A VIEW OF MINI-BATCH SGD VIA GENERATING FUNCTIONS: CONDITIONS OF CONVERGENCE, PHASE TRANSITIONS, BENEFIT FROM NEGATIVE MOMENTA

Anonymous authors
Paper under double-blind review

ABSTRACT

Mini-batch SGD with momentum is a fundamental algorithm for learning large predictive models. In this paper we develop a new analytic framework to analyze noise-averaged properties of mini-batch SGD for linear models at constant learning rates, momenta and sizes of batches. Our key idea is to consider the dynamics of the second moments of model parameters for a special family of "Spectrally Expressible" approximations. This allows to obtain an explicit expression for the generating function of the sequence of loss values. By analyzing this generating function, we find, in particular, that 1) the SGD dynamics exhibits several convergent and divergent regimes depending on the spectral distributions of the problem; 2) the convergent regimes admit explicit stability conditions, and explicit loss asymptotics in the case of power-law spectral distributions; 3) the optimal convergence rate can be achieved at negative momenta. We verify our theoretical predictions by extensive experiments with MNIST and synthetic problems, and find a good quantitative agreement.

1 INTRODUCTION

The fundamental algorithm of learning predictive models is the mini-batch Stochastic Gradient Descent (SGD) (Robbins & Monro, 1951; Bottou & Bousquet, 2007) with momentum (Polyak, 1964):

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{v}_{t+1}, \quad \mathbf{v}_{t+1} = -\alpha_t \nabla_{\mathbf{w}} L_{B_t}(\mathbf{w}_t) + \beta_t \mathbf{v}_t. \tag{1}$$

Here, $L_B(\mathbf{w}) = \frac{1}{b}\sum_{i=1}^{b} l(f(\mathbf{x}_i, \mathbf{w}), y_i)$ is the sampled loss of a model $\widehat{y} = f(\mathbf{x}, \mathbf{w})$, computed using a pointwise loss $l(\widehat{y}, y)$ on a mini-batch $B = \{(\mathbf{x}_i, y_i)\}_{s=1}^{b}$ of b data points representing the target function $y = f^*(\mathbf{x})$. The sampled loss $L_B(\mathbf{w})$ is a noisy estimate of the true loss $L(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)} l(f(\mathbf{x}, \mathbf{w}), y)$ recovered in the limit $b \to \infty$. This *sampling noise* decreases with the increase of batch size. We will assume that the stochasticity in the iterations (1) is only associated with this sampling noise (note that this assumption is very different from a popular assumption of *additive* noise which may be, e.g., due to numerical errors, see Section G).

The term $\mathbf{v}_n$ represents the *momentum* which is widely used and well-known to significantly improve convergence both generally (Polyak, 1987) and in the context of deep learning (Sutskever et al., 2013). Typically, β_t are assigned values from $[0, 1)$ close to 1, e.g. 0.9 or 0.99. For constant learning rates α and momenta β, the ratio $\alpha_{\text{eff}} = \alpha/(1-\beta) = \alpha + \alpha\beta + \alpha\beta^2 + \ldots$ is the *effective learning rate* reflecting the accumulated effect of multiple iterations (Tugay & Tanik, 1989; Yuan et al., 2016). We will always assume that $|\beta| < 1$, but allow β to be negative.

The goal of the present work is to derive explicit and easy-to-use analytic evolution laws for the SGD (1) applicable to modern deep neural networks trained on real world datasets. We argue now which further assumptions would be reasonable for that. First, in many (though not all) cases the nonlinear network training is reasonably well approximated (Lee et al., 2020) by the linearized NTK regime (Jacot et al., 2018), in which the loss L is quadratic with respect to parameters $\mathbf{w}$. Accordingly, from now on we restrict our attention to quadratic problems. Next, modern networks are usually overparameterized, and the respective problems are highly ill-conditioned. Moreover, one can often observe particular laws – typically, power laws – governing the spectral distribution of the problem:

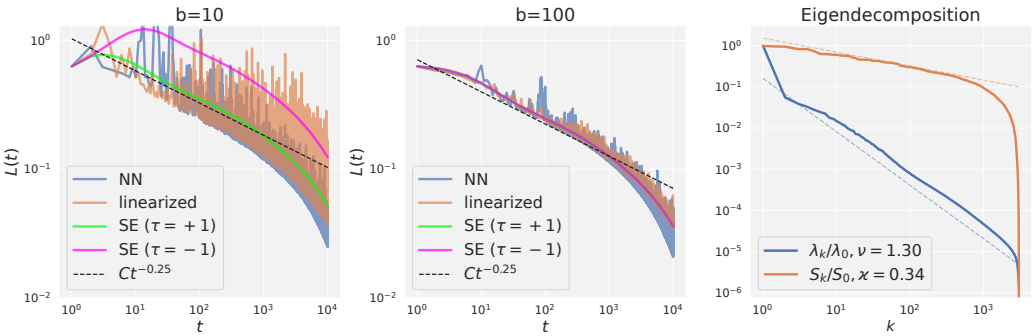

Figure 1: **Left, center**: Loss trajectories of different training regimes on MNIST for batch sizes 10 and 100. Actual nonlinear neural network (NN) is reasonably well approximated by the linearized model and its SE approximation with $\tau = 1$ (see Section 2). **Right**: eigenvalues λ_k and partial sums of coefficients $S_k = \sum_{l=k}^{N} c_l^2$. Experimental curves are fitted to the asymptotic power laws $L(t) \propto t^{-\xi}, \lambda_k \propto k^{-\nu}, S_k \propto k^{-\varkappa}$ resulting in $\nu \approx 1.30, \varkappa \approx 0.34, \xi \approx 0.25 \approx \frac{\varkappa}{\nu}$.

NTK eigenvalues λ_k and coefficients c_k of expansion of target function over NTK eigenvectors. (see Cui et al. (2021); Bahri et al. (2021); Lee et al. (2020); Canatar et al. (2021); Kopitkov & Indelman (2020); Dou & Liang (2020); Atanasov et al. (2021); Bordelon & Pehlevan (2021); Basri et al. (2020); Bietti (2021) and Figure 1). Taking these spectral distributions into account is expected to lead to reasonably accurate evolution laws.

We stress that our approach is different from the more common approach focused on general upper bounds for convergence rates. To illustrate this, recall the classical bounds for the basic (non-stochastic) gradient descent (Polyak, 1987). Assuming $\inf_{\mathbf{w}} L(\mathbf{w}) = 0$, the bound for the non-strongly-convex quadratic optimization reads $L(\mathbf{w}_t) \leq \|\mathbf{w}_0 - \mathbf{w}_*\|^2/(4\alpha e t) = O(1/t)$, while in the strongly convex case $L(\mathbf{w}_t) \leq (\lambda_{\max}/2)\|\mathbf{w}_0 - \mathbf{w}_*\|^2(\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}})^{2t} = O(q^t), q < 1$. One might expect these formulas to reasonably describe the training loss trajectory of a MNIST classifier at large batch sizes. In reality, however, this trajectory looks rather differently – it rather falls off as $\propto t^{-0.25}$ up to a large number of iterations ($\sim 10^3 - 10^4$, see Figure 1). This effect, along with the exponent 0.25, can be explained using the spectral power laws associated with this problem (Nemirovskiy & Polyak, 1984; Velikanov & Yarotsky, 2021). We aim to achieve this level of quantitative agreement with experiment for SGD with momentum.

Main contributions. The central contribution of our paper is an analytic framework for the study of mini-batch SGD. Its first step (Section 2) is a "spectrally expressible" approximation which allows to express SGD dynamics only through spectral distributions of the problem . In the next step (Section 3) we express generating function of the loss values through these spectral distributions. Finally, in Section 4 we use this framework to address various practically important questions:

1. **Convergence criteria.** We derive a precise condition of loss convergence in terms of the Hessian spectrum λ_k and show that for small $1 - \beta$ it simplifies to $\alpha_{\text{eff}} \lesssim 2b/\lambda_{\text{crit}}$, where λ_{crit} is the critical regularization at which the effective dimension of the problem equals 1.

2. **Phase transitions.** For non-strongly convex problems we show that our model exhibits distinct "signal-dominated" and "noise-dominated" convergence regimes (previously known for SGD without momenta (Varre et al., 2021)) depending on the spectral tails of eigenvalues and target expansion coefficients. We derive exact loss asymptotics for these regimes. We show that the noise-dominated regime is typically preceded by a time interval of signal-dominated regime, with transition occuring on a time scale t_{trans} that we predict analytically. We also estimate the time scale t_{blowup} of loss divergence for divergent models.

3. **Hyperparameter and budget optimization.** We show that in the signal-dominated regime it is favorable to choose large effective learning rates α_{eff}. However, because of the stability constraint $\alpha_{\text{eff}} \lesssim 2b/\lambda_{\text{crit}}$, admissible α_{eff} are limited if batch size $b < \infty$ (in stark contrast to noiseless GD). In agreement with earlier works, we find that the loss obtained with an optimal choice of α, β can be approximately expressed as a function of the total

computation budget bt. In contrast to the signal-dominated phase, in the noise-dominated phase it is counter-productive to use large effective learning rates.

4. **Negative momenta.** It is common to use SGD with momentum $\beta \in (0, 1)$. We find that in the signal-dominated regime it is always advantageous to include in SGD a momentum $\beta > 0$, but in the noise-dominated regime it may actually be advantageous to choose $\beta < 0$.

For lack of space, related work is discussed in Appendix A and experimental details [1] in Appendix F.

2 THE SETTING AND THE *Spectrally Expressible* APPROXIMATION

SGD. We consider a linear model $f(\mathbf{x}, \mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}) \rangle$ with non-linear features $\boldsymbol{\psi}(\mathbf{x})$ trained to approximate a target function $y^*(\mathbf{x})$ by minimizing quadratic loss $L(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^{N} (\langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}_i) \rangle - y^*(\mathbf{x}_i))^2$ over training dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$. The model's Jacobian is given by $\boldsymbol{\Psi} = \left(\boldsymbol{\psi}(\mathbf{x}_i)\right)_{i=1}^N$, the Hessian $\mathbf{H} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\psi}(\mathbf{x}_i) \otimes \boldsymbol{\psi}(\mathbf{x}_i) = \frac{1}{N} \boldsymbol{\Psi} \boldsymbol{\Psi}^T$ and the tangent kernel $K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle$. Assuming that the target function is representable as $y^*(\mathbf{x}) = \langle \mathbf{w}^*, \boldsymbol{\psi}(\mathbf{x}) \rangle$, the loss takes the form $L(\mathbf{w}) = \frac{1}{2} \langle \mathbf{w} - \mathbf{w}^*, \mathbf{H}(\mathbf{w} - \mathbf{w}^*) \rangle$. We allow model parameters to be either finite dimensional vectors or belong to a Hilbert space $\mathcal{H}$. Similarly, dataset $\mathcal{D}$ can be either finite ($N < \infty$) or infinite, and in the latter case we only require a finite norm of the target function $y^*(\mathbf{x})$ but not of the solution $\mathbf{w}^*$. Note that this setting is quite rich and can accommodate kernel methods, linearization of real neural networks, and infinitely wide neural networks in the NTK regime. See Appendix B for details. In the experimental results, we refer to this setting as *linearized* (to be distinguished from nonlinear NN's and the SE approximation introduced below).

Denoting deviation from the optimum as $\Delta \mathbf{w} \equiv \mathbf{w} - \mathbf{w}^*$, we can write a single SGD step at iteration t with a randomly chosen batch B_t of size b as

$$\begin{pmatrix} \Delta \mathbf{w}_{t+1} \\ \mathbf{v}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H}(B_t) & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \Delta \mathbf{w}_t \\ \mathbf{v}_t \end{pmatrix}, \quad \mathbf{H}(B_t) \equiv \frac{1}{b} \sum_{i \in B_t} \boldsymbol{\psi}(\mathbf{x}_i) \otimes \boldsymbol{\psi}(\mathbf{x}_i). \tag{2}$$

An important feature of the multiplicative noise introduced by the random choice of B_t is that the n'th moments of parameter and momentum vectors $\Delta \mathbf{w}_{t+1}, \mathbf{v}_{t+1}$ in dynamics (2) are fully determined from the same moments at the previous step t. As our loss function is quadratic, we focus on the second moments

$$\mathbf{C} \equiv \mathbb{E}[\Delta \mathbf{w} \otimes \Delta \mathbf{w}], \quad \mathbf{J} \equiv \mathbb{E}[\Delta \mathbf{w} \otimes \mathbf{v}], \quad \mathbf{V} \equiv \mathbb{E}[\mathbf{v} \otimes \mathbf{v}], \quad \mathbf{M} \equiv \begin{pmatrix} \mathbf{C} & \mathbf{J} \\ \mathbf{J}^\dagger & \mathbf{V} \end{pmatrix}. \tag{3}$$

Then the average loss is $\mathbb{E}[L(\mathbf{w})] = \frac{1}{2} \operatorname{Tr}(\mathbf{HC})$ and for combined second moment matrix we have

Proposition 1. *Consider SGD* (1) *with learning rates α_t, momentum β_t, batch size b and random uniform choice of the batch B_t. Then the update of second moments* (3) *is*

$$\begin{pmatrix} \mathbf{C}_{t+1} & \mathbf{J}_{t+1} \\ \mathbf{J}_{t+1}^\dagger & \mathbf{V}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{C}_t & \mathbf{J}_t \\ \mathbf{J}_t^\dagger & \mathbf{V}_t \end{pmatrix} \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix}^T + \gamma \alpha_t^2 \begin{pmatrix} \boldsymbol{\Sigma}(\mathbf{C}_t) & \boldsymbol{\Sigma}(\mathbf{C}_t) \\ \boldsymbol{\Sigma}(\mathbf{C}_t) & \boldsymbol{\Sigma}(\mathbf{C}_t) \end{pmatrix} \tag{4}$$

Here $\gamma \boldsymbol{\Sigma}(\mathbf{C}_t)$ is the covariance of gradient noise due to mini-batch sampling, with $\gamma = \frac{N-b}{(N-1)b}$ and

$$\boldsymbol{\Sigma}(\mathbf{C}) = \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{\psi}(\mathbf{x}_i), \mathbf{C} \boldsymbol{\psi}(\mathbf{x}_i) \rangle \boldsymbol{\psi}(\mathbf{x}_i) \otimes \boldsymbol{\psi}(\mathbf{x}_i) - \mathbf{HCH}. \tag{5}$$

SE Approximation. An important property of non-stochastic GD is that it can be "solved" by diagonalizing the Hessian $\mathbf{H}$. Namely, typical quantities of interest in the context of optimization have the form $\operatorname{Tr}[\phi(\mathbf{H})\mathbf{C}]$ with some function ϕ: e.g., $\phi(x) = \frac{x}{2}$ for the loss, $\phi(x) = 1$ for the parameter displacement, $\phi(\mathbf{x}) = x^2$ for the squared gradient norm. Given a Hessian $\mathbf{H}$, by Riesz–Markov theorem we can associate to a positive definite $\mathbf{C}$ the respective *spectral measure* $\rho_{\mathbf{C}}(d\lambda)$:

$$\operatorname{Tr}[\phi(\mathbf{H})\mathbf{C}] = \int \phi(\lambda) \rho_{\mathbf{C}}(d\lambda), \quad \forall \phi \in C(\operatorname{spec}(\mathbf{H})). \tag{6}$$

We will assume that $\mathbf{H}$ has an eigenbasis $\mathbf{u}_k$ with eigenvalues λ_k; in this case $\rho_{\mathbf{C}} = \sum_k C_{kk} \delta_{\lambda_k}$, where $C_{kk} = \langle \mathbf{u}_k, \mathbf{C} \mathbf{u}_k \rangle$. For vanilla GD, the spectral measure of the second moment matrix $\mathbf{C}_t$ at

[1] Our code: `https://anonymous.4open.science/r/PowerLawOptimization-1401/`

iteration t is given by $\rho_{\mathbf{C}_t}(d\lambda) = p_t^2(\lambda)\rho_{\mathbf{C}_0}(d\lambda)$, where $p_t(x)$ is a polynomial fully determined by the learning algorithm (Fischer, 1996). Thus, the information encoded in the initial spectral measure $\rho_{\mathbf{C}_0}$ is, in principle, sufficient to compute main characteristics of the optimization trajectories.

