-(1+r)^2\right).$$

The iterates of SGD satsify if $\gamma\le\gamma_*$

$$f(\boldsymbol{x}_{\lfloor\frac{n}{\beta}t\rfloor})\xrightarrow[n\to\infty]{\Pr}\quad R\cdot\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty\frac{xe^{-2\gamma xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x)$$

and if $\gamma>\gamma_*$, the iterates of SGD satisfy

$$f(\boldsymbol{x}_{\lfloor\frac{n}{\beta}t\rfloor})\xrightarrow[n\to\infty]{\Pr}\quad R\cdot\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty\frac{xe^{-2\gamma xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x)$$
$$+R\cdot\frac{1}{4\sqrt{|\omega|}}\cdot\left[\varrho+\sqrt{|\omega|}-\left(\frac{2}{\gamma}\right)^2\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho-\sqrt{|\omega|}}{r(\varrho^2-|\omega|)}\right]e^{-2\gamma(\varrho+\sqrt{|\omega|})t}.$$

Here the convergence is locally uniformly.

Proof. The result follows immediately from Lemma E.4 after noting that when $\gamma_*<\gamma$ we have $\omega<0$ and that $f(\boldsymbol{x}_{\lfloor\frac{n}{\beta}t\rfloor})\xrightarrow[n\to\infty]{\Pr}\psi_0(t)=\frac{1}{2}\widehat{\psi}_0(2\gamma t)$. □

E.2.2 Noisy term

We now turn to solving the noisy term in (104) (*i.e.* the term with $\widetilde{R}$), namely

$$\widetilde{R}\cdot\frac{r\mathcal{L}\{T(t)\}(p)+\frac{1-r}{p}}{1-K(p)}.$$

First we rewrite the Laplace transform of T in terms of the point $q = \frac{-p-1-r}{\sqrt{r}}$. Recall the function $m(z)$ in (105) as in Lemma E.2 by $m(z) = \frac{-z+\sqrt{z^2-4}}{2}$ so that the Laplace transform of T becomes

$$r\mathcal{L}\{T(t)\}(p) = \frac{\sqrt{r}}{p} \cdot \frac{q - \sqrt{q^2-4}}{2} + \frac{r}{p} = \frac{r}{p} - \frac{\sqrt{r}}{p} m(q).$$

We note again from Lemma E.2 that $m(q)\hat{m}(q) = 1$ and $\hat{m}(q) = -q - m(q)$. Using the definition of $K(p)$ from (118), we get the following equality for the noisy term

$$\begin{aligned}
\widetilde{R} \cdot \frac{r\mathcal{L}\{T(t)\}(p) + \frac{1-r}{p}}{1-K(p)} &= \widetilde{R} \cdot \frac{-\sqrt{r} \cdot m(q) + 1}{p(1-K(p))} \cdot \frac{1 - \frac{r\gamma}{2} + \frac{r\gamma p}{2\sqrt{r}}\hat{m}(q)}{1 - \frac{r\gamma}{2} + \frac{r\gamma p}{2\sqrt{r}}\hat{m}(q)} \\
&= \widetilde{R} \cdot \frac{-\sqrt{r}\big(1-\frac{r\gamma}{2}\big)m(q) - \frac{r\gamma p}{2} + 1 - \frac{r\gamma}{2} + \frac{r\gamma p}{2\sqrt{r}}\hat{m}(q)}{p\big[\big(1-\frac{r\gamma}{2}\big)^2 + \frac{rp\gamma(1+r)}{2r}\big(1-\frac{r\gamma}{2}\big) + \frac{r\gamma p^2}{2r}\big]} \\
&= \widetilde{R} \cdot \frac{-\big[\frac{2}{\gamma}\sqrt{r}\big(1-\frac{r\gamma}{2}\big) + \sqrt{r}p\big]m(q) + p^2 + p + \frac{2}{\gamma}\big(1-\frac{r\gamma}{2}\big)}{p((p+\varrho)^2+\omega)},
\end{aligned} \tag{126}$$

where ϱ and ω are defined in (109). With this, we able to conclude derive an expression for the noisy term in $\widehat{\psi}_0(t)$ of (100).

Lemma E.5. *Fix the stepsize $0 < \gamma < \frac{2}{r}$ and set the following constants*

$$\gamma_* = \frac{2}{\sqrt{r}(r-\sqrt{r}+1)}, \;\; \varrho = \frac{1+r}{2}\left(1 - \frac{r\gamma}{2}\right), \;\; \textit{and} \quad \omega = \frac{1}{4}\left(1-\frac{r\gamma}{2}\right)^2\left(\frac{8}{\gamma} - (1+r)^2\right).$$

If $\gamma \leq \gamma_*$, then the following holds

$$\mathcal{L}^{-1}\left\{\frac{r\mathcal{L}\{T(t)\} + \frac{1-r}{p}}{1-K(p)}\right\} = \frac{\frac{2}{\gamma}(1-r)\big(1-\frac{r\gamma}{2}\big)}{\varrho^2+\omega} + \int_0^\infty \frac{-rx + \frac{2}{\gamma}r\big(1-\frac{r\gamma}{2}\big)}{(x-\varrho)^2+\omega} e^{-xt}\, \mathrm{d}\mu_{\mathrm{MP}}(x)$$

and in the case that $\gamma_* < \gamma < \frac{2}{r}$, one gets that

$$\begin{aligned}
\mathcal{L}^{-1}\left\{\frac{r\mathcal{L}\{T(t)\} + \frac{1-r}{p}}{1-K(p)}\right\} &= \frac{\frac{2}{\gamma}(1-r)\big(1-\frac{r\gamma}{2}\big)}{\varrho^2+\omega} + \int_0^\infty \frac{-rx + \frac{2}{\gamma}r\big(1-\frac{r\gamma}{2}\big)}{(x-\varrho)^2+\omega} e^{-xt}\, \mathrm{d}\mu_{\mathrm{MP}}(x) \\
&+ \frac{(2i\sqrt{\omega})r\big[\frac{2}{\gamma}\big(1-\frac{r\gamma}{2}\big) - (\varrho - i\sqrt{\omega})\big]}{4\omega(\varrho - i\sqrt{\omega})} \cdot \left[\frac{\varrho - i\sqrt{\omega}}{\frac{2}{\gamma}\big(1-\frac{r\gamma}{2}\big)} - \frac{\frac{2}{\gamma}(\varrho + i\sqrt{\omega})\big(1-\frac{r\gamma}{2}\big)}{r(\varrho^2+\omega)}\right] e^{(-\varrho + i\sqrt{\omega})t}.
\end{aligned}$$