This conclusion, however, does not hold for SGD, since the first component of the noise term (5) uses finer, non-spectral details of the problem. Our key idea is then to approximate actual transformation (5) by the class of spectral measure transformations parameterized by $\tau_1, \tau_2 \in \mathbb{R}, \tau_1 \geq \tau_2$:

$$\rho_{\mathbf{\Sigma}}(d\lambda) = \tau_1\Big(\int \lambda' \rho_{\mathbf{C}}(d\lambda')\Big)\rho_{\mathbf{H}}(d\lambda) - \tau_2\lambda^2\rho_{\mathbf{C}}(d\lambda), \quad \rho_{\mathbf{H}} = \sum\nolimits_k \lambda_k \delta_{\lambda_k}. \tag{7}$$

We refer to this as the *Spectrally Expressible (SE)* approximation. Under this approximation we again can, in principle, reconstruct the trajectory of observables $\mathrm{Tr}[\phi(\mathbf{H})\mathbf{C}_t]$ from the initial $\mathbf{C}_0$.

There are multiple reasons justifying approximation (7). To begin with, there are scenarios where Eq. (7) holds exactly. First, we prove this for translation-invariant models, with $\tau_1 = 1, \tau_2 = 1$:

Proposition 2. *Consider a problem with regular grid dataset* $\mathcal{D} = \{\mathbf{x_i} = (\frac{2\pi i_1}{N_1}, \dots, \frac{2\pi i_d}{N_d})\}, \quad \mathbf{i} \in (\mathbb{Z}/N_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/N_d\mathbb{Z})$ *on the d-dimensional torus* $\mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$ *and with translation invariant kernel* $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x} - \mathbf{x}')$. *Then* (7) *holds exactly with* $\tau_1 = 1, \tau_2 = 1$.

Second, Bordelon & Pehlevan (2021) show that when the features $\boldsymbol{\psi}(\mathbf{x})$ are Gaussian w.r.t. $\mathbf{x} \sim \mathcal{D}$, the SE approximation is exact with $\tau_1 = 1, \tau_2 = -1$.

On the other hand, Eq. (7) can be derived by generalizing the map (5) and requiring it to be "spectrally expressible". Namely, replace summation over i in (5) by a general linear combination with coefficients $\widehat{R} = (R_{i_1 i_2 i_3 i_4})$ only assuming the symmetries $R_{i_1 i_2 i_3 i_4} = R_{i_2 i_1 i_3 i_4} = R_{i_1 i_2 i_4 i_3}$:

$$\mathbf{\Sigma}_{\widehat{R}}(\mathbf{C}) = \frac{1}{N^2} \sum_{i_1, i_2, i_3, i_4 = 1}^{N} R_{i_1 i_2 i_3 i_4} \langle \boldsymbol{\psi}(\mathbf{x}_{i_3}), \mathbf{C}\boldsymbol{\psi}(\mathbf{x}_{i_4})\rangle \boldsymbol{\psi}(\mathbf{x}_{i_1}) \otimes \boldsymbol{\psi}(\mathbf{x}_{i_2}) - \mathbf{HCH}. \tag{8}$$

Assuming the coefficients $\widehat{R}$ to be fixed and independent of $\mathbf{\Psi}, \mathbf{H}, \mathbf{C}$, we prove

Theorem 1. *For an SGD dynamics* (4) *with noise term* (8)*, the following statements are equivalent:* ***(1)*** *Spectral measures* $\rho_{\mathbf{C}_t}$ *occurring during SGD are uniquely determined by the initial spectral measure* $\rho_{\mathbf{C}_0}$ *and by the eigenvalues* λ_k *of Hessian* $\mathbf{H}$*;* ***(2)*** $R_{i_1 i_2 i_3 i_4} = \tau_1 \delta_{i_1 i_2}\delta_{i_3 i_4} - \frac{\tau_2 - 1}{2}(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3})$ *for some* τ_1, τ_2*, and Eq.* (7) *holds;* ***(3)*** *The trajectory* $\{\mathbf{C}_t\}_{t=0}^{\infty}$ *is invariant under transformations* $\widetilde{\mathbf{\Psi}} = \mathbf{\Psi}\mathbf{U}^T$ *of the feature matrix* $\mathbf{\Psi}$ *by any* $\mathbf{U} \in \mathrm{O}(N)$.

Finally, we observe empirically that SE approximation works well on realistic problems such as MNIST for predicting loss evolution (figure 1 and section F.4), determination of convergence conditions (figure 2) and approximation of original noise term (5) in terms of trace norm (section F.4).

In the reminder we focus on analytical treatment of SGD dynamics under SE approximation (7). Note that parameters γ, τ_1, τ_2 enter evolution equations in combinations $\gamma\tau_1, \gamma\tau_2$. This allows to lighten the notation in the sequel by setting $\tau_1 = 1$ and $\tau_2 = \tau$ without restricting generality.

3 Reduction to generating functions

Gas of rank-1 operators. Iterations (4) are linear and can be compactly written as $\mathbf{M}_{t+1} = F_t \mathbf{M}_t$ with a linear operator F_t. Under SE approximation (7), for all four blocks of $\mathbf{M}_t$ we only need to consider the diagonal components in the eigenbasis of $\mathbf{H}$, e.g. $C_{kk,t} \equiv \langle \mathbf{u}_k, \mathbf{C}_t \mathbf{u}_k \rangle$. Then

$$L(T) = \tfrac{1}{2}\,\mathrm{Tr}\,\mathbf{HC}_T = \tfrac{1}{2}\langle \left(\begin{smallmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 0 \end{smallmatrix}\right), F_T F_{T-1} \cdots F_1 \left(\begin{smallmatrix} \mathbf{C}_0 & 0 \\ 0 & 0 \end{smallmatrix}\right)\rangle. \tag{9}$$

Here $\left(\begin{smallmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 0 \end{smallmatrix}\right) = \Big(\left(\begin{smallmatrix} \lambda_1 & 0 \\ 0 & 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} \lambda_2 & 0 \\ 0 & 0 \end{smallmatrix}\right), \dots\Big), \left(\begin{smallmatrix} \mathbf{C}_0 & 0 \\ 0 & 0 \end{smallmatrix}\right) = \Big(\left(\begin{smallmatrix} C_{11,0} & 0 \\ 0 & 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} C_{22,0} & 0 \\ 0 & 0 \end{smallmatrix}\right), \dots\Big) \in \mathbb{R}^{\mathbb{N}} \otimes \mathbb{R}^{2\times 2}$ are (2×2)-matrix-valued sequences indexed by the eigenvalues λ_k, and $\langle \cdot, \cdot \rangle$ is the natural scalar product in $l^2 \otimes \mathbb{R}^{2\times 2}$. The evolution operator F_t can be represented as the sum $F_t = Q_t + A_t$ of a rank-one operator Q_t (noise term) and an operator A_t (main term) that acts independently at each eigenvalue λ_k as a 4×4 matrix:

$$Q_t = \gamma_t \alpha_t^2 \boldsymbol{\lambda} \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) \langle \left(\begin{smallmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 0 \end{smallmatrix}\right), \cdot \rangle, \tag{10}$$

$$A_t = (A_{t,\lambda}), \quad A_{t,\lambda} M = \left(\begin{smallmatrix} 1-\alpha_t\lambda & \beta_t \\ -\alpha_t\lambda & \beta_t \end{smallmatrix}\right) M \left(\begin{smallmatrix} 1-\alpha_t\lambda & \beta_t \\ -\alpha_t\lambda & \beta_t \end{smallmatrix}\right)^T - \tau\gamma_t\alpha_t^2\lambda^2 \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right) M \left(\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right)^T, \tag{11}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)^T$. Let us expand the loss by the binomial formula, with m terms Q_t chosen at positions $t_1, \ldots, t_m$ and $T - m$ terms A_t at the remaining positions:

$$L(T) = \frac{1}{2} \sum_{m=0}^{T} \sum_{0<t_1<\ldots<t_m<T+1} U_{T+1,t_m} U_{t_m,t_{m-1}} \cdots U_{t_2,t_1} V_{t_1}, \tag{12}$$

where

$$U_{t,s} = \langle \left(\begin{smallmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 0 \end{smallmatrix}\right), A_{t-1} A_{t-2} \cdots A_{s+1} \boldsymbol{\lambda} \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) \rangle \gamma_s \alpha_s^2, \tag{13}$$

$$V_t = \langle \left(\begin{smallmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 0 \end{smallmatrix}\right), A_{t-1} A_{t-2} \cdots A_1 \left(\begin{smallmatrix} \mathbf{C}_0 & 0 \\ 0 & 0 \end{smallmatrix}\right) \rangle. \tag{14}$$

Expansion (12) has a suggestive interpretation as a partition function of a gas of rank-1 operators interacting with nearest neighbors (via $U_{t,s}$) and the origin (via V_t). Interactions via $U_{t.s}$ contain the factor γ so that their strength depends on the sampling noise. We expect the model to be in different phases depending on the amount of noise: in the "signal-dominated" regime $L(T)$ is primarily determined by V_t at large T, while in the "noise-dominated" regime $L(T)$ is primarily determined by $U_{t,s}$. Below we show that such a phase transition indeed occurs for non-strongly convex problems.

Generating functions. Expansion (12) allows to compute $L(T)$ iteratively:

$$L(T) = \frac{1}{2} V_{T+1} + \sum_{t_m=1}^{T} U_{T+1,t_m} L(t_m - 1). \tag{15}$$

For constant learning rates and momenta $\alpha_t \equiv \alpha, \beta_t \equiv \beta$ we have $F_t \equiv F$, the value $U_{t,s}$ becomes translation invariant, $U_{t,s} = U_{t-s}$, and the loss can be conveniently described by generating functions (or equivalently, Laplace transform):

$$\widetilde{L}(z) = \sum_{t=0}^{\infty} L(t) z^t = \frac{\widetilde{V}(z)/2}{1 - z\widetilde{U}(z)}, \tag{16}$$

where $\widetilde{U}$ and $\widetilde{V}$ are the "noise" and "signal" generating functions:

$$\widetilde{U}(z) = \sum_{t=0}^{\infty} U_{t+1} z^t = \gamma \alpha^2 \langle \left(\begin{smallmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 0 \end{smallmatrix}\right), (1 - zA)^{-1} \boldsymbol{\lambda} \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) \rangle \tag{17}$$

$$= \gamma \alpha^2 \sum_k \lambda_k^2 (\beta z + 1) / S(\alpha, \beta, \tau\gamma, \lambda_k, z), \tag{18}$$

$$\widetilde{V}(z) = \sum_{t=0}^{\infty} V_{t+1} z^t = \langle \left(\begin{smallmatrix} \boldsymbol{\lambda} & 0 \\ 0 & 0 \end{smallmatrix}\right), (1 - zA)^{-1} \left(\begin{smallmatrix} \mathbf{C}_0 & 0 \\ 0 & 0 \end{smallmatrix}\right) \rangle \tag{19}$$

$$= \sum_k \lambda_k C_{kk,0} (2\alpha\beta\lambda_k z + \beta^3 z^2 - \beta^2 z - \beta z + 1) / S(\alpha, \beta, \tau\gamma, \lambda_k, z) \tag{20}$$

$$\begin{aligned} S(\alpha, \beta, \gamma, \lambda, z) = {} & \alpha^2\beta\gamma\lambda^2 z^2 + \alpha^2\beta\lambda^2 z^2 + \alpha^2\gamma\lambda^2 z - \alpha^2\lambda^2 z - 2\alpha\beta^2\lambda z^2 - 2\alpha\beta\lambda z^2 \\ & + 2\alpha\beta\lambda z + 2\alpha\lambda z - \beta^3 z^3 + \beta^3 z^2 + \beta^2 z^2 - \beta^2 z + \beta z^2 - \beta z - z + 1. \end{aligned}$$

In the remainder we derive various properties of the loss evolution by analyzing these formulas.

4 Convergence, phase transitions, parameter optimization, negative momenta

In this section we assume a non-strongly-convex scenario with an infinite-dimensional and compact $\mathbf{H}$, i.e., $\lambda_k \to 0$. Also, we assume that $0 < \tau \leq 1$ in SE approximations: this simplifies some statements and fits the experiment for several practical problems (see Figure 1 and Appendix F).

Conditions of loss convergence (stability). First note that as $\lambda_k \to 0$ the "main" components A_{λ_k} of the evolution operator F appearing in Eq. (11) have their action in trial matrix M converge to $\left(\begin{smallmatrix} 1 & \beta \\ 0 & \beta \end{smallmatrix}\right) M \left(\begin{smallmatrix} 1 & \beta \\ 0 & \beta \end{smallmatrix}\right)^T$, and from (13) we get $U_t = const \sum_{k=1}^{\infty} \lambda_k^2 (1 + o(1))$, for each t. This means that if $\sum_k \lambda_k^2 = \infty$ then each $U_t = \infty$ and loss diverges already at first step ($L(t = 1) = \infty$) – we call this effect *immediate divergence*. On the other hand, assuming $\sum_k \lambda_k^2 < \infty$, loss stability can be related to the first positive singularity of the loss generating function $\widetilde{L}(z)$ given by (16). Indeed, let r_L be the convergence radius of power series (16). Since $L(t) \geq 0$, $\widetilde{L}(z)$ must be monotone

increasing on $[0, r_L)$ and have a singularity at $z = r_L$. If $r_L < 1$, then $L(t) \not\to 0$ as $t \to \infty$, while if $r_L > 1$, then $L(t) \to 0$ exponentially fast. Thus, large-t loss stability can be characterized by the condition $r_L \geq 1$, which can be further related to the generating function $\widetilde{U}(z)$.

Proposition 3. *Let $\beta \in (-1, 1)$, $0 < \alpha < 2(\beta+1)/\lambda_{\max}$ and $\gamma \in [0, 1]$. Then $r_L < 1$ iff $\widetilde{U}(1) > 1$.*

This result is proved by showing that $\widetilde{U}(z), \widetilde{V}(z)$ have no singularities for $z \in (0, 1)$ and that $z\widetilde{U}(z)$ is monotone increasing on $(0, 1)$. The only source of singularity in $\widetilde{L}(z)$ is then the denominator in (16). Note that conditions $\beta \in (-1, 1),\ 0 < \alpha < 2(\beta + 1)/\lambda_{\max}$ exactly specify the convergence region for the non-stochastic problem with $\gamma = 0$ (see (Roy & Shynk, 1990; Tugay & Tanik, 1989) and Appendix H). This is also a necessary condition in the presence of sampling noise since the second term in (4) is a positive semi-definite matrix.

Evaluating $\widetilde{U}(z)$ at $z = 1$, we find

$$\widetilde{U}(1) = \sum_k \frac{\gamma\alpha\lambda_k(\beta + 1)}{\alpha\beta\tau\gamma\lambda_k + \alpha\beta\lambda_k + \alpha\tau\gamma\lambda_k - \alpha\lambda_k - 2\beta^2 + 2} = \frac{\alpha\gamma}{2(1-\beta)} \sum_k \lambda_k(1 + O(\lambda_k)). \quad (21)$$

This shows that $\sum_k \lambda_k = \infty$ iff $\widetilde{U}(1) = \infty > 1$ (regardless of values of α, β, γ), so that $r_L < 1$ and $L(t)$ diverges exponentially as $t \to \infty$. We refer to this scenario as *eventual divergence*.

We examine next the practically common case of momentum close to 1, when $\delta := 1 - \beta \to 0+$. From the first equality in (21) we get $\widetilde{U}(1) = \sum_k \frac{\lambda_k(1+O(\alpha\lambda_k))}{\tau\lambda_k + 2\delta/(\alpha\gamma)}$. Dropping the $O(\alpha\lambda_k)$ term and assuming for simplicity that $\tau = 1$, the convergence condition $\widetilde{U}(1) < 1$ can be approximately written in terms of the regularization parameter $\lambda = \lambda_{\text{crit}}$ at which the "effective dimension" $d_{\text{eff}} = \sum_k \lambda_k/(\lambda_k + \lambda)$ of $\mathbf{H}$ equals 1:

$$\alpha_{\text{eff}}\gamma \lesssim \frac{2}{\lambda_{\text{crit}}}, \text{ where } \sum_k \frac{\lambda_k}{\lambda_k + \lambda_{\text{crit}}} = 1. \quad (22)$$

Both conditions (21), (22) agree very well with experiments on MNIST and synthetic data (Fig. 2).

Exponential divergence. Our stability analysis above shows that loss diverges at large t when $\widetilde{U}(1) > 1$. This divergence is primarily characterized by the convergence radius r_L determined from the equation $r_L\widetilde{U}(r_L) = 1$ according to (16). Indeed, assuming an asymptotically exponential ansatz $L(t) \sim Ce^{t/t_{\text{div}}}$, we immediately get the divergence time scale $t_{\text{div}} = -1/\log r_L$. A closer inspection of $\widetilde{L}(z)$ near r_L also allows to determine the constant C (see appendix E.4):

$$L(t) \sim \frac{\widetilde{V}(r_L)}{2(1 + r_L^2\widetilde{U}'(r_L))} r_L^{-t}. \quad (23)$$

Note that r_L can be very close to 1, which is naturally observed in *eventual divergent* phase by taking sufficiently small α and/or γ. In this case the characteristic divergence time t_{div} is large and we expect divergent behavior to occur only for $t \gtrsim t_{\text{blowup}}$, while for $t \ll t_{\text{blowup}}$ the loss converges roughly with the rate of noiseless model $\widetilde{U}(z) \equiv 0$. As t_{blowup} should have a meaning of a time moment where loss starts to significantly deviate from noiseless trajectory we define it by equating divergent (23) and noiseless losses. We confirm these effects experimentally in Figure 3 (right).

Power laws and two convergence regimes. Note that if $\lambda_k \to 0$ and $C_{kk,0} > 0$ for infinitely many k, then $r_L \leq 1$, thus excluding the possibility of exponential loss convergence. Indeed, $S(\alpha, \beta, \gamma, \lambda, z) \xrightarrow{\lambda \to 0} (1 - z)(1 - \beta z)(1 - \beta^2 z)$. It follows that for small λ_k the respective terms in expansion (18) have three poles $z_{k,0}, z_{k,1}, z_{k,2}$ converging to $1, \beta^{-1}, \beta^{-2}$, respectively. By Eq. (16), $\widetilde{L}(z)$ has a removable singularity at $z_{k,0}$, moreover $\widetilde{L}(z_{k,0}) < 0$ if $C_{kk,0} > 0$ and k is large enough, meaning that $r_L < z_{k,0}$. We then get $r_L \leq 1$ by taking the limit in k.

To get a more detailed picture of the loss evolution let us assume now that the eigenvalues λ_k and the second moments $C_{kk,0}$ of initial ($t = 0$) approximation error are subject to large-k power laws

$$\lambda_k = \Lambda k^{-\nu}(1 + o(1)), \quad \sum_{l \geq k} \lambda_l C_{ll,0} = K k^{-\varkappa}(1 + o(1)) \quad (24)$$

with some exponents $\nu, \varkappa > 0$ (also denote $\zeta = \varkappa/\nu$) and coefficients Λ, K. Such (or similar) power laws are empirically observed in many high-dimensional problems (Cui et al. (2021); Bahri et al. (2021); Lee et al. (2020); Canatar et al. (2021); Kopitkov & Indelman (2020); Dou & Liang (2020); Atanasov et al. (2021); Bordelon & Pehlevan (2021); Basri et al. (2020); Bietti (2021) and Figure 1), can be derived theoretically (Basri et al., 2020; Velikanov & Yarotsky, 2021), and are often assumed for theoretical optimization guarantees (Nemirovskiy & Polyak, 1984; Brakhage, 1987; Hanke, 1991; 1996; Gilyazov & Gol'dman, 2013; Varre et al., 2021; Caponnetto & De Vito, 2007; Steinwart et al., 2009; Berthier et al., 2020; Zou et al., 2021). If $\nu \le \frac{1}{2}$, then $\sum_k \lambda_k^2 = \infty$ and we are in the "immediate divergence" regime mentioned earlier, and if $\frac{1}{2} < \nu \le 1$, then $\sum_k \lambda_k = \infty$ and we are in "eventual divergence" regime. Away from these two divergent regimes, we precisely characterize the late time loss asymptotics:

Theorem 2. *Assume spectral conditions* (24) *with* $\nu > 1$ *and that parameters* α, β, γ *are as in Proposition 3 and such that convergence condition* $\widetilde{U}(1) < 1$ *is satisfied. Then*

$$\sum_{t=1}^{T} tL(t) = (1+o(1)) \begin{cases} \frac{K\Gamma(\zeta+1)}{2(1-\widetilde{U}(1))(2-\zeta)} \left(\frac{2\alpha\Lambda}{1-\beta}\right)^{-\zeta} t^{2-\zeta}, & 2-\zeta > 1/\nu, \\ \frac{\gamma \widetilde{V}(1)\Gamma(2-1/\nu)}{8(1-\widetilde{U}(1))^2} \left(\frac{2\alpha\Lambda}{1-\beta}\right)^{1/\nu} t^{1/\nu}, & 2-\zeta < 1/\nu. \end{cases} \tag{25}$$

The idea of the proof is to 1) observe by Proposition 3 that $r_L = 1$; 2) show that for $z = 1-\varepsilon$ close to the singularity $z = 1$ the derivative $\frac{d}{dz}\widetilde{L}(z) = \sum_{t=1}^{\infty} tL(t) z^{t-1}$ of the loss generating function diverges as $\frac{d}{dz}\widetilde{L}(1-\varepsilon) \propto \varepsilon^{\xi-2}$; 3) apply the Tauberian theorem to relate the asymptotic $\frac{d}{dz}\widetilde{L}(1-\varepsilon) \approx C\varepsilon^{\xi-2}$ to the asymptotic of the partial sum $\sum_t^T tL(t) \approx \frac{C}{\Gamma(3-\xi)} T^{2-\xi}$. The second step can be done by writing

$$\frac{d}{dz}\widetilde{L}(z) = \frac{\frac{d}{dz}\widetilde{V}(z)}{2(1-z\widetilde{U}(z))} + \frac{\widetilde{V}(z)\frac{d}{dz}(z\widetilde{U}(z))}{2(1-z\widetilde{U}(z))^2} \tag{26}$$

One can then show using (18), (19) that $\widetilde{U}'(z)$ diverges as $\widetilde{U}'(1-\varepsilon) \propto \varepsilon^{-\frac{1}{\nu}}$ and $\widetilde{V}'(z)$ diverges as $\widetilde{V}'(1-\varepsilon) \propto \varepsilon^{\zeta-2}$ for $\zeta < 2$. Depending on which of the two terms dominates in Eq. (26) at small ε, we observe two qualitatively different phases (previously known in SGD without momenta (Varre et al., 2021)): the "signal-dominated" phase at $\zeta - 2 < -1/\nu$ and the "noise-dominated" phase at $\zeta - 2 > -1/\nu$.

In the sequel we will assume that not only the cumulative sums $\sum_1^T tL(t)$, but also the individual values $L(t)$ obey a power law $L(t) = C(1+o(1))t^{-\xi}$. Then from (25) we immediately get

$$L(t) = (1+o(1)) \begin{cases} \frac{K\Gamma(\zeta+1)}{2(1-\widetilde{U}(1))} \left(\frac{2\alpha\Lambda}{1-\beta}\right)^{-\zeta} t^{-\zeta}, & \zeta < 2-1/\nu, \quad \text{(27a)} \\ \frac{\gamma \widetilde{V}(1)\Gamma(2-1/\nu)}{8\nu(1-\widetilde{U}(1))^2} \left(\frac{2\alpha\Lambda}{1-\beta}\right)^{1/\nu} t^{1/\nu-2}, & \zeta > 2-1/\nu. \quad \text{(27b)} \end{cases}$$

See Figure 3(left) for the resulting full phase diagram.

Transitions between phases. Explicit result (25) allow us to go further in understanding the "noise-dominated" phase $\zeta > 2 - 1/\nu$. Representation (26) and linearity of Laplace transform allows to write loss asymptotic as $L(t) \approx C_{\text{signal}} t^{-\zeta} + C_{\text{noise}} t^{-2+\frac{1}{\nu}}$ with two terms given by (27a),(27b). As C_{noise} vanishes in the noiseless limit $\gamma \to 0$, we expect that for sufficiently small γ the signal term here dominates the noise term up to time scale t_{trans} given by

$$t_{\text{trans}} = \left(\frac{C_{\text{signal}}}{C_{\text{noise}}}\right)^{1/(\zeta-2+1/\nu)} = \left(\left(\frac{1-\beta}{2\alpha\Lambda}\right)^{1/\nu+\zeta} \frac{4\nu K\Gamma(\zeta+1)(1-\widetilde{U}(1))}{\gamma\Gamma(2-1/\nu)\widetilde{V}(1)}\right)^{1/(\zeta-2+1/\nu)}. \tag{28}$$

This conclusion is fully confirmed by our experiments with simulated data, see Fig. 3(center). Note that the time scale t_{trans} can vary widely depending on α, β, γ; in particular, t_{trans} might not be reached at all in realistic optimization time if the noise γ and effective learning rate $\frac{\alpha}{1-\beta}$ are small.

A similar reasoning can be used to estimate the time t_{blowup} of transition from early convergence to subsequent divergence (see Figure 3 (right)). In Section E.4 we consider the scenario with $\beta = 0, \gamma = \tau = 1$ and power laws (24) in which $\frac{1}{2} < \nu < 1$ (eventual divergence) and $\zeta < 1$. We show that at small α we have $t_{\text{blowup}} \approx a_* t_{\text{div}} \approx a_* \left(\frac{1}{4\nu}\Gamma(2-1/\nu)\Gamma(1/\nu-1)\right)^{\nu/(\nu-1)} (2\alpha\Lambda)^{1/(\nu-1)}$, where $a_* = a_*(\nu, \zeta)$ is the solution of the equation $\frac{1/\nu-1}{\Gamma(1-\zeta)} a_*^{-\zeta} = e^{a_*}$.

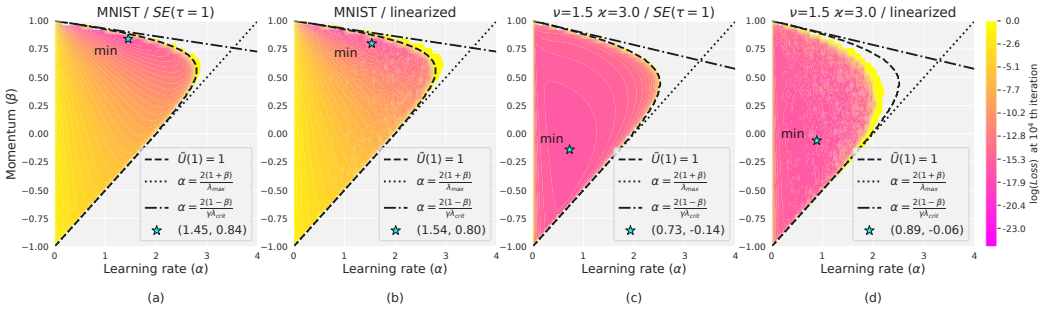

Figure 2: $L(t = 10^4)$ for different learning rates α and momenta β, in the SE($\tau = 1$) and linearized regimes, for MNIST and synthetic data given by Eq. (24) with $\nu = 1.5, \varkappa = 3$, for $b = 10$. Black lines show convergence boundaries from Proposition (3) and Eqs. (21) and (22). MNIST experiment is in the signal-dominated phase and synthetic in the noise-dominated. The stars show $\arg\min_{\alpha,\beta} L$.