Proof. We first consider the setting where $\gamma \leq \gamma_*$ and we define the functions

$$y(t) \overset{\text{def}}{=} \int_0^\infty \frac{-rx + \frac{2r}{\gamma}\big(1-\frac{r\gamma}{2}\big)}{(x-\varrho)^2+\omega} e^{-xt}\, \mathrm{d}\mu_{\mathrm{MP}}(x) \quad \text{and} \quad j(t) \overset{\text{def}}{=} \frac{\frac{2}{\gamma}(1-r)\big(1-\frac{r\gamma}{2}\big)}{\varrho^2+\omega}.$$

Then the Laplace transform of this function is given by the equation

$$Y(p) \overset{\text{def}}{=} \mathcal{L}\{y(t)\} = \int_0^\infty \frac{-rx + \frac{2r}{\gamma}\big(1-\frac{r\gamma}{2}\big)}{x(x+p)[(x-\varrho)^2+\omega]}\, x\, \mathrm{d}\mu_{\mathrm{MP}}(x). \tag{127}$$

The roots of $(x-\varrho)^2 + \omega$ are precisely $\varrho \pm i\sqrt{\omega}$ so by partial fractions, we have that

$$\begin{aligned}
\frac{-rx + \frac{2r}{\gamma}\big(1-\frac{r\gamma}{2}\big)}{x(x+p)[(x-\varrho)^2+\omega]} \cdot x &= \frac{\frac{2r}{\gamma}\big(1-\frac{r\gamma}{2}\big)}{p(\varrho^2+\omega)} - \frac{rp + \frac{2r}{\gamma}\big(1-\frac{r\gamma}{2}\big)}{\big[(p+\varrho)^2+\omega\big]p} \cdot \frac{x}{x+p} \\
&+ \frac{\frac{2r}{\gamma}\big(1-\frac{r\gamma}{2}\big) - r(\varrho + i\sqrt{\omega})}{(p+\varrho+i\sqrt{\omega})(\varrho+i\sqrt{\omega})(2i\sqrt{\omega})} \cdot \frac{x}{x-\varrho-i\sqrt{\omega}} \\
&+ \frac{\frac{2r}{\gamma}\big(1-\frac{r\gamma}{2}\big) - r(\varrho - i\sqrt{\omega})}{(p+\varrho-i\sqrt{\omega})(\varrho-i\sqrt{\omega})(-2i\sqrt{\omega})} \cdot \frac{x}{x-\varrho+i\sqrt{\omega}}.
\end{aligned} \tag{128}$$

We will consider each term individual. By using Lemma E.2, we deduce that

$$\int_0^\infty \frac{rp + \frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right)}{\left[(p+\varrho)^2 + \omega\right]p} \cdot \frac{x}{x+p} \, \mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{m(q)}{\sqrt{r}} \cdot \frac{rp + \frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right)}{p((p+\varrho)^2 + \omega)}. \tag{129}$$

This matches the term in front of the $m(q)$ in (126). For the last two terms, we apply Lemma E.3 and some simple computations to conclude that

$$\begin{aligned} L(p) \stackrel{\text{def}}{=} \int_0^\infty & \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho + i\sqrt{\omega})}{(p + \varrho + i\sqrt{\omega})(\varrho + i\sqrt{\omega})(2i\sqrt{\omega})} \cdot \frac{x}{x - \varrho - i\sqrt{\omega}} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \\ & + \int_0^\infty \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho - i\sqrt{\omega})}{(p + \varrho - i\sqrt{\omega})(\varrho - i\sqrt{\omega})(-2i\sqrt{\omega})} \cdot \frac{x}{x - \varrho + i\sqrt{\omega}} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \\ & = \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho + i\sqrt{\omega})}{(p + \varrho + i\sqrt{\omega})(2i\sqrt{\omega})\frac{2}{\gamma}\left(1 - \frac{r\gamma}{2}\right)} + \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho - i\sqrt{\omega})}{(p + \varrho - i\sqrt{\omega})(-2i\sqrt{\omega})\frac{2}{\gamma}\left(1 - \frac{r\gamma}{2}\right)} \\ & = \frac{-r\gamma(\frac{2}{\gamma} + p - r)}{2((p+\varrho)^2 + \omega)\left(1 - \frac{r\gamma}{2}\right)}. \end{aligned} \tag{130}$$

We next observe that $\mathcal{L}\{1\} = \frac{1}{p}$. This observation applied to the function $j(t)$ together with the terms not associated with $m(q)$ in (128) gives the following result

$$\frac{2(1-r)\left(1 - \frac{r\gamma}{2}\right)}{p\gamma(\varrho^2 + \omega)} + L(p) + \frac{2r\left(1 - \frac{r\gamma}{2}\right)}{p\gamma(\varrho^2 + \omega)} = \frac{p^2 + p + \frac{2}{\gamma}\left(1 - \frac{r\gamma}{2}\right)}{p((p+\varrho)^2 + \omega)}.$$

This combined with (129) and (126) shows that result holds when $\gamma \leq \gamma_*$.