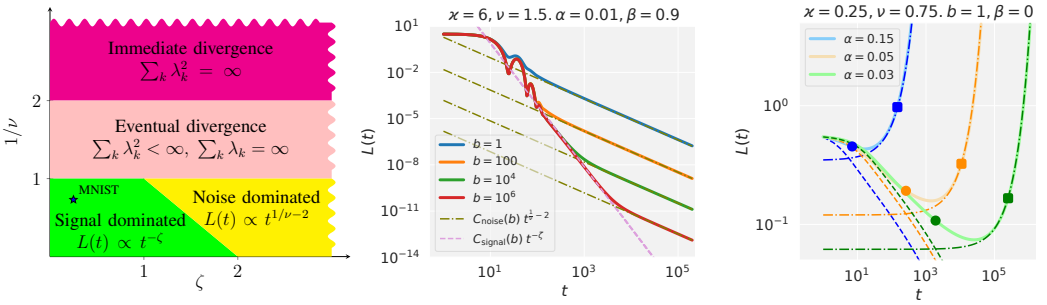

Figure 3: **Left**: Phase diagram for $L(t)$ under power-law spectral conditions (24). **Center**: Transition from signal-dominated to noise-dominated regime during training. Asymptotic power-law lines are calculated from (27a) and (27b). **Right**: Divergent $L(t)$ in the phase $\frac{1}{2} < \nu < 1$: experimental trajectories (solid) with marked points at positions $t = t_{\text{div}}$ (squares) and $t = t_{\text{blowup}}$ (circles), theoretical late-time asymptotics (23) (dash-dotted) and noiseless early time trajectories (dashed).

Hyperparameter and budget optimization. The two natural concepts of budget in our setting are the *time budget* T (the total number of iterations) and the *computation budget* bT. We discuss now minimization of the loss value for available budgets. It is convenient to consider separately the two phases and assume that T is large enough for the asymptotic formulas (27a),(27b) to be applicable. Note that near the divergence boundary $\widetilde{U}(1) = 1$ the loss is high due to denominator $1 - \widetilde{U}(1)$ in Eqs. (27a),(27b). This suggests that at optimal α, β the difference $\widetilde{U}(1) - 1$ is of order 1, which allows us to neglect dependence of loss on $\widetilde{U}(1)$ as long as convergence condition $\widetilde{U}(1) < 1$ holds.

"Signal-dominated" phase ($\zeta < 2 - 1/\nu$). By Eq. (27a), $L(t) \gtrsim C(\alpha_{\text{eff}} t)^{-\zeta}$ with a problem-dependent constant C. Thus to minimize $L(t)$ for a fixed time budget t, we need to maximize α_{eff}. However, we are limited by convergence condition (22) stating that $\alpha_{\text{eff}} \lesssim \frac{2}{\lambda_{\text{crit}}\gamma}$, and implying in particular that $L(t) \gtrsim C'\gamma^{\zeta} t^{-\zeta}$. Since γ is determined by the batch size (see Proposition 1), once the batch size is limited, we get a general lower bound on the asymptotic of $L(t)$ that cannot be improved by any adjustment of α, β. This contrasts markedly with the non-stochastic GD ($\gamma = 0$), in which one can improve loss convergence by bringing β arbitrarily close to 1 and hence making α_{eff} arbitrarily large (Nemirovskiy & Polyak, 1984; Flammarion & Bach, 2015). See the two left subplots in Figure 2 for an illustration on MNIST.

Recalling that $\gamma \approx \frac{1}{b}$ for $N \gg b$, we see that at the optimal α_{eff} the loss $L(t) \approx C'(bt)^{-\xi}$, i.e. it is (approximately) a function of the total computation budget bt regardless of how it is factorized over b and t. This conclusion agrees with existing results on approximately linear scalability of SGD with batch size (Ma et al., 2018; Shallue et al., 2018). (We remark that this scalability breaks down at larger batch sizes; there is no contradiction with above power law since Eq. (27a) and hence the power law are established for $t \to \infty$). See Fig. 5 in Appendix for an illustration on MNIST.

"Noise-dominated" phase ($\zeta > 2 - 1/\nu$). Here Eq. (27b) implies $L(t) \approx C\gamma\widetilde{V}(1)\alpha_{\text{eff}}^{1/\nu} t^{1/\nu - 2}$. We see in particular that α_{eff} appears with the *positive* power $1/\nu$, so that (in contrast to the signal-

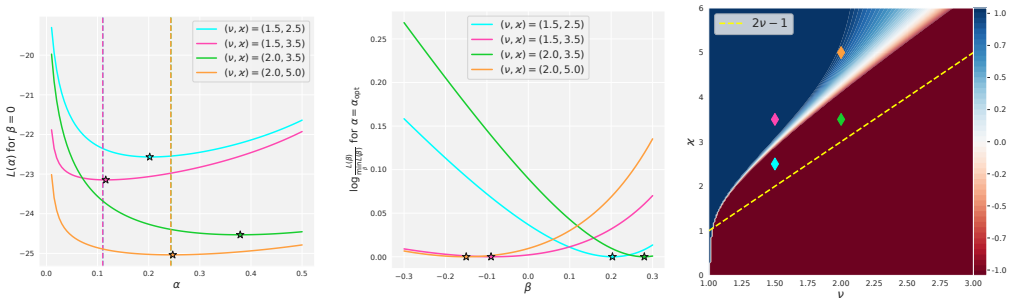

Figure 4: **Left**: Dependence of the loss on the learning rate α at $\beta = 0$, for different ν and $\varkappa$ (synthetic data) in the "noisy" phase. The stars show the experimentally determined optimal learning rates $\arg\min_\alpha L(\alpha)$, while the dashed vertical lines show the respective theoretical values $\alpha_{\text{opt}} = (\sum_k \lambda_k)^{-1} \frac{2(\nu-1)}{3\nu-1}$. The experimental and theoretical optima agree well for two pairs $(\nu, \varkappa)$ deep in the noisy phase, but not so much for $(\nu, \varkappa)$ near the phase boundary (cf. right figure). **Center**: Dependence of the loss log-ratio $\log \frac{L(\beta)}{\min_\beta L(\beta)}$ on β at experimental α_{opt}. The stars denote $\text{argmin}_\beta L(\beta)$. In two scenarios the minimum is attained at $\beta < 0$, and in the other two at $\beta > 0$. **Right**: The value (29) characterizing improvement from $\beta < 0$ at large t by Proposition 9, computed for the range $(\nu, \varkappa) \in [1, 3] \times [0, 6]$ of power-law models (24). Diamonds show the pairs $(\nu, \varkappa)$ in the experiments on the left and central figures. In agreement with our theoretical prediction, the value (29) is positive in the scenarios where the optimal β is negative, and vice versa. The dashed yellow line $\varkappa = 2\nu - 1$ separates the noise-dominated (top) and the signal-dominated (bottom) regimes.

dominated scenario) high effective learning rates are disadvantageous. At the same time, it can be shown that $\widetilde{V}(1) \propto \frac{1}{\alpha}$ at small α, so it is neither favorable to use small α. Experiments with synthetic data in the noise-dominated regime in Fig. 2 show that the loss is indeed minimized when both α is kept away from 0 and α_{eff} from the maximum admissible value $\frac{2}{\lambda_{\text{crit}}\gamma}$.

Negative momenta. It is known that in some noisy settings momenta $\beta > 0$ do not improve convergence compared to $\beta = 0$ (Polyak, 1987; Kidambi et al., 2018). We show now that in our noise-dominated regime the optimal configuration of α, β may be achieved with a *negative* momentum β (two right subplots in Fig. 2). To clarify this effect, assume that approximation (27b) is applicable and can be differentiated w.r.t. α and β. Assume also that $\gamma = 1$ (this allows to simplify some formulas). Then, by computing $\partial_\alpha L(t)|_{\beta=0}$ using approximation (27b), it can be shown that $L(t)|_{\beta=0}$ is minimized at $\alpha = \alpha_{\text{opt}} \approx \left(\sum_k \lambda_k\right)^{-1} \frac{2(\nu-1)}{3\nu-1}$. Now we compute $\partial_\beta L(t)|_{\beta=0,\alpha=\alpha_{\text{opt}}}$ and find that its sign is equal to the sign of the expression

$$\nu \sum_k \lambda_k \sum_k \lambda_k C_{kk,0} - (\nu - 1) \sum_k \lambda_k^2 \sum_k C_{kk,0}. \tag{29}$$

It follows that if this expression is positive, then pair $(\alpha_{\text{opt}}, \beta)$ with a small negative β provides a lower loss than any pair (α, β) with $\beta = 0$. We verify this conclusion experimentally in Figure 4.

Interestingly, we do not observe (both experimentally and theoretically) this effect for the signal-dominated phase. Retracing above steps with Eq. (27b) replaced by Eq. (27a), we find $\partial_\beta L(t)|_{\beta=0,\alpha=\alpha_{\text{opt}}}$ to be always negative. We conclude that in the signal-dominated phase the loss is minimized at $\beta > 0$. See Appendix E.6 and Proposition 9 there for details.

5 Conclusion

We have proposed a two step framework which allowed us to get a precise characterization of mini-batch SGD dynamics with momentum. In the first step we performed a *spectrally expressible* approximation for the exact dynamics of the second moments of model weights. This approximation makes the dynamics tractable and is even exact in a number of scenarios. In the second step we developed a formalism based on generating functions that clearly distinguishes the roles of noiseless part and sampling noise in the overall SGD evolution with a quantitative agreement with experiments. Finally, we applied the formalism to SGD with fixed learning rate, momentum and batch size and non-strongly convex problems to answer several questions including convergence conditions, different phases and transitions between them, and optimal SGD settings.

6 BIBLIOGRAPHY

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. *arXiv preprint arXiv:2111.00034*, 2021.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *International Conference on Machine Learning*, pp. 685–694. PMLR, 2020.

Saugata Basu. Algorithms in real algebraic geometry: a survey. *arXiv preprint arXiv:1409.1534*, 2014.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *arXiv preprint arXiv:2006.08212*, 2020.

Alberto Bietti. Approximation and learning with deep convolutional models: a kernel perspective. *arXiv preprint arXiv:2102.10032*, 2021.

Blake Bordelon and Cengiz Pehlevan. Learning curves for sgd on structured features. *arXiv preprint arXiv:2106.02713*, 2021.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/0d3180d672e08b4c5312dcdafdf6ef36-Paper.pdf.

Benjamin Bowman and Guido Montufar. Spectral bias outside the training set for deep networks in the kernel regime. *arXiv preprint arXiv:2206.02927*, 2022.

Helmut Brakhage. On ill-posed problems and the method of conjugate gradients. In *Inverse and ill-posed Problems*, pp. 165–175. Elsevier, 1987.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1), may 2021. doi: 10.1038/s41467-021-23103-1. URL https://doi.org/10.1038%2Fs41467-021-23103-1.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34, 2021.

Andreas Dolzmann and Thomas Sturm. Redlog: Computer algebra meets computer logic. *Acm Sigsam Bulletin*, 31(2):2–9, 1997.

Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *Journal of the American Statistical Association*, 116(535):1507–1520, apr 2020. doi: 10.1080/01621459.2020.1745812. URL https://doi.org/10.1080%2F01621459.2020.1745812.

William Feller. *An introduction to probability theory and its applications. Vol. II.* Second edition. John Wiley & Sons Inc., New York, 1971.

B. Fischer. Polynomial based iteration methods for symmetric linear systems. 1996.

Nicolas Flammarion and Francis Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pp. 658–695. PMLR, 2015.

Sergei Farshatovich Gilyazov and Nataliâ L'vovna Gol'dman. *Regularization of ill-posed problems by iteration methods*, volume 499. Springer Science & Business Media, 2013.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Martin Hanke. Accelerated landweber iterations for the solution of ill-posed equations. *Numerische mathematik*, 60(1):341–373, 1991.

Martin Hanke. Asymptotics of orthogonal polynomials and the numerical solution of ill-posed problems. *Numerical Algorithms*, 11(1):203–213, 1996.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Prateek Jain, Dheeraj Nagaraj, and Praneeth Netrapalli. Making the last iterate of sgd information theoretically optimal. In *Conference on Learning Theory*, pp. 1752–1755. PMLR, 2019.

Harold Jeffreys. On isotropic tensors. *Mathematical Proceedings of the Cambridge Philosophical Society*, 73(1):173–176, 1973. doi: 10.1017/S0305004100047587.

Hui Jin, Pradeep Kr Banerjee, and Guido Montúfar. Learning curves for gaussian process regression with power-law priors and targets. *arXiv preprint arXiv:2110.12231*, 2021.

Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9. IEEE, 2018.

Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

Hermann König. *Eigenvalue distribution of compact operators*, volume 16. Birkhäuser, 2013.

Dmitry Kopitkov and Vadim Indelman. Neural spectrum alignment: Empirical study. In *International Conference on Artificial Neural Networks*, pp. 168–179. Springer, 2020.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jaehoon Lee, Samuel Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. *Advances in Neural Information Processing Systems*, 33:15156–15172, 2020.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL https://doi.org/10.7717/peerj-cs.103.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

Arkadi S Nemirovskiy and Boris T Polyak. Iterative methods for solving linear ill-posed problems under precise information. *Eng. Cyber.*, (4):50–56, 1984.

Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate o(kˆ2). In *Doklady Akademii Nauk*, volume 269, pp. 543–547. Russian Academy of Sciences, 1983.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

B.T. Polyak. *Introduction to Optimization*. Optimization Software, New York, 1987.

John Proakis. Channel identification for high speed digital communications. *IEEE Transactions on Automatic Control*, 19(6):916–922, 1974.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Sumit Roy and John J Shynk. Analysis of the momentum lms algorithm. *IEEE transactions on acoustics, speech, and signal processing*, 38(12):2088–2098, 1990.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600*, 2018.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 71–79, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/shamir13.html`.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pp. 79–93, 2009.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Mehmet Ali Tugay and Yalcin Tanik. Properties of the momentum lms algorithm. *Signal Processing*, 18(2):117–127, 1989.

Aditya Varre and Nicolas Flammarion. Accelerated sgd for non-strongly-convex least squares. *arXiv preprint arXiv:2203.01744*, 2022.

Aditya Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *arXiv preprint arXiv:2102.03183*, 2021.

Maksim Velikanov and Dmitry Yarotsky. Explicit loss asymptotics in the gradient descent training of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Hermann Weyl. The classical groups. 1939.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, pp. 10367–10376. PMLR, 2020.

Kun Yuan, Bicheng Ying, and Ali H Sayed. On the influence of momentum acceleration on online learning. *The Journal of Machine Learning Research*, 17(1):6602–6667, 2016.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize sgd for linear regression. *arXiv preprint arXiv:2103.12692*, 2021.

CONTENTS

1 Introduction **1**

2 The setting and the *Spectrally Expressible* approximation **3**

3 Reduction to generating functions **4**

4 Convergence, phase transitions, parameter optimization, negative momenta **5**

5 Conclusion **9**

6 Bibliography **10**

A Related work **14**

B Details of the setting **15**

B.1 Basics 15

B.2 Eigendecompositions and other details 16

B.3 Output and parameter spaces 17

C Dynamics of second moments **18**

D SE approximation **21**

D.1 Dynamics of spectral measures 21

D.2 Theorem 1 22

D.3 Translation invariant models 25

E Generating functions and their applications **26**

E.1 Reduction to generating functions 26

E.2 Convergence conditions 27

E.3 Loss asymptotic under power-laws 29

E.4 Exponential divergence of the loss and convergence-divergence transition 32

E.5 Budget optimization 35

E.6 Negative momenta 36

F Experiments **38**

F.1 Training regimes 38

F.2 Datasets 39

F.3 Experimental details 39

F.4 Validity of spectrally expressible approximation 39

F.5 Additional experiments 41

G Mini-batch SGD vs. SGD with additive noise **43**

H Noiseless convergence condition **43**

A RELATED WORK

SGD. Stochastic Gradient Descent (SGD) was introduced in Robbins & Monro (1951), and its various versions have been extensively studied since then, see e.g. Kiefer & Wolfowitz (1952); Rumelhart et al. (1986); Bottou & Bousquet (2007); Shamir & Zhang (2013); Jain et al. (2019); Harvey et al. (2019); Moulines & Bach (2011). The version (1) considered in the present paper is motivated by applications to training neural networks and is widely used (LeCun et al., 2015; Goodfellow et al., 2016; Sutskever et al., 2013). Its key feature is the presence of momentum, which is both theoretically known to improve convergence and widespread in practice (Polyak, 1964; 1987; Nesterov, 1983; Sutskever et al., 2013; Shallue et al., 2018).

Mini-batch SGD and spectrally expressible representation. In the present paper we consider the *mini-batch* SGD, meaning that the stochasticity in SGD is specifically associated with the random choice of different batches used to approximate the full loss function at each iteration (not to be confused with SGD modeled as gradient descent with additive noise as e.g. in the gradient Langevin dynamics; see e.g. Wu et al. (2020); Zhu et al. (2018) for some discussion and comparison, as well as our Section G for an illustration in our case of quadratic problems). We study this SGD in terms of how the second moments of the errors evolve with time. Our Proposition 1 generalizes existing evolution equations (Varre et al., 2021) to SGD with momentum. A serious difficulty in the study of this evolution is the presence of higher order mixing terms coupling different spectral components (the first term on the r.h.s. in our Eq. (5)). In Varre et al. (2021), it was shown that the effects of these terms can be controlled by suitable inequalities, allowing to establish upper bounds on error rates. Our approach is different and is close to the approach of recent paper Bordelon & Pehlevan (2021): we rather *assume* that the full evolution can be replaced by what we call *spectrally expressible* approximation (see Section 2 for details and discussion). We consider a family of such approximations parameterized by scalar τ, while Bordelon & Pehlevan (2021) consider a particular case (in our notation, $\tau = -1$). The SE assumption allows us to go beyond upper bounds and obtain explicit coefficients in the asymptotics of large-t loss evolution (see Theorem 2).

Spectral power laws. In Section 4 we adopt at some point very convenient *power law spectral assumptions* (24) on the asymptotic behavior of eigenvalues and target expansion coefficients. Such power laws are empirically observed in many high-dimensional problems (Cui et al., 2021; Bahri et al., 2021; Lee et al., 2020; Canatar et al., 2021; Kopitkov & Indelman, 2020; Dou & Liang, 2020; Atanasov et al., 2021; Bordelon & Pehlevan, 2021; Basri et al., 2020; Bietti, 2021), can be derived theoretically (Basri et al., 2020; Velikanov & Yarotsky, 2021), and closely related conditions are often assumed for theoretical optimization guarantees (Nemirovskiy & Polyak, 1984; Brakhage, 1987; Hanke, 1991; 1996; Gilyazov & Gol'dman, 2013; Varre et al., 2021; Caponnetto & De Vito, 2007; Steinwart et al., 2009; Berthier et al., 2020; Zou et al., 2021; Jin et al., 2021; Bowman & Montufar, 2022). An important aspect of our work is that, in contrast to most other works, we allow the exponent ζ characterizing initial second moments $C_{kk,0}$ to have values in the interval $(0, 1)$. In this case, $\sum_k \lambda_k C_{kk} < \infty$ while $\sum_k C_{kk,0} = \infty$. This effectively means that the target function $y^*(\mathbf{x})$ has a finite euclidean norm, while the norm of the respective optimal weight vector $\mathbf{w}^*$ is infinite (i.e., the loss does not attain a minimum value). This broader assumption allows us, in particular, to correctly describe the loss evolution on MNIST for which $\zeta \approx 0.25$ (see Figures 1 and 3): without assuming $\|\mathbf{w}^*\| = \infty$, the SGD is proved to have error rate $O(t^{-\xi})$ with $\xi \geq 1$ (Varre et al., 2021), which clearly contradicts the experiment (see Figure 1(left)).

Phase diagram. It is known that performance of SGD can be decomposed into "bias" and "variance" terms, with the dominant term determining what we call "signal-dominated" or "noise-dominated" regime (Moulines & Bach, 2011; Varre et al., 2021; Varre & Flammarion, 2022). In particular, see Varre et al. (2021) for respective upper bounds for SGD without momentum. The main novelty of our paper compared to these previous results is the analytic framework allowing to obtain large-t loss asymptotics with explicit constants for SGD with momentum (see Theorem 2) and use them to derive various quantitative conclusions related to model stability, phase transitions, and parameter optimization (see Section 4).

Noisy GD with momentum. GD with momentum (Heavy Ball) (Polyak, 1964) was extensively studied, but mostly in a noiseless setting or for generic (non-sampling) kinds of noise; see e.g. the monograph Polyak (1987). See Proakis (1974); Roy & Shynk (1990); Tugay & Tanik (1989) for a discussion of stability on quadratic problems. Polyak (1987) analyses various methods with noisy gradients for different kinds of noise and concludes that the fast convergence advantage of Heavy Ball and Conjugate Gradients (compared to GD) are not preserved under noise unless it decreases to 0 as the optimization trajectory approaches the solution. See also Kidambi et al. (2018) for further comparisons of noisy GD with or without momentum. Our work refines this picture in the case of sampling noise. Our results in Section 4 suggest that including positive momentum in SGD always improves convergence in the signal-dominated phase, but generally not in the noise-dominated phase (where it may even be beneficial to use a negative momentum).

Optimization of batch size. It is known that for SGD with not too large batches, there is approximately linear scalability of learning w.r.t. batch size, in the sense that to reach a given loss value with an increased batch size the number of iterations should be decreased proportionally (Ma et al., 2018; Shallue et al., 2018). This law is naturally no longer valid at very large batch sizes, since in the limit of infinitely large batch size SGD just converges to the non-stochastic GD, so that further increasing batch size when it it already large brings little benefit (the effect of "diminishing returns"). Our theoretical analysis in Section 4 based on large-t loss asymptotics is consistent with linear scalability.

B Details of the setting

B.1 Basics

In our problem setting we always consider regression tasks of fitting target function $y^\star(\mathbf{x})$ using linear models $f(\mathbf{x}, \mathbf{w})$ of the form

$$f(\mathbf{x}, \mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}) \rangle \tag{30}$$

where $\langle \cdot, \cdot \rangle$ stands for a scalar product and $\boldsymbol{\psi}(\mathbf{x})$ are some (presumably non-linear) features of inputs $\mathbf{x}$. The space $\mathcal{X}$ containing inputs $\mathbf{x}$ is not important in our context. Next, the model will be always trained to minimize a quadratic loss function over training dataset $\mathcal{D}$:

$$L(\mathbf{w}) = \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} (\langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}) \rangle - y^\star(\mathbf{x}))^2. \tag{31}$$

Denote by B_t a set of $b = |B_t|$ training samples $\mathbf{x}$ drawn independently and uniformly without replacement from $\mathcal{D}$. Then a single step of SGD with learning rate α_t and momentum parameter β_t takes the form

$$\begin{pmatrix} \mathbf{w}_{t+1} \\ \mathbf{v}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H}(B_t) & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{w}_t \\ \mathbf{v}_t \end{pmatrix} + \begin{pmatrix} \alpha_t \mathbf{h}(B_t) \\ \alpha_t \mathbf{h}(B_t) \end{pmatrix} \tag{32}$$

$$\mathbf{H}(B_t) \equiv \frac{1}{b} \sum_{\mathbf{x}_i \in B_t} \boldsymbol{\psi}(\mathbf{x}_i) \otimes \boldsymbol{\psi}(\mathbf{x}_i) \tag{33}$$

$$\mathbf{h}(B_t) \equiv \frac{1}{b} \sum_{\mathbf{x}_i \in B_t} \boldsymbol{\psi}(\mathbf{x}_i) y^\star(\mathbf{x}_i) \tag{34}$$

Note that expression (32) of SGD step is equivalent to more compact expression (2) when there exist optimal parameters $\mathbf{w}^\star$ such that $\langle \mathbf{w}^\star, \boldsymbol{\psi}(\mathbf{x}) \rangle = y^\star(\mathbf{x})$. As we have already noted, existence of optimal parameters with $\|\mathbf{w}^\star\| < \infty$ may not be the case for models relevant to practical problems (e.g. MNIST). Thus we will consider (32) to be the basic form of SGD step in our work. However, the formal usage of (2) instead of (32) always leads to the same conclusions and therefore we focus on the case $\|\mathbf{w}^\star\| < \infty$ in the main paper. In subsection B.3 we consider formulation of SGD in the output space, which allows to treat cases $\|\mathbf{w}^\star\| = \infty$ and $\|\mathbf{w}^\star\| < \infty$ withing the same framework.

It is convenient to distinguish several cases of our problem setting differing by finite/infinite dimensionality d of parameters $\mathbf{w}$, and finite/infinite size $|\mathcal{D}| = N$ of the dataset used for training. In our work we take into account all these cases, as each of them is used in either experimental or theoretical part of the paper.

B.2 EIGENDECOMPOSITIONS AND OTHER DETAILS

Case 1: $d, N < \infty$**.** This is the simplest case and we will often repeat the respective paragraph of section 2. First, recall that matrix elements $d \times N$ Jacobian matrix $\mathbf{\Psi}$ are

$$\Psi_{ij} = \psi_i(\mathbf{x}_j), \ \mathbf{x}_j \in \mathcal{D} \tag{35}$$

We will assume without loss of generality that $\mathbf{\Psi}$ has a full rank $r = \min(d, N)$ (the general case reduces to this after a suitable projection). Then by (31) the Hessian of the model is

$$\mathbf{H} = \frac{1}{N}\mathbf{\Psi}\mathbf{\Psi}^T = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{\psi}(\mathbf{x}_j)\boldsymbol{\psi}^T(\mathbf{x}_j). \tag{36}$$

Another important matrix is the kernel $K(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^T\psi(\mathbf{x}')$ calculated on the training dataset $\mathcal{D}$

$$\mathbf{K} = \frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi} = \frac{1}{N}\left(\sum_{i=1}^{d} \psi_i(\mathbf{x}_j)\psi_i(\mathbf{x}_{j'})\right)_{j,j'=1}^{N} \tag{37}$$

Next, consider the SVD decomposition of Jacobian $\mathbf{\Psi} = \sqrt{N}\mathbf{U}\mathbf{\Lambda}\mathbf{V}$ with $\mathbf{U} = (\mathbf{u}_k)_{k=1}^d$ being the matrix of left eigenvectors, $\mathbf{V} = (\boldsymbol{\phi}_k^T)_{k=1}^N$ being the matrix of right eigenvectors, and rectangular $d \times N$ diagonal matrix with singular values $\mathbf{\Lambda}_{kk} = \sqrt{\lambda_k}, \ k = 1 \dots r$ on the diagonal. The Hessian and kernel matrices are symmetric and have eigendecompositions with shared spectrum of non-zero eigenvalues: $\lambda_k, \mathbf{u}_k$ for $\mathbf{H}$ and $\lambda_k, \mathbf{v}_k$ for $\mathbf{K}$.

Now let us proceed to the target function $y^\star(\mathbf{x})$. As in this paper we focus on the training error (31), we may also restrict ourselves to the target vector $\mathbf{y}^\star$ of target function values at training points: $y_i^\star = \frac{1}{\sqrt{N}}y^\star(\mathbf{x}_i), \ \mathbf{x}_i \in \mathcal{D}$. The normalization by $\sqrt{N}$ is done here to allow for direct correspondence with the $N = \infty$ case. If $d \leq N$, we recall our assumption of $\mathbf{\Psi}$ having full rank and therefore the target function is reachable since the equation $\frac{1}{\sqrt{N}}\mathbf{w}^T\mathbf{\Psi} = \mathbf{y}^\star$ always has a solution. However, if $d < N$, there are many possible solutions of $\frac{1}{\sqrt{N}}\mathbf{w}^T\mathbf{\Psi} = \mathbf{y}^\star$, and we choose $\mathbf{w}^*$ to be the one with the same projection on $\ker \mathbf{H}$ as the initial model parameters $\mathbf{w}_0$. Later, this choice will guarantee that during SGD dynamic (32) parameter deviation always lies in the image of the Hessian: $\Delta\mathbf{w}_t \equiv \mathbf{w}_t - \mathbf{w}^\star \in \operatorname{Im}\mathbf{H}$. Finally, if $N > d$, the target may by unreachable: $\mathbf{y}^\star \notin \operatorname{Im}\mathbf{\Psi}^{\boldsymbol{T}}$. In this case we consider the decomposition of the output space $\mathbb{R}^N = \operatorname{Im}\mathbf{K} \oplus \ker\mathbf{K}$ and the respective decomposition of target vector $\mathbf{y}^\star = \mathbf{y}_\parallel^\star + \mathbf{y}_\perp^\star$. Then $\mathbf{y}_\parallel^\star$ is always reachable $\mathbf{y}_\parallel^\star \in \operatorname{Im}\mathbf{\Psi}^{\boldsymbol{T}}$ and in the rest we simply consider reachable part of the target $\mathbf{y}_\parallel^\star$ instead of the original target $\mathbf{y}^\star$.