Next, we consider the setting where $\gamma > \gamma_*$. In this case, we always have that $\omega < 0$. Let A_1 be an indeterminate and define

$$w(t) \stackrel{\text{def}}{=} y(t) + j(t) + \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho - i\sqrt{\omega})}{(\varrho - i\sqrt{\omega})(-2i\sqrt{\omega})} A_1 e^{(-\varrho + i\sqrt{\omega})t}.$$

The Laplace transform of $w(t)$ is

$$W(p) \stackrel{\text{def}}{=} \mathcal{L}\{w(t)\} = Y(p) + \frac{2(1-r)\left(1 - \frac{r\gamma}{2}\right)}{\gamma(\varrho^2 + \omega)} \cdot \frac{1}{p} + \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho - i\sqrt{\omega})}{(\varrho - i\sqrt{\omega})(-2i\sqrt{\omega})} A_1 \cdot \frac{1}{p + \varrho - i\sqrt{\omega}}.$$

As in the previous case, we have that (127), (128), and (129) all still hold. The only difference occurs in the second term in the function $L(p)$ in (130). In this case using Lemma E.3, we get that

$$\begin{aligned} \widehat{L}(p) \stackrel{\text{def}}{=} \int_0^\infty & \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho + i\sqrt{\omega})}{(p + \varrho + i\sqrt{\omega})(\varrho + i\sqrt{\omega})(2i\sqrt{\omega})} \cdot \frac{x}{x - \varrho - i\sqrt{\omega}} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \\ & + \int_0^\infty \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho - i\sqrt{\omega})}{(p + \varrho - i\sqrt{\omega})(\varrho - i\sqrt{\omega})(-2i\sqrt{\omega})} \cdot \frac{x}{x - \varrho + i\sqrt{\omega}} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \\ & = \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho + i\sqrt{\omega})}{(p + \varrho + i\sqrt{\omega})(2i\sqrt{\omega})\frac{2}{\gamma}\left(1 - \frac{r\gamma}{2}\right)} \\ & \quad + \frac{\frac{2r}{\gamma}\left(1 - \frac{r\gamma}{2}\right) - r(\varrho - i\sqrt{\omega})}{(p + \varrho - i\sqrt{\omega})(\varrho - i\sqrt{\omega})(-2i\sqrt{\omega})} \frac{2(\varrho + i\sqrt{\omega})\left(1 - \frac{r\gamma}{2}\right)}{r\gamma(\varrho^2 + \omega)}. \end{aligned} \tag{131}$$

Using the term A_1, we are able to make $\widehat{L}(p)$ exactly equal to the RHS of $L(p)$. In particular, we choose the constant A_1 as

$$A_1 = \frac{\varrho - i\sqrt{\omega}}{\frac{2}{\gamma}\left(1 - \frac{r\gamma}{2}\right)} - \frac{2(\varrho + i\sqrt{\omega})\left(1 - \frac{r\gamma}{2}\right)}{r\gamma(\varrho^2 + \omega)}. \tag{132}$$

By this choice of A_1, we guarantee that the sum of $\widehat{L}(p)$ and the A_1 term equals the RHS of $L(p)$:

$$\widehat{L}(p) + \frac{\frac{2r}{\gamma}\left(1-\frac{r\gamma}{2}\right) - r(\varrho - i\sqrt{\omega})}{(\varrho - i\sqrt{\omega})(-2i\sqrt{\omega})} A_1 \cdot \frac{1}{p+\varrho-i\sqrt{\omega}} = \frac{-r\gamma(\frac{2}{\gamma}+p-r)}{2((p+\varrho)^2+\omega)\left(1-\frac{r\gamma}{2}\right)}.$$

The result then immediately follows from the previous case when $\gamma > \gamma_*$. □

From this lemma, we can now derive the main result which shows that the function values concentrate.

Theorem E.5 (Dynamics of SGD, noisy setting)**.** *Suppose the batchsize satisfies $\beta(n) \leq n^{1/5-\delta}$ for some $\delta > 0$ and the stepsize is $0 < \gamma < \frac{2}{r}$. Define the critical stepsize γ_* and constants $\varrho > 0$ and $\omega \in \mathbb{C}$ by*

$$\gamma_* = \frac{2}{\sqrt{r}(r-\sqrt{r}+1)},\ \varrho = \frac{1+r}{2}\left(1-\frac{r\gamma}{2}\right),\ \text{and } \omega = \frac{1}{4}\left(1-\frac{r\gamma}{2}\right)^2\left(\frac{8}{\gamma}-(1+r)^2\right).$$

The iterates of SGD satisfy if $\gamma \leq \gamma_*$

$$f(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t \rfloor}) \xrightarrow[d\to\infty]{\Pr} R\cdot\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-2\gamma xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x) + \widetilde{R}\cdot\frac{1}{2}\cdot\left[\frac{\frac{2}{\gamma}(1-r)\left(1-\frac{r\gamma}{2}\right)}{\varrho^2+\omega} + \int_0^\infty \frac{-rx+\frac{2r}{\gamma}\left(1-\frac{r\gamma}{2}\right)}{(x-\varrho)^2+\omega}e^{-2\gamma xt}\,\mathrm{d}\mu_{\mathrm{MP}}(x)\right].$$

and if $\gamma > \gamma_*$, the iterates of SGD satisfy

$$\begin{aligned} f(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t \rfloor}) &\xrightarrow[d\to\infty]{\Pr} R\cdot\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right)\int_0^\infty \frac{xe^{-2\gamma xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x) \\ &+ R\cdot\frac{1}{4\sqrt{|\omega|}}\cdot\left[\varrho+\sqrt{|\omega|}-\frac{4}{\gamma^2}\left(1-\frac{r\gamma}{2}\right)^2\frac{\varrho-\sqrt{|\omega|}}{r(\varrho^2-|\omega|)}\right]e^{-2\gamma(\varrho+\sqrt{|\omega|})t} \\ &+ \widetilde{R}\cdot\frac{1}{2}\left[\frac{\frac{2}{\gamma}(1-r)\left(1-\frac{r\gamma}{2}\right)}{\varrho^2+\omega} + \int_0^\infty \frac{-rx+\frac{2r}{\gamma}\left(1-\frac{r\gamma}{2}\right)}{(x-\varrho)^2+\omega}e^{-2\gamma xt}\,\mathrm{d}\mu_{\mathrm{MP}}(x)\right. \\ &\left. + \frac{(-2\sqrt{|\omega|})r\left[\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)-(\varrho+\sqrt{|\omega|})\right]}{4\omega(\varrho+\sqrt{|\omega|})}\cdot\left(\frac{\varrho+\sqrt{|\omega|}}{\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)} - \frac{\frac{2}{\gamma}(\varrho-\sqrt{|\omega|})(1-\frac{r\gamma}{2})}{r(\varrho^2+\omega)}\right)e^{-2\gamma(\varrho+\sqrt{|\omega|})t}\right]. \end{aligned}$$

Here the convergence is locally uniformly.