Apart from general finite regression problems, the case $d, N < \infty$ described above applies to linearization of practical neural networks (and other parametric models). Specifically, consider a model $f(\mathbf{w}, \mathbf{x})$ initialized at $\mathbf{w}_0$ with initial predictions $f_0(\mathbf{x}) \equiv f(\mathbf{w}_0, \mathbf{x})$. The training of linearized model $f(\mathbf{w}, \mathbf{x}) = f_0(\mathbf{x}) + \langle \mathbf{w} - \mathbf{w}_0, \nabla_{\mathbf{w}} f(\mathbf{w}_0, \mathbf{x})\rangle$ can be mapped to out basic problem setting (30) with replacements $\boldsymbol{\psi}(\mathbf{x}) \longleftarrow \nabla_{\mathbf{w}} f(\mathbf{w}_0, \mathbf{x}), \ \ \mathbf{w} \longleftarrow \mathbf{w} - \mathbf{w}_0, \ \ y^\star(\mathbf{x}) \longleftarrow y^\star(\mathbf{x}) - f_0(\mathbf{x})$.

Case 2: $d = \infty, N < \infty$**.** In this case we assume that model parameters $\mathbf{w}$ and features $\boldsymbol{\psi}(\mathbf{x})$ belong to a Hilbert space $\mathcal{H}$, and $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathcal{H}$. For the Hessian we have

$$\mathbf{H} = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{\psi}(\mathbf{x}_j) \otimes \boldsymbol{\psi}(\mathbf{x}_j), \quad \mathbf{x}_j \in \mathcal{D} \tag{38}$$

We assume that the features $\boldsymbol{\psi}(\mathbf{x}_j), \ \mathbf{x}_j \in \mathcal{D}$ are linearly independent (as is usually the case in practice), then $\dim \operatorname{Im}\mathbf{H} = N < \infty$. As SGD updates occur in $\operatorname{Im}\mathbf{H}$, we again use space decomposition $\mathcal{H} = \operatorname{Im}\mathbf{H} \oplus \ker\mathbf{H}$ and respective decomposition of parameters $\mathbf{w} = \mathbf{w}_\parallel + \mathbf{w}_\perp$ and features $\boldsymbol{\psi}(\mathbf{x}) = \boldsymbol{\psi}(\mathbf{x})_\parallel + \boldsymbol{\psi}(\mathbf{x})_\perp$. As $\boldsymbol{\psi}(\mathbf{x}_j)_\perp = 0, \mathbf{x}_j \in \mathcal{D}$ and $\mathbf{w}_\perp$ do not affect the loss (31) we can restrict ourselves to $\operatorname{Im}\mathbf{H}$ which bring us to finite dimensional case $d = N < \infty$ fully described above.

The case $d = \infty, N < \infty$ is primarily used for kernel regression problems defined by a kernel function $K(\mathbf{x}, \mathbf{x}')$ which can to be, for example, one of the "classical" kernels (e.g. Gaussian) or

Neural Tangent Kernel (NTK) of some infinitely wide neural network. In this case $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of the kernel $K(\mathbf{x}, \mathbf{x}')$, and features $\boldsymbol{\psi}(\mathbf{x})$ are the respective mappings $\boldsymbol{\psi} : \mathcal{X} \to \mathcal{H}$ with the property $K(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle$.

Case 3: $d, N = \infty$**.** We are particularly interested in the case $d = N = \infty$ because in practice N is quite large and, as previously discussed, the eigenvalues λ_k and target expansion coefficients often are distributed according to asymptotic power laws, which all suggests working in an infinite-dimensional setting. A standard approach to rigorously accommodate $N = \infty$, which we sketch now, is based on Mercer's theorem.

We describe the infinite ($N = \infty$) training dataset by a probability measure $\mu(\mathbf{x})$ so that the loss takes the form

$$L(\mathbf{w}) = \frac{1}{2} \int (\langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}) \rangle - y^\star(\mathbf{x}))^2 d\mu(\mathbf{x}) = \frac{1}{2} \| \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}) \rangle - y^\star(\mathbf{x}) \|_{\mathcal{F}}^2, \tag{39}$$

assuming that $\|y^\star(\mathbf{x})\|_{\mathcal{F}}^2 < \infty$ for the loss to be well defined. Here in addition to the Hilbert space of parameters $\mathcal{H} \ni \mathbf{w}$ we introduced the Hilbert space $\mathcal{F}$ of square integrable functions $f(\mathbf{x}) : \|f\|_{\mathcal{F}}^2 \equiv \int f(\mathbf{x})^2 d\mu(\mathbf{x}) < \infty$. The Hessian $\mathbf{H}$, and the counterparts of the kernel and feature matrices $\mathcal{K}$ (37) and $\boldsymbol{\Psi}$ (35) are now operators

$$\mathbf{H} : \mathcal{H} \to \mathcal{H}, \quad \mathbf{H} = \int \boldsymbol{\psi}(\mathbf{x}) \otimes \boldsymbol{\psi}(\mathbf{x}) d\mu(\mathbf{x}) \tag{40}$$

$$\mathbf{K} : \mathcal{F} \to \mathcal{F}, \quad \mathbf{K}g(x) = \int K(\mathbf{x}, \mathbf{x}') g(\mathbf{x}') d\mu(\mathbf{x}') \tag{41}$$

$$\boldsymbol{\Psi} : \mathcal{F} \to \mathcal{H}, \quad \boldsymbol{\Psi}g(x) = \int \boldsymbol{\psi}(\mathbf{x}') g(\mathbf{x}') d\mu(\mathbf{x}') \tag{42}$$

Again, due to our interest in convergence of the loss (31) during SGD (32), we project parameters $\mathbf{w}$ to $\mathcal{H} \ominus \ker \mathbf{H}$ and target function $y^\star(\mathbf{x})$ to $\mathcal{F} \ominus \ker \mathbf{K}$. After this projection the target is reachable in the sense that $y^\star(\mathbf{x}) \in \mathcal{F} \ominus \ker \mathbf{K} = \overline{\operatorname{Im} \boldsymbol{\Psi}^\dagger}$. In particular, this means that an optimum such that $\|\langle \mathbf{w}^\star, \boldsymbol{\psi}(\mathbf{x}) \rangle - y^\star(\mathbf{x})\| = 0$ may not exist, but there is always a sequence $\mathbf{w}_n$ such that $\lim_{n\to\infty} \|\langle \mathbf{w}^\star, \boldsymbol{\psi}(\mathbf{x}) \rangle - y^\star(\mathbf{x})\| = 0$

Now we proceed to eigendecompositions of $\mathbf{H}$ and $\mathbf{K}$. According to the Mercer's theorem, the restriction of the operator $\mathbf{K}$ given by (41) to $\mathcal{F} \ominus \ker \mathbf{K}$ admits an eigendecomposition of the form

$$\mathbf{K}g(\mathbf{x}) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\mathbf{x}) \int \phi_k(\mathbf{x}') g(\mathbf{x}') d\mu(\mathbf{x}') \tag{43}$$

with $\lambda_k > 0$, $\int \phi_k(\mathbf{x}) \phi_l(\mathbf{x}) d\mu(\mathbf{x}) = \delta_{kl}$, and $\{\phi_k(\mathbf{x})\}$ being a complete basis in $\mathcal{F} \ominus \ker \mathbf{K}$. The latter allows to decompose features as

$$\boldsymbol{\psi}(\mathbf{x}) = \sum_k \rho_k \mathbf{u}_k \phi_k(\mathbf{x}) \tag{44}$$

with $\|\mathbf{u}_k\|^2 = 1$. Substituting (44) into (41) gives $\langle \mathbf{u_k}, \mathbf{u}_l \rangle = \delta_{kl}$ and $\rho_k^2 = \lambda_k$. Substituting then (44) into (40) gives

$$\mathbf{H} = \sum_{k=1}^{\infty} \lambda_k \mathbf{u}_k \otimes \mathbf{u}_k, \tag{45}$$

which also makes $\{\mathbf{u}_k\}$ an orthonormal basis in $\mathcal{H} \ominus \ker \mathbf{H}$.

Under mild regularity assumptions, one can show (König, 2013) that the eigenvalues λ_k converge to 0, and we will assume this in the sequel. Note that if $\lambda_k \to 0$, then the range of the operator $\boldsymbol{\Psi}^\dagger$ is not closed so that, exactly as pointed out above, for some $y^*(\mathbf{x}) \in \mathcal{F} \ominus \ker \mathbf{K}$ there are no exact finite-norm minimizers $\mathbf{w}^* \in \mathcal{H}$ satisfying $\langle \mathbf{w}^\star, \boldsymbol{\psi}(\mathbf{x}) \rangle = y^\star(\mathbf{x})$ for μ-almost all $\mathbf{x}$.

B.3 Output and parameter spaces

Recall that we only assume $\int y^\star(\mathbf{x})^2 d\mu(\mathbf{x}) < \infty$ but do not guarantee existence of the optimum $\mathbf{w}^\star$ in the case $N = d = \infty$. This means that considering SGD dynamics directly in the space of

model outputs $f(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\psi}(\mathbf{x}) \rangle$ may be advantageous when $\mathbf{w}^\star$ is not available. Moreover, it will allow to completely bypass construction of parameter space when original problem is formulated in terms of kernel $K(\mathbf{x}, \mathbf{x}')$ (e.g. for an infinitely wide neural network in the NTK regime). First, let us rewrite SGD recursion in the output space by taking the scalar product of (32) with $\boldsymbol{\psi}(\mathbf{x})$:

$$\begin{pmatrix} f_{t+1}(\mathbf{x}) \\ v_{t+1}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} f_t(\mathbf{x}) \\ v_t(\mathbf{x}) \end{pmatrix} + \begin{pmatrix} \alpha_t \mathbf{K}(B_t) y^\star(\mathbf{x}) \\ \alpha_t \mathbf{K}(B_t) y^\star(\mathbf{x}) \end{pmatrix} \tag{46}$$

$$\mathbf{K}(B_t) g(\mathbf{x}) = \frac{1}{b} \sum_{\mathbf{x}_i \in B_t} K(\mathbf{x}, \mathbf{x}_i) g(\mathbf{x}_i) \tag{47}$$

Here $v_t(\mathbf{x}) = \langle \mathbf{v}_t, \boldsymbol{\psi}(\mathbf{x}) \rangle = f_t(\mathbf{x}) - f_{t+1}(\mathbf{x})$. Considering the approximation error $\Delta f(\mathbf{x}) \equiv f(\mathbf{x}) - y^\star(\mathbf{x})$ we get

$$\begin{pmatrix} \Delta f_{t+1}(\mathbf{x}) \\ v_{t+1}(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \Delta f_t(\mathbf{x}) \\ v_t(\mathbf{x}) \end{pmatrix} \tag{48}$$

As (46) is equivalent to (32), we also get that (48) is equivalent to (2) when $\mathbf{w}^\star$ is available. Thus (48) can be considered as the most general and convenient form of writing an SGD iteration.

Similarly to second moments (3) defined for parameter space, we define second moments in the output space:

$$\begin{aligned} C^{\text{out}}(\mathbf{x}, \mathbf{x}') &\equiv \mathbb{E}[\Delta f(\mathbf{x}) \Delta f(\mathbf{x}')] \\ J^{\text{out}}(\mathbf{x}, \mathbf{x}') &\equiv \mathbb{E}[\Delta f(\mathbf{x}) v(\mathbf{x}')] \\ V^{\text{out}}(\mathbf{x}, \mathbf{x}') &\equiv \mathbb{E}[v(\mathbf{x}) v(\mathbf{x}')] \\ \mathbf{M}^{\text{out}}(\mathbf{x}, \mathbf{x}') &\equiv \left(\begin{smallmatrix} C^{\text{out}}(\mathbf{x},\mathbf{x}') & J^{\text{out}}(\mathbf{x},\mathbf{x}') \\ J^{\text{out}}(\mathbf{x}',\mathbf{x}) & V^{\text{out}}(\mathbf{x},\mathbf{x}') \end{smallmatrix} \right), \end{aligned} \tag{49}$$

where the superscript ^{out} indicates the output space (to avoid confusion with the respective moment counterparts defined in the parameter space). The same moments can be written in the eigenbasis $\{\phi(\mathbf{x})\}$ of the operator $\mathbf{K}$ according to the rule

$$G^{\text{out}}_{kl} = \int \phi_k(\mathbf{x}) G^{\text{out}}(\mathbf{x}, \mathbf{x}') \phi_l(\mathbf{x}) d\mu(\mathbf{x}), \tag{50}$$

where G stands for either C,F or V. When $\mathbf{w}^\star$ is available, the moments C^{out}_{kl} in the output space are related to the moments C_{kl} in the parameter space simply by

$$C^{\text{out}}_{kl} = \sqrt{\lambda_k \lambda_l} C_{kl}. \tag{51}$$

In particular,

$$C^{\text{out}}_{kk} = \lambda_k C_{kk} \tag{52}$$

and

$$\sum_k \lambda_k C_{kk} = \sum_k C^{\text{out}}_{kk} = \mathbb{E}[\|\Delta f\|^2] < \infty. \tag{53}$$

Even when $\mathbf{w}^\star$ is not available we can still define C_{kl} by Eq. (51), but in this case $\sum_k C_{kk} = \infty$.

C Dynamics of second moments

The purpose of this section is to prove proposition 1. However, as discussed in section B.3, the description in terms of moments in output space (49) is more widely applicable than description in terms of parameter moments (3). Thus we first prove an analogue of proposition 1 for output space moments. For convinience, let us denote the moments (49) as $\mathbf{C}^{\text{out}}, \mathbf{J}^{\text{out}}, \mathbf{V}^{\text{out}} \in \mathcal{F} \otimes \mathcal{F}$.