Proof. The result follows immediately from Lemma E.5 after noting that when $\gamma_* < \gamma$ we have $\omega < 0$ and that $f(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t \rfloor}) \xrightarrow[d\to\infty]{\Pr} \psi_0(t) = \frac{1}{2}\widehat{\psi}_0(2\gamma t)$. □

E.2.3 Computing average-complexity

With our explicit expressions for ψ_0, we can now derive the complexity results.

Theorem E.6 (Asymptotic convergence rates isotropic features)**.** *Suppose Assumptions 1.1 and 1.2 hold with $d\mu(x) = d\mu_{\mathrm{MP}}(x)$. Fix the stepsize $0 < \gamma < \frac{2}{r}$ and let the batchsize satisfies $\beta(n) \leq n^{1/5-\delta}$ for some $\delta > 0$. Define the constants $\varrho > 0, \omega \in \mathbb{C}$, and critical stepsize $\gamma_* \in \mathbb{R}$ as in (E.5). If the ratio $r = 1$, the iterates of SGD satisfy*

$$\lim_{n\to\infty} f(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t \rfloor}) \overset{\Pr}{\sim} \frac{1}{\gamma^{3/2}}\cdot\left(1-\frac{r\gamma}{2}\right)\cdot\frac{\sqrt{\lambda^+}}{2\sqrt{2\pi}}\cdot\frac{1}{\varrho^2+\omega}\left[R\cdot\frac{1}{4\gamma}\cdot\frac{1}{t^{3/2}} + \widetilde{R}\cdot r\cdot\frac{1}{t^{1/2}}\right]. \tag{133}$$

If the ratio $r \neq 1$ and $\gamma < \gamma_*$, the iterates of SGD satisfy

$$\begin{aligned} &\lim_{n\to\infty} f(\boldsymbol{x}_{\lfloor \frac{n}{\beta}t \rfloor}) \overset{Pr}{\sim} \widetilde{R}\cdot\frac{\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right)}{(\varrho^2+\omega)}\max\{0,1-r\} \\ &+ \frac{1}{8\sqrt{2\pi}r}\cdot\frac{(\lambda^+-\lambda^-)^{1/2}}{\gamma^{3/2}\left[(\lambda^--\varrho)^2+\omega\right]}\left[R\cdot\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right) + \frac{\widetilde{R}}{2}\left(\frac{2r}{\gamma\lambda^-}\left(1-\frac{r\gamma}{2}\right)-r\right)\right]\cdot e^{-2\gamma\lambda^- t}\cdot\frac{1}{t^{3/2}}. \end{aligned}$$

If the ratio $r \neq 1$ and $\gamma = \gamma_*$, the iterates of SGD satisfy

$$\lim_{n\to\infty} f(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}) \overset{Pr}{\sim} \widetilde{R} \cdot \frac{\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right)}{(\varrho^2+\omega)} \max\{0, 1-r\}$$
$$+ \frac{1}{2\sqrt{2\pi r}} \cdot \frac{1}{\gamma^{1/2} r_2 (\lambda^+ - \lambda^-)^{1/2}} \left[R \cdot \frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right) + \widetilde{R} \cdot \left(\frac{2r}{\gamma\lambda^-}\left(1-\frac{r\gamma}{2}\right) - r \right) \right] \cdot e^{-2\gamma\lambda^- t} \cdot \frac{1}{t^{1/2}}.$$

If the ratio $r \neq 1$ and $\gamma > \gamma_*$, the iterates of SGD satisfy

$$\lim_{n\to\infty} f(\boldsymbol{x}_{\lfloor \frac{n}{\beta} t \rfloor}) \overset{Pr}{\sim} \widetilde{R} \cdot \frac{\frac{1}{\gamma}\left(1-\frac{r\gamma}{2}\right)}{(\varrho^2+\omega)} \max\{0, 1-r\}$$
$$+ R \cdot \frac{1}{4\sqrt{|\omega|}} \cdot \left[\varrho + \sqrt{|\omega|} - \frac{4}{\gamma^2}\left(1-\frac{r\gamma}{2}\right)^2 \frac{\varrho - \sqrt{|\omega|}}{r(\varrho^2 - |\omega|)} \right] e^{-2\gamma(\varrho+\sqrt{|\omega|})t}$$
$$+ \widetilde{R} \cdot \frac{(-2\sqrt{|\omega|}) r \left[\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right) - (\varrho + \sqrt{|\omega|}) \right]}{8\omega(\varrho+\sqrt{|\omega|})} \cdot \left(\frac{\varrho+\sqrt{|\omega|}}{\frac{2}{\gamma}\left(1-\frac{r\gamma}{2}\right)} - \frac{\frac{2}{\gamma}(\varrho - \sqrt{|\omega|})(1-\frac{r\gamma}{2})}{r(\varrho^2+\omega)} \right) e^{-2\gamma(\varrho+\sqrt{|\omega|})t}.$$

Here $t = 1$ corresponds to computing n stochastic gradients, the convergence is locally uniformly, and the notation $\overset{Pr}{\sim}$ means that you take the limit in probability and then compute the asymptotic.