Proposition 4. *Consider SGD* (48) *with learning rates α_t, momentum β_t, batch size b and random uniform choice of the batch B_t. Then the update of second moments* (49) *is*

$$\begin{aligned} \begin{pmatrix} \mathbf{C}^{out}_{t+1} & \mathbf{J}^{out}_{t+1} \\ (\mathbf{J}^{out}_{t+1})^\dagger & \mathbf{V}^{out}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K} & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{C}^{out}_{t} & \mathbf{J}^{out}_{t} \\ (\mathbf{J}^{out}_{t})^\dagger & \mathbf{V}^{out}_{t} \end{pmatrix} \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K} & \beta_t \mathbf{I} \end{pmatrix}^T \\ + \gamma \alpha_t^2 \begin{pmatrix} \boldsymbol{\Sigma}^{out}_t & \boldsymbol{\Sigma}^{out}_t \\ \boldsymbol{\Sigma}^{out}_t & \boldsymbol{\Sigma}^{out}_t \end{pmatrix} \end{aligned} \tag{54}$$

The second term represents covariance of function gradients induced mini-batch sampling noise and is given by

1) for finite dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$:

$$\mathbf{\Sigma}_t^{out}(\mathbf{x}, \mathbf{x}') = \frac{1}{N} \sum_{\mathbf{x}_i \in \mathcal{D}} K(\mathbf{x}, \mathbf{x}_i) C_t^{out}(\mathbf{x}_i, \mathbf{x}_i) K(\mathbf{x}_i, \mathbf{x}') - \mathbf{K}\mathbf{C}_t^{out}\mathbf{K}(\mathbf{x}, \mathbf{x}'), \tag{55}$$

and the amplitude $\gamma = \frac{N-b}{(N-1)b}$;

2) for infinite dataset $\mathcal{D}$ with density $d\mu(\mathbf{x})$:

$$\mathbf{\Sigma}_t^{out}(\mathbf{x}, \mathbf{x}') = \int K(\mathbf{x}, \mathbf{x}'') C_t^{out}(\mathbf{x}'', \mathbf{x}'') K(\mathbf{x}'', \mathbf{x}') d\mu(\mathbf{x}'') - \mathbf{K}\mathbf{C}_t^{out}\mathbf{K}(\mathbf{x}, \mathbf{x}'), \tag{56}$$

and the amplitude $\gamma = \frac{1}{b}$.

Proof. First, we take expression of single SGD (48) and use it to express $\boldsymbol{M}_{t+1}^{\text{out}}$ through $\boldsymbol{M}_t^{\text{out}}$.

$$\begin{aligned}
\boldsymbol{M}_{t+1}^{\text{out}}(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\left[\begin{pmatrix} \Delta f_{t+1}(\mathbf{x})\Delta f_{t+1}(\mathbf{x}') & \Delta f_{t+1}(\mathbf{x}) v_{t+1}(\mathbf{x}') \\ v_{t+1}(\mathbf{x})\Delta f_{t+1}(\mathbf{x}') & v_{t+1}(\mathbf{x}) v_{t+1}(\mathbf{x}') \end{pmatrix}\right] \\
&= \mathbb{E}\left[\begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \Delta f_t(\mathbf{x}) \\ v_t(\mathbf{x}) \end{pmatrix} \begin{pmatrix} \Delta f_t(\mathbf{x}') & v_t(\mathbf{x}') \end{pmatrix} \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \end{pmatrix}^T\right] \\
&\overset{(1)}{=} \mathbb{E}\left[\begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{C}_t^{\text{out}} & \mathbf{J}_t^{\text{out}} \\ (\mathbf{J}_t^{\text{out}})^\dagger & \mathbf{V}_t^{\text{out}} \end{pmatrix} \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K}(B_t) & \beta_t \mathbf{I} \end{pmatrix}^T\right] \\
&\overset{(2)}{=} \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K} & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{C}_t^{\text{out}} & \mathbf{J}_t^{\text{out}} \\ (\mathbf{J}_t^{\text{out}})^\dagger & \mathbf{V}_t^{\text{out}} \end{pmatrix} \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{K} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{K} & \beta_t \mathbf{I} \end{pmatrix}^T \\
&+ \begin{pmatrix} \mathbb{E}\Big[\delta\mathbf{K}(B_t)\mathbf{C}^{\text{out}}\delta\mathbf{K}(B_t)\Big] & \mathbb{E}\Big[\delta\mathbf{K}(B_t)\mathbf{C}^{\text{out}}\delta\mathbf{K}(B_t)\Big] \\ \mathbb{E}\Big[\delta\mathbf{K}(B_t)\mathbf{C}^{\text{out}}\delta\mathbf{K}(B_t)\Big] & \mathbb{E}\Big[\delta\mathbf{K}(B_t)\mathbf{C}^{\text{out}}\delta\mathbf{K}(B_t)\Big] \end{pmatrix}
\end{aligned} \tag{57}$$

Here in (1) we used that $\Delta f_t, v_t$ are independent from B_t and therefore the average factorizes into product of averages. In (2) we introduced notation $\delta\mathbf{K}(B_t) = \mathbf{K}(B_t) - \mathbf{K}$ and used that $\mathbb{E}[\delta\mathbf{K}(B_t)] = 0$.

Now we proceed to the calculation of the average from the last line in (57):

$$\gamma \mathbf{\Sigma}_t^{\text{out}} \equiv \mathbb{E}\Big[\delta\mathbf{K}(B_t)\mathbf{C}_t^{\text{out}}\delta\mathbf{K}(B_t)\Big] \tag{58}$$

1) Finite $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$.

Denoting for convenience function values calculated at $\mathbf{x}_i, \mathbf{x}_j$ with subscripts $_{ij}$, and omitting time index t, we get

$$\begin{aligned}
\gamma\left(\mathbf{\Sigma}^{\text{out}}\right)_{ij} =& \mathbb{E}\left[\left(\delta\mathbf{K}(B)\mathbf{C}^{\text{out}}\delta\mathbf{K}(B)\right)_{ij}\right] \\
=& \mathbb{E}\left[\left(\frac{1}{b}\sum_{i'\in B} - \frac{1}{N}\sum_{i'}\right)\left(\frac{1}{b}\sum_{j'\in B} - \frac{1}{N}\sum_{j'}\right) K_{ii'}K_{jj'}C^{\text{out}}_{i'j'}\right] \\
=& \mathbb{E}\left[\left(\frac{1}{b^2}\sum_{i',j'\in B} - \frac{1}{N^2}\sum_{i',j'}\right) K_{ii'}K_{jj'}C^{\text{out}}_{i'j'}\right] \\
\overset{(1)}{=}& \left(\left(\frac{1}{b^2}\frac{b}{N} - \frac{1}{N^2}\right)\sum_{i'=j'} + \left(\frac{1}{|B|^2}\frac{b(b-1)}{N(N-1)} - \frac{1}{N^2}\right)\sum_{i'\neq j'}\right) K_{ii'}K_{jj'}C^{\text{out}}_{i'j'} \\
=& \left(\frac{N-b}{b}\sum_{i'=j'} - \frac{N-b}{b(N-1)}\sum_{i'\neq j'}\right)\frac{1}{N^2}K_{ii'}K_{jj'}C^{\text{out}}_{i'j'} \\
=& \frac{N-b}{(N-1)b}\sum_{i'j'}(N\delta_{i'j'}-1)\frac{1}{N^2}K_{ii'}K_{jj'}C^{\text{out}}_{i'j'}
\end{aligned} \tag{59}$$

Here unspecified sums run over dataset $\mathcal{D}$. In (1) we used that the fraction of batches B which contain two indices $i' \neq j'$ is $\binom{N-2}{|B|-2}/\binom{N}{|B|} = \frac{|B|(|B-1|)}{N(N-1)}$ and the fraction of batches containing index j' is $\binom{N-1}{|B|-1}/\binom{N}{|B|} = \frac{|B|}{N}$. Taking $\gamma = \frac{N-b}{(N-1)b}$ we get (55)

2) Infinite $\mathcal{D}$ with density $d\mu(\mathbf{x})$. Proceeding similarly to (59) we get

$$\begin{aligned}
\mathbf{\Sigma}^{\text{out}}(\mathbf{x},\mathbf{x}') =& \mathbb{E}\left[\mathbf{K}(B)\mathbf{C}^{\text{out}}\mathbf{K}(B)\right](\mathbf{x},\mathbf{x}') - \mathbf{K}\mathbf{C}^{\text{out}}\mathbf{K}(\mathbf{x},\mathbf{x}') \\
=& \frac{b}{b^2}\int K(\mathbf{x},\mathbf{x}'')C^{\text{out}}(\mathbf{x}'',\mathbf{x}'')K(\mathbf{x}'',\mathbf{x}')d\mu(\mathbf{x}'') \\
&+ \frac{b(b-1)}{b^2}\mathbf{K}\mathbf{C}^{\text{out}}\mathbf{K}(\mathbf{x},\mathbf{x}') - \mathbf{K}\mathbf{C}^{\text{out}}\mathbf{K}(\mathbf{x},\mathbf{x}') \\
=& \frac{1}{b}\left(\int K(\mathbf{x},\mathbf{x}'')C^{\text{out}}(\mathbf{x}'',\mathbf{x}'')K(\mathbf{x}'',\mathbf{x}')d\mu(\mathbf{x}'') - \mathbf{K}\mathbf{C}^{\text{out}}\mathbf{K}(\mathbf{x},\mathbf{x}')\right)
\end{aligned} \tag{60}$$

which gives (56) after setting $\gamma = \frac{1}{b}$. □

Let us write second moments dynamics for both parameter space (4) and output space (54) in eigenbasises $\{\mathbf{u}_k\}$ and $\{\phi_k(\mathbf{x})\}$ using decompositions (44) and (43). For parameter space we get

$$\begin{aligned}
\begin{pmatrix} C_{kl,t+1} & J_{kl,t+1} \\ J_{lk,t+1} & V_{kl,t+1}\end{pmatrix} =& \begin{pmatrix} 1-\alpha_t\lambda_k & \beta_t \\ -\alpha_t\lambda_k & \beta_t\end{pmatrix}\begin{pmatrix} C_{kl,t} & J_{kl,t} \\ J_{lk,t} & V_{kl,t}\end{pmatrix}\begin{pmatrix} 1-\alpha_t\lambda_l & \beta_t \\ -\alpha_t\lambda_l & \beta_t\end{pmatrix}^T \\
&+\gamma\alpha_t^2\sqrt{\lambda_k\lambda_l}\left(\sum_{k'l'}\sqrt{\lambda_{k'}\lambda_{l'}}C_{k'l',t}\int\phi_k(\mathbf{x})\phi_l(\mathbf{x})\phi_{k'}(\mathbf{x})\phi_{l'}(\mathbf{x})d\mu(\mathbf{x}) - C_{kl,t}\right)\begin{pmatrix} 1 & 1 \\ 1 & 1\end{pmatrix}
\end{aligned} \tag{61}$$

And for output space

$$\begin{aligned}
\begin{pmatrix} C^{\text{out}}_{kl,t+1} & J^{\text{out}}_{kl,t+1} \\ J^{\text{out}}_{lk,t+1} & V^{\text{out}}_{kl,t+1}\end{pmatrix} =& \begin{pmatrix} 1-\alpha_t\lambda_k & \beta_t \\ -\alpha_t\lambda_k & \beta_t\end{pmatrix}\begin{pmatrix} C^{\text{out}}_{kl,t} & J^{\text{out}}_{kl,t} \\ J^{\text{out}}_{lk,t} & V^{\text{out}}_{kl,t}\end{pmatrix}\begin{pmatrix} 1-\alpha_t\lambda_l & \beta_t \\ -\alpha_t\lambda_l & \beta_t\end{pmatrix}^T \\
&+\gamma\alpha_t^2\lambda_k\lambda_l\left(\sum_{k'l'}C^{\text{out}}_{k'l',t}\int\phi_k(\mathbf{x})\phi_l(\mathbf{x})\phi_{k'}(\mathbf{x})\phi_{l'}(\mathbf{x})d\mu(\mathbf{x}) - C^{\text{out}}_{kl,t}\right)\begin{pmatrix} 1 & 1 \\ 1 & 1\end{pmatrix}
\end{aligned} \tag{62}$$

Now we are ready to prove proposition 1

Proposition 1. *Consider SGD* (1) *with learning rates* α_t*, momentum* β_t*, batch size* b *and random uniform choice of the batch* B_t*. Then the update of second moments* (3) *is*

$$\begin{pmatrix} \mathbf{C}_{t+1} & \mathbf{J}_{t+1} \\ \mathbf{J}_{t+1}^{\dagger} & \mathbf{V}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{C}_t & \mathbf{J}_t \\ \mathbf{J}_t^{\dagger} & \mathbf{V}_t \end{pmatrix} \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix}^T + \gamma \alpha_t^2 \begin{pmatrix} \mathbf{\Sigma}(\mathbf{C}_t) & \mathbf{\Sigma}(\mathbf{C}_t) \\ \mathbf{\Sigma}(\mathbf{C}_t) & \mathbf{\Sigma}(\mathbf{C}_t) \end{pmatrix} \tag{4}$$

Here $\gamma \mathbf{\Sigma}(\mathbf{C}_t)$ *is the covariance of gradient noise due to mini-batch sampling, with* $\gamma = \frac{N-b}{(N-1)b}$ *and*

$$\mathbf{\Sigma}(\mathbf{C}) = \frac{1}{N} \sum_{i=1}^{N} \langle \boldsymbol{\psi}(\mathbf{x}_i), \mathbf{C} \boldsymbol{\psi}(\mathbf{x}_i) \rangle \boldsymbol{\psi}(\mathbf{x}_i) \otimes \boldsymbol{\psi}(\mathbf{x}_i) - \mathbf{HCH}. \tag{5}$$

Proof. Note that in the eigenbasis $\{\phi_k(\mathbf{x})\}$ of $\mathbf{K}$ and eigenbasis $\{\mathbf{u}_k\}$ of $\mathbf{H}$ output and parameter second moments are connected as

$$\sqrt{\lambda_k \lambda_l} \boldsymbol{M}_{kl} = \boldsymbol{M}_{kl}^{\text{out}} \tag{63}$$

Using this connection rule we see that parameter dynamics (61) in eigenbasis of $\mathbf{H}$ is equivalent to output space second moments dynamics (62) in the eigenbasis of $\mathbf{K}$. Finally, (62) is equivalent to (54) proved in 4. □

For completeness, let us also write a formula of SE approximation in output space:

$$\begin{aligned} \begin{pmatrix} C_{kk,n+1}^{\text{out}} & J_{kk,n+1}^{\text{out}} \\ J_{kk,n+1}^{\text{out}} & V_{kk,n+1}^{\text{out}} \end{pmatrix} = & \begin{pmatrix} 1 - \alpha_t \lambda_k & \beta_t \\ -\alpha_t \lambda_k & \beta_t \end{pmatrix} \begin{pmatrix} C_{kk,t}^{\text{out}} & J_{kk,t}^{\text{out}} \\ J_{kk,t}^{\text{out}} & V_{kk,t}^{\text{out}} \end{pmatrix} \begin{pmatrix} 1 - \alpha_t \lambda_k & \beta_t \\ -\alpha_t \lambda_k & \beta_t \end{pmatrix}^T \\ & + \gamma (\alpha_t \lambda_k)^2 \left(\tau_1 \sum_l C_{ll,t}^{\text{out}} - \tau_2 C_{kk,t} \right) \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \end{aligned} \tag{64}$$

D SE APPROXIMATION

D.1 DYNAMICS OF SPECTRAL MEASURES

In the section 2 we mention that SE approximation allows to reconstruct the trajectories of observables $\mathrm{Tr}[\phi(\mathbf{H})\mathbf{C}_t]$ from initial state $\mathbf{C}_0$. The way to do it is, of course, through considering dynamics of spectral measures, which we explicitly write below. Denote, for simplicity, $\rho_{\mathbf{J}_t} \equiv \rho_{\frac{1}{2}(\mathbf{J}_t + \mathbf{J}_t^{\dagger})}$.