Proof. Throughout the proof we use $2\gamma t \mapsto t$ and we define ϱ and ω as in Theorem E.5. Suppose we have that $r = 1$. Because $1 - \frac{\gamma}{2} > 0$, we know that $\gamma < \gamma_*$. By construction of γ_* in (112), we have that

$$(1+r)^2 - 8\gamma^{-1} < (1+r)^2 - 8\gamma_* = (1-\sqrt{r})^4 = 0.$$

In particular, this means that $\omega > 0$ and the roots of $(x-\varrho)^2 + \omega$ are precisely $x = \varrho \pm \sqrt{-\omega}$. As $\omega > 0$, these roots are complex with positive imaginary part and thus, $(x-\varrho)^2 + \omega \neq 0$ for any $x \in [0, \lambda^+]$. So there exists a constant $C > 0$ such that $(x-\varrho)^2 + \omega > C$ for all $x \in [0, \lambda^+]$ and consequently,

$$\int_0^\infty \frac{x}{(x-\varrho)^2+\omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \quad \text{is bounded.} \tag{134}$$

We begin by computing $\int_0^\infty \frac{xe^{-xt}}{(x-\varrho)^2+\omega} \, \mathrm{d}\mu_{\mathrm{MP}}$. If $x \geq \log^2(t)/t$, then it is clear that e^{-tx} decays faster than any polynomial in t. Combining this with (134), we get that

$$\int_{\log^2(t)/t}^\infty \frac{xe^{-xt}}{(x-\varrho)^2+\omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \quad \text{decays faster than any polynomial in } t.$$

Now we suppose that $0 \leq x < \log^2(t)/t$. Using a simple change of variables, we deduce that

$$\int_0^{\log^2(t)/t} \frac{xe^{-tx}}{(x-\varrho)^2+\omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{1}{2\pi} \cdot \frac{1}{t^{3/2}} \int_0^{\min\{\log^2(t), t\lambda^+\}} \frac{\sqrt{u} \cdot e^{-u}}{(\frac{1}{t}u - \varrho)^2 + \omega} \sqrt{\lambda^+ - \tfrac{u}{t}} \, \mathrm{d}u.$$

We have that $\sqrt{\lambda^+ - \frac{u}{t}} \leq \sqrt{\lambda^+}$ and $0 < \omega \leq (x-\varrho)^2 + \omega$ so dominated convergence theorem holds

$$\lim_{t\to\infty} \int_0^{\min\{\log^2(t), t\lambda^+\}} \frac{\sqrt{u} \cdot e^{-u}}{(\frac{1}{t}u - \varrho)^2 + \omega} \sqrt{\lambda^+ - \tfrac{u}{t}} \, \mathrm{d}u = \int_0^\infty \frac{\sqrt{u} \cdot e^{-u}}{\varrho^2+\omega} \sqrt{\lambda^+} \, \mathrm{d}u = \frac{\sqrt{\lambda^+ \pi}}{2(\varrho^2+\omega)}.$$

Consequently, we have that

$$\int_0^\infty \frac{xe^{-tx}}{(x-\varrho)^2+\omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \sim \frac{1}{4\sqrt{\pi}} \cdot \frac{\sqrt{\lambda^+}}{\varrho^2+\omega} \cdot \frac{1}{t^{3/2}}. \tag{135}$$

Now we turn to $\int_0^\infty \frac{e^{-xt}}{(x-\varrho)^2+\omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x)$. As before, we know that

$$\int_{\log^2(t)/t}^\infty \frac{e^{-xt}}{(x-\varrho)^2+\omega} \, \mathrm{d}\mu_{\mathrm{MP}}(x) \quad \text{decays faster than any polynomial in } t.$$

Again using a simple change of variables, we deduce that

$$\int_0^{\log^2(t)/t} \frac{e^{-tx}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x) = \frac{1}{2\pi}\cdot\frac{1}{t^{1/2}}\int_0^{\min\{\log^2(t),t\lambda^+\}} \frac{e^{-u}\sqrt{\lambda^+-\frac{u}{t}}}{\sqrt{u}\big((\frac{u}{t}-\varrho)^2+\omega\big)}\,\mathrm{d}u.$$

By using dominated convergence theorem, we know that

$$\lim_{t\to\infty}\int_0^{\min\{\log^2(t),\lambda^+t\}} \frac{e^{-u}\sqrt{\lambda^+-\frac{u}{t}}}{\sqrt{u}\big((\frac{u}{t}-\varrho)^2+\omega\big)}\,\mathrm{d}u = \int_0^{\infty}\frac{e^{-u}\sqrt{\lambda^+}}{\sqrt{u}(\varrho^2+\omega)}\,\mathrm{d}u = \frac{\sqrt{\pi\lambda^+}}{\varrho^2+\omega}.$$

Consequently, we have that

$$\int_0^{\infty}\frac{e^{-xt}}{(x-\varrho)^2+\omega}\,\mathrm{d}\mu_{\mathrm{MP}}(x) \sim \frac{1}{2\sqrt{\pi}}\cdot\frac{\sqrt{\lambda^+}}{\varrho^2+\omega}\cdot\frac{1}{t^{1/2}}. \tag{136}$$

After noting that $t = 2\gamma t$ and Theorem E.5, the result immediately follows for the case when $r = 1$ in (133).

Next we consider the setting where $r \neq 1$. Suppose $\ell \in \{0,1\}$. By a simple change of variables $x = \lambda^- + (\lambda^+ - \lambda^-)u$, we have

$$\begin{aligned}&\frac{1}{2\pi r}\int_{\lambda^-}^{\lambda^+}\frac{e^{-xt}}{x^{1-\ell}((x-\varrho)^2+\omega)}\sqrt{(x-\lambda^-)(\lambda^+-x)}\,\mathrm{d}x \\ &= \frac{1}{2\pi r}\left(\int_0^{\log^2(t)/t}+\int_{\log^2(t)/t}^1\right)\frac{(\lambda^+-\lambda^-)^2e^{-t\lambda^-}e^{-(\lambda^+-\lambda^-)ut}\sqrt{u(1-u)}}{(\lambda^-+(\lambda^+-\lambda^-)u)^{1-\ell}\big[(\lambda^-+(\lambda^+-\lambda^-)u-\varrho)^2+\omega\big]}\,\mathrm{d}u.\end{aligned} \tag{137}$$

Let's first consider where $u \geq \log^2(t)/t$. As $r \neq 1$, we have that $\lambda^- > 0$ and therefore, $(\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell} \geq (\lambda^-)^{1-\ell}$. If $\omega > 0$, then $0 < \omega < (\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega$ so that

$$\int_{\log^2(t)/t}^1 \frac{1}{(\lambda^-+(\lambda^+-\lambda^-)u)^{1-\ell}\big[(\lambda^-+(\lambda^+-\lambda^-)u-\varrho)^2+\omega\big]}\sqrt{u(1-u)}\,\mathrm{d}u \leq C, \tag{138}$$

or equivalently this integral is bounded by some $C > 0$. Now we suppose that $\omega \leq 0$. The roots of $(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega$ are precisely given by $u = \frac{\varrho - \lambda^- \pm \sqrt{-\omega}}{\lambda^+ - \lambda^-}$. By (112) if $\gamma \neq \gamma_*$, we have that $\varrho + \sqrt{-\omega} < (1-\sqrt{r})^2 = \lambda^-$. Hence there exists a constant $\widehat{C} > 0$ such that $\widehat{C} < (\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega$ for all $u \in [0,1]$. It immediately follows that (138) holds. Now suppose that $\gamma = \gamma_*$. Then $\varrho + \sqrt{|\omega|} = (1-\sqrt{r})^2 = \lambda^-$ so the polynomial $(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega$ is 0 when $u = 0$. We observe that $u = 0$ is not a double root since there does not exist an r with $\gamma = \gamma_*, \omega = 0$, and $\gamma_* < \frac{2}{r}$ (here $\gamma = \gamma_*$ and $\omega = 0$ imply that $r = 1$, but then $\gamma_* = 2$ which violates $\gamma_* < \frac{2}{r}$). Consequently, we can write

$$(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega = (\lambda^+ - \lambda^-)^2 u(u + r_2),$$

where $r_2 > 0$ as the second root is negative. Since the Marchenko-Pastur measure has $\sqrt{u}$-behavior near 0, it immediately follows that

$$\begin{aligned}&\int_{\log^2(t)/t}^1 \frac{1}{(\lambda^-+(\lambda^+-\lambda^-)u)^{1-\ell}\big[(\lambda^-+(\lambda^+-\lambda^-)u-\varrho)^2+\omega\big]}\sqrt{u(1-u)}\,\mathrm{d}u \\ &\qquad\leq \frac{1}{r_2(\lambda^-)^{\ell-1}(\lambda^+-\lambda^-)^2}\int_0^1\frac{\sqrt{1-u}}{\sqrt{u}}\,\mathrm{d}u < C.\end{aligned}$$

Hence in all cases we have that

$$\begin{aligned}&\frac{(\lambda^+-\lambda^-)^2}{2\pi r}\int_{\log^2(t)/t}^1 \frac{e^{-(\lambda^+-\lambda^-)ut}}{(\lambda^-+(\lambda^+-\lambda^-)u)^{1-\ell}\big[(\lambda^-+(\lambda^+-\lambda^-)u-\varrho)^2+\omega\big]}\sqrt{u(1-u)}\,\mathrm{d}u \\ &\qquad\leq e^{-(\lambda^+-\lambda^-)\log^2(t)}C\end{aligned}$$

where $C > 0$ is some constant. Since $e^{-(\lambda^+-\lambda^-)\log^2(t)}$ decays faster than polynomial, this part of the integral does not have the interesting asymptotic. Now returning to (137) the interesting asymptotic occurs near $u = 0$. First we consider the setting where $\gamma \neq \gamma_*$. As we saw for $u \in [\log^2(t)/t, 1]$, we also know that there exists a constant $\widehat{C} > 0$ such that

$$\widehat{C} < (\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell}\big[(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega\big] \quad \text{for all } u \in [0, 1].$$

Using a change of variables $u = v/t$, we deduce that

$$\begin{aligned}
&\int_0^{\log^2(t)/t} \frac{e^{-(\lambda^+-\lambda^-)ut}}{(\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell}\big[(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega\big]} \sqrt{u(1-u)}\, \mathrm{d}u \\
&\qquad \leq \frac{1}{\widehat{C}} \cdot \frac{1}{t^{3/2}} \int_0^{\log^2(t)} e^{-(\lambda^+-\lambda^-)v} \sqrt{v(1-\tfrac{v}{t})}\, \mathrm{d}v \\
&\qquad \leq \frac{1}{\widehat{C}} \cdot \frac{1}{t^{3/2}} \int_0^{\infty} e^{-(\lambda^+-\lambda^-)v} \sqrt{v}\, \mathrm{d}v
\end{aligned}$$

Since the last integral is bounded, we can apply dominated convergence theorem. Using the change of variables, $u = \frac{v}{t}$, we have that

$$\begin{aligned}
&\int_0^{\log^2(t)/t} \frac{e^{-(\lambda^+-\lambda^-)ut}}{(\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell}\big[(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega\big]} \sqrt{u(1-u)}\, \mathrm{d}u \\
&= \frac{1}{t^{3/2}} \cdot \int_0^{\log^2(t)} \frac{e^{-(\lambda^+-\lambda^-)v}}{(\lambda^- + (\lambda^+ - \lambda^-)\frac{v}{t})^{1-\ell}\big[(\lambda^- + (\lambda^+ - \lambda^-)\frac{v}{t} - \varrho)^2 + \omega\big]} \sqrt{v(1-\tfrac{v}{t})}\, \mathrm{d}v \\
&\sim \frac{1}{t^{3/2}} \int_0^{\infty} \frac{e^{-(\lambda^+-\lambda^-)v}}{(\lambda^-)^{1-\ell}\big[(\lambda^- - \varrho)^2 + \omega\big]} \sqrt{v}\, \mathrm{d}v \\
&\sim \frac{1}{t^{3/2}} \cdot \frac{\sqrt{\pi}}{2(\lambda^+ - \lambda^-)^{3/2}(\lambda^-)^{1-\ell}((\lambda^- - \varrho)^2 + \omega)}.
\end{aligned} \tag{139}$$