Proposition 5. *Assume SE approximation* (7) *holds for all* $\mathbf{C}_t$ *during SGD dynamics* (4)*. Then spectral measures of all second moment matrices* $\rho_{\mathbf{C}_t}, \rho_{\mathbf{J}_t}, \rho_{\mathbf{V}_t}$ *can be written as*

$$\begin{pmatrix} \rho_{\mathbf{C}_t}(d\lambda) \\ \rho_{\mathbf{J}_t}(d\lambda) \\ \rho_{\mathbf{V}_t}(d\lambda) \end{pmatrix} = \mathbf{G}_t(\lambda) \begin{pmatrix} \rho_{\mathbf{C}_0}(d\lambda) \\ \rho_{\mathbf{J}_0}(d\lambda) \\ \rho_{\mathbf{V}_0}(d\lambda) \end{pmatrix} + \mathbf{s}_t(\lambda) \rho_{\mathbf{H}}(d\lambda) \tag{65}$$

where $\mathbf{G}_t(\lambda)$ *is* λ *dependent* 3×3 *matrix, and* $\mathbf{s}_t(\lambda)$ *is* λ *dependent* 3*-dimensional vector.*

During SGD iterations $\mathbf{G}_t(\lambda)$ *and* $\mathbf{s}_t(\lambda)$ *are transformed as*

$$\mathbf{G}_{t+1}(\lambda) = \begin{pmatrix} (1 - \alpha_t \lambda)^2 - \gamma \tau_2 (\alpha_t \lambda)^2 & 2\beta_t (1 - \alpha_t \lambda) & \beta_t^2 \\ -\alpha_t \lambda (1 - \alpha_t \lambda) - \gamma \tau_2 (\alpha_t \lambda)^2 & 1 - 2\beta_t \alpha_t \lambda & \beta_t^2 \\ (\alpha_t \lambda)^2 - \gamma \tau_2 (\alpha_t \lambda)^2 & -2\beta_t \alpha_t \lambda & \beta_t^2 \end{pmatrix} \mathbf{G}_t(\lambda) \tag{66}$$

$$\begin{aligned} \mathbf{s}_{t+1}(\lambda) = & \begin{pmatrix} (1 - \alpha_t \lambda)^2 - \gamma \tau_2 (\alpha_t \lambda)^2 & 2\beta_t (1 - \alpha_t \lambda) & \beta_t^2 \\ -\alpha_t \lambda (1 - \alpha_t \lambda) - \gamma \tau_2 (\alpha_t \lambda)^2 & 1 - 2\beta_t \alpha_t \lambda & \beta_t^2 \\ (\alpha_t \lambda)^2 - \gamma \tau_2 (\alpha_t \lambda)^2 & -2\beta_t \alpha_t \lambda & \beta_t^2 \end{pmatrix} \mathbf{s}_t(\lambda) \\ & + \gamma \tau_1 \alpha_t^2 \int \lambda' s_{t,1}(\lambda) \rho_{\mathbf{H}}(d\lambda) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ & + \gamma \tau_1 \alpha_t^2 \int \lambda' \Big[G_{t,11}(\lambda) \rho_{\mathbf{C}_0}(d\lambda') + G_{t,12}(\lambda) \rho_{\mathbf{J}_0}(d\lambda') G_{t,13}(\lambda) \rho_{\mathbf{V}_0}(d\lambda') \Big] \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \end{aligned} \tag{67}$$

Proof. Let us rewrite dynamics of second moments (4) as linear transformation of on $(\mathbf{C}_t, \mathbf{J}_t, \mathbf{V}_t)$ and then take traces $\mathrm{Tr}[\phi(\mathbf{H})\#]$ of both sides with arbitrary test function $\phi(\lambda)$. Since arbitrariness of $\phi(\lambda)$ in the obtained relation implies the same relation but for spectral measures instead of traces

$$\begin{pmatrix} \rho_{\mathbf{C}_{t+1}}(d\lambda) \\ \rho_{\mathbf{J}_{t+1}}(d\lambda) \\ \rho_{\mathbf{V}_{t+1}}(d\lambda) \end{pmatrix} = \begin{pmatrix} (1-\alpha_t\lambda)^2 & 2\beta_t(1-\alpha_t\lambda) & \beta_t^2 \\ -\alpha_t\lambda(1-\alpha_t\lambda) & 1-2\beta_t\alpha_t\lambda & \beta_t^2 \\ (\alpha_t\lambda)^2 & -2\beta_t\alpha_t\lambda & \beta_t^2 \end{pmatrix} \begin{pmatrix} \rho_{\mathbf{C}_t}(d\lambda) \\ \rho_{\mathbf{J}_t}(d\lambda) \\ \rho_{\mathbf{V}_t}(d\lambda) \end{pmatrix} + \gamma\alpha_t^2 \begin{pmatrix} \rho_{\mathbf{\Sigma}_t} \\ \rho_{\mathbf{\Sigma}_t} \\ \rho_{\mathbf{\Sigma}_t} \end{pmatrix} \tag{68}$$

Now we use SE approximation (7) to express $\rho_{\mathbf{\Sigma}_t}$ through $\rho_{\mathbf{C}_t}$ and $\rho_{\mathbf{H}}$. The result is

$$\begin{aligned} \begin{pmatrix} \rho_{\mathbf{C}_{t+1}}(d\lambda) \\ \rho_{\mathbf{J}_{t+1}}(d\lambda) \\ \rho_{\mathbf{V}_{t+1}}(d\lambda) \end{pmatrix} &= \begin{pmatrix} (1-\alpha_t\lambda)^2 - \gamma\tau_2(\alpha_t\lambda)^2 & 2\beta_t(1-\alpha_t\lambda) & \beta_t^2 \\ -\alpha_t\lambda(1-\alpha_t\lambda) - \gamma\tau_2(\alpha_t\lambda)^2 & 1-2\beta_t\alpha_t\lambda & \beta_t^2 \\ (\alpha_t\lambda)^2 - \gamma\tau_2(\alpha_t\lambda)^2 & -2\beta_t\alpha_t\lambda & \beta_t^2 \end{pmatrix} \begin{pmatrix} \rho_{\mathbf{C}_t}(d\lambda) \\ \rho_{\mathbf{J}_t}(d\lambda) \\ \rho_{\mathbf{V}_t}(d\lambda) \end{pmatrix} \\ &+ \gamma\tau_1\alpha_t^2 \int \lambda' \rho_{\mathbf{C}_t}(d\lambda') \begin{pmatrix} \rho_{\mathbf{H}}(d\lambda) \\ \rho_{\mathbf{H}}(d\lambda) \\ \rho_{\mathbf{H}}(d\lambda) \end{pmatrix} \end{aligned} \tag{69}$$

Now we use obtained form of recursion of spectral measures prove representation (65) and transition formulas (66),(67). We proceed by induction. At $t = 0$ representation (65) indeed holds with $\mathbf{G}_0(\lambda) = \mathbf{I}$ and $\mathbf{s}_0(\lambda) = 0$. Now assume that (65) holds for iteration t and substitute it into (69). The result is

$$\begin{aligned} \begin{pmatrix} \rho_{\mathbf{C}_{t+1}}(d\lambda) \\ \rho_{\mathbf{J}_{t+1}}(d\lambda) \\ \rho_{\mathbf{V}_{t+1}}(d\lambda) \end{pmatrix} &= \begin{pmatrix} (1-\alpha_t\lambda)^2 - \gamma\tau_2(\alpha_t\lambda)^2 & 2\beta_t(1-\alpha_t\lambda) & \beta_t^2 \\ -\alpha_t\lambda(1-\alpha_t\lambda) - \gamma\tau_2(\alpha_t\lambda)^2 & 1-2\beta_t\alpha_t\lambda & \beta_t^2 \\ (\alpha_t\lambda)^2 - \gamma\tau_2(\alpha_t\lambda)^2 & -2\beta_t\alpha_t\lambda & \beta_t^2 \end{pmatrix} \mathbf{G}_t(\lambda) \begin{pmatrix} \rho_{\mathbf{C}_0}(d\lambda) \\ \rho_{\mathbf{J}_0}(d\lambda) \\ \rho_{\mathbf{V}_0}(d\lambda) \end{pmatrix} \\ &+ \begin{pmatrix} (1-\alpha_t\lambda)^2 - \gamma\tau_2(\alpha_t\lambda)^2 & 2\beta_t(1-\alpha_t\lambda) & \beta_t^2 \\ -\alpha_t\lambda(1-\alpha_t\lambda) - \gamma\tau_2(\alpha_t\lambda)^2 & 1-2\beta_t\alpha_t\lambda & \beta_t^2 \\ (\alpha_t\lambda)^2 - \gamma\tau_2(\alpha_t\lambda)^2 & -2\beta_t\alpha_t\lambda & \beta_t^2 \end{pmatrix} \mathbf{s}_t(\lambda)\rho_{\mathbf{H}}(d\lambda) \\ &+ \gamma\tau_1\alpha_t^2 \int \lambda' \Big[G_{t,11}(\lambda)\rho_{\mathbf{C}_0}(d\lambda') + G_{t,12}(\lambda)\rho_{\mathbf{J}_0}(d\lambda') G_{t,13}(\lambda)\rho_{\mathbf{V}_0}(d\lambda') \Big] \begin{pmatrix} \rho_{\mathbf{H}}(d\lambda) \\ \rho_{\mathbf{H}}(d\lambda) \\ \rho_{\mathbf{H}}(d\lambda) \end{pmatrix} \\ &+ \gamma\tau_1\alpha_t^2 \int \lambda' s_{t,1}(\lambda')\rho_{\mathbf{H}}(d\lambda') \begin{pmatrix} \rho_{\mathbf{H}}(d\lambda) \\ \rho_{\mathbf{H}}(d\lambda) \\ \rho_{\mathbf{H}}(d\lambda) \end{pmatrix} \end{aligned} \tag{70}$$

Now we simply observe that right-hand side of (70) has exactly the form (65) with $\mathbf{G}_{t+1}(\lambda), \mathbf{s}_{t+1}(\lambda)$ expressed through $\mathbf{G}_t(\lambda), \mathbf{s}_t(\lambda)$ according to (66),(67).

□

Note that updates of matrix $\mathbf{G}_t(\lambda)$ and vector $\mathbf{s}_t(\lambda)$ in representation (65) depend only on learning algorithm parameters α_t, β_t, combinations $\gamma\tau_1, \gamma\tau_2$, and non-zero eigenvalues of Hessian $\mathbf{H}$. Also, we stress that although in this proposition we considered all possible initial conditions with $\mathbf{J}_0 \neq 0, \mathbf{V} \neq 0$, in all other parts of the paper we focus on the case $\mathbf{J}_0 = \mathbf{V} = 0$.

D.2 Theorem 1

In this section we provide full version of theorem 1 and its proof, and also elaborate on how spectral measures $\rho_{\mathbf{C}_t}$ change during training. We keep the notations from previous sections, in particular the model Jacobian is $\mathbf{\Psi}$, the Hessian $\mathbf{H} = \frac{1}{N}\mathbf{\Psi}\mathbf{\Psi}^T$, the kernel matrix $\mathbf{K} = \frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi}$, and output space second moments $\mathbf{C}^{\text{out}} = \mathbf{\Psi}^T\mathbf{C}\mathbf{\Psi}$. However, for simplicity, in this section we consider only finite training datasets $N < \infty$ and features dimension $d \geq N$ (including $d = \infty$). The extension of the results to all values of d, N seems to be rather technical and we expect it to lead to effectively the same conclusions.

We start with an analogue of theorem 1 but for the isolated noise term (8) containing the tensor $\widehat{R}$. It turns out that the SE approximation is closely related to what we call the *SE family* of tensors $\widehat{R}$:

$$R_{i_1i_2i_3i_4} = \tau_1\delta_{i_1i_2}\delta_{i_3i_4} - \frac{\tau_2 - 1}{2}(\delta_{i_1i_3}\delta_{i_2i_4} + \delta_{i_1i_4}\delta_{i_2i_3}) \tag{71}$$

It is easy to check that the respective $\mathbf{\Sigma}(\widehat{R})$ satisfies not only (7), but also a stronger version

$$\mathbf{\Sigma}_{\widehat{R}}(\mathbf{C}) = \tau_1 \mathbf{H}\,\mathrm{Tr}[\mathbf{HC}] - (\tau_2)\mathbf{HCH} \tag{72}$$

Let's focus only on the $\widehat{R}$-dependent term in (8), and for convenience rewrite it in terms $\mathbf{\Psi}$ as

$$\mathbf{\Sigma}'_{\widehat{R}}(\mathbf{C}) = \frac{1}{N^2} \sum_{i_1,i_2,i_3,i_4=1}^{N} R_{i_1 i_2 i_3 i_4} \langle \mathbf{\Psi}_{i_3}, \mathbf{C}\mathbf{\Psi}_{i_4} \rangle \mathbf{\Psi}_{i_1} \otimes \mathbf{\Psi}_{i_2}, \tag{73}$$

where $\mathbf{\Psi}_i = \boldsymbol{\psi}(\mathbf{x}_i)$ denotes the i'th column of $\mathbf{\Psi}$. Then we have

Proposition 6. *Consider a fixed tensor $\widehat{R}$, and let operators $\mathbf{\Sigma}'$, $\mathbf{C}$ be related to each over as $\mathbf{\Sigma}' = \mathbf{\Sigma}'_{\widehat