Consequently, we get for any $\ell \in \{0, 1\}$

$$\begin{aligned}
&\frac{1}{2\pi r} \int_{\lambda^-}^{\lambda^+} \frac{e^{-xt}}{x^{1-\ell}\big((x - \varrho)^2 + \omega\big)} \sqrt{(x - \lambda^-)(\lambda^+ - x)}\, \mathrm{d}x \\
&\qquad\qquad \sim \frac{1}{4\sqrt{\pi} r} \cdot \frac{(\lambda^+ - \lambda^-)^{1/2}}{(\lambda^-)^{1-\ell}\big[(\lambda^- - \varrho)^2 + \omega\big]} \cdot e^{-\lambda^- t} \cdot \frac{1}{t^{3/2}}.
\end{aligned}$$

Now consider the setting where $\gamma = \gamma_*$. As we saw for $u \in [\log^2(t)/t, 1]$, we know that the polynomial has a root at $u = 0$ (not a double root). In particular, we get that

$$\begin{aligned}
&(\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell}\big[(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega\big] \\
&\qquad\qquad = (\lambda^+ - \lambda^-)^2(\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell} u(u + r_2),
\end{aligned}$$

where $-r_2$ is the second root of the quadratic $(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega$. We know this root is negative (i.e. $r_2 > 0$). Using a change of variables $u = v/t$ and a simple lower bound on $(\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell}$, we get that

$$\begin{aligned}
&\int_0^{\log^2(t)/t} \frac{e^{-(\lambda^+-\lambda^-)ut}}{(\lambda^- + (\lambda^+ - \lambda^-)u)^{1-\ell}\big[(\lambda^- + (\lambda^+ - \lambda^-)u - \varrho)^2 + \omega\big]} \sqrt{u(1-u)}\, \mathrm{d}u \\
&\qquad \leq \frac{1}{(\lambda^-)^{1-\ell}(\lambda^+ - \lambda^-)^2 r_2} \cdot \frac{1}{t^{1/2}} \int_0^{\log^2(t)} \frac{e^{-(\lambda^+-\lambda^-)v}}{v} \sqrt{v(1-\tfrac{v}{t})}\, \mathrm{d}v \\
&\qquad \leq \frac{1}{(\lambda^-)^{1-\ell}(\lambda^+ - \lambda^-)^2 r_2} \cdot \frac{1}{t^{1/2}} \int_0^{\infty} \frac{e^{-(\lambda^+-\lambda^-)v}}{\sqrt{v}}\, \mathrm{d}v
\end{aligned}$$

Since the last integral is bounded, we can apply dominated convergence theorem. Using the change of variables, $u = \frac{v}{t}$, we have that

$$\begin{aligned}
&\int_0^{\log^2(t)/t} \frac{e^{-(\lambda^+-\lambda^-)ut}}{(\lambda^- + (\lambda^+-\lambda^-)u)^{1-\ell}\big[(\lambda^+-\lambda^-)^2 u(u+r_2)\big]} \sqrt{u(1-u)}\,\mathrm{d}u \\
&= \frac{1}{t^{1/2}} \cdot \int_0^{\log^2(t)} \frac{e^{-(\lambda^+-\lambda^-)v}}{(\lambda^- + (\lambda^+-\lambda^-)\frac{v}{t})^{1-\ell}\big[(\lambda^+-\lambda^-)^2 v(\frac{v}{t}+r_2)\big]} \sqrt{v(1-\frac{v}{t})}\,\mathrm{d}v \\
&\sim \frac{1}{t^{1/2}} \int_0^{\infty} \frac{e^{-(\lambda^+-\lambda^-)v}}{(\lambda^-)^{1-\ell}\big[(\lambda^+-\lambda^-)^2 r_2\big]} \cdot \frac{1}{\sqrt{v}}\,\mathrm{d}v \\
&\sim \frac{1}{t^{1/2}} \cdot \frac{\sqrt{\pi}}{(\lambda^+-\lambda^-)^{5/2}(\lambda^-)^{1-\ell} r_2}.
\end{aligned} \tag{140}$$

Consequently, we have when $\gamma = \gamma_*$ that

$$\begin{aligned}
&\frac{1}{2\pi r} \int_{\lambda^-}^{\lambda^+} \frac{e^{-xt}}{x^{1-\ell}\big((x-\varrho)^2+\omega\big)} \sqrt{(x-\lambda^-)(\lambda^+-x)}\,\mathrm{d}x \\
&\qquad\qquad \sim \frac{1}{2\sqrt{\pi} r} \cdot \frac{1}{r_2(\lambda^+-\lambda^-)^{1/2}(\lambda^-)^{1-\ell}} \cdot e^{-\lambda^- t} \cdot \frac{1}{t^{1/2}}.
\end{aligned}$$

We note that the Dirac delta from the Marchenko-Pastur terms combined with the constant term yields that

$$\frac{\frac{2}{\gamma}(1-r)\big(1-\frac{r\gamma}{2}\big)}{\varrho^2+\omega} + \frac{\frac{2r}{\gamma}(1-r)\big(1-\frac{r\gamma}{2}\big)}{\varrho^2+\omega} \max\big\{0, 1-\tfrac{1}{r}\big\} = \frac{\frac{2}{\gamma}\big(1-\frac{r\gamma}{2}\big)}{\varrho^2+\omega} \max\{0, 1-r\}.$$

The result immediately follows. □

F Numerical simulation details and extra experiments

Problem setup. The vectors $\boldsymbol{x}_0$, $\widetilde{\boldsymbol{x}}$ and are sampled i.i.d. from the Gaussian $N(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I})$. The objective function in which we run SGD is in all cases the least squares objective function $f(\boldsymbol{x}) = \frac{1}{2n}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$, where $\boldsymbol{b}$ is generated as in (2). The definition of $\boldsymbol{A}$ is different depending on the data-generating process considered:

- For the isotropic features model, the rows of $\boldsymbol{A}$ are generated i.i.d. from a standard Gaussian distribution.
- In the one-hidden layer model these are generated following (25), with both $\boldsymbol{W}$ and and $\boldsymbol{Y}$ are i.i.d. from a standard Gaussian and $n/m = X$, $m/d = Y$, and g is a shifted hinge loss:

$$g(z) = \max(x, 0) - \frac{1}{\sqrt{2\pi}}\,. \tag{141}$$

  The substraction of $-\frac{1}{\sqrt{2\pi}}$ ensures that the zero Gaussian mean assumption is verified (27).

Throughout the experiments r is fixed to 1.5 and $\tilde{R} = 0$. We experimented with different values, and always obtained very similar qualitative results.

Algorithms. We simulate the **SME and SDE models** (see (17) for description) using Euler-Murayama discretization with stepsize 10^{-3}. The SDE model discretizes the same as equation as the SME model (Eq. (17)) with a scaled identity covariance ($\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$). In this model σ^2 is a free parameter which for the experiments we set to 0.1, as this value was giving the closest fit to SGD across the log-space grid of parameters $10^{-i}, i = 0, 1, ...$

For the **streaming model**, we use SGD updates and regenerate $\boldsymbol{a}_i$ (following the same model as SGD) at every step.

Extra experiments. We provide extra experiments following the same setting as in Figure 4 but with different choices of the ratio $r = \frac{d}{n}$ parameter.

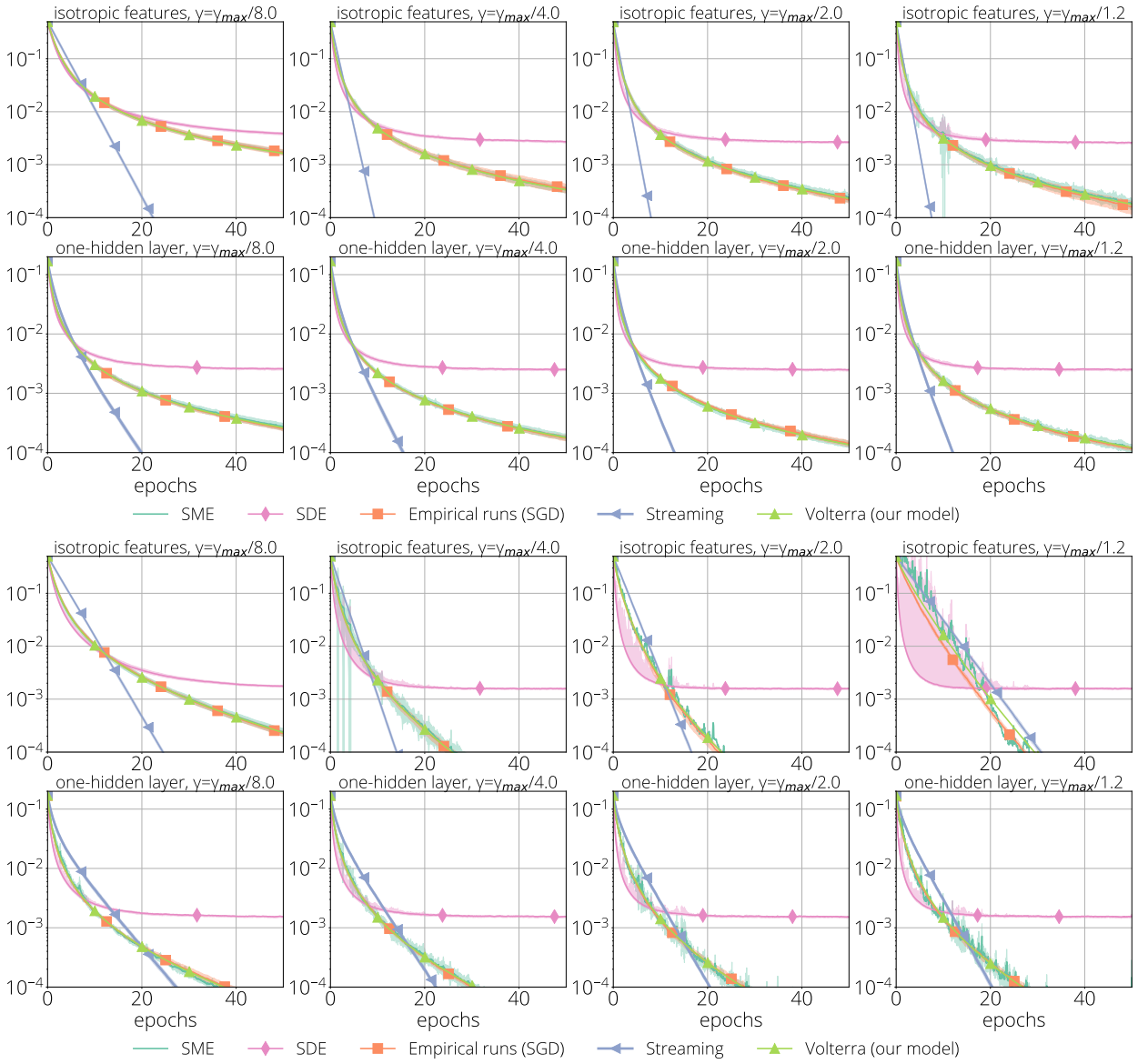

Figure 5: **Comparison of different SGD models with $r = 0.8$ and $r = 1.6$**: isotropic features (top) and one-hidden layer network (bottom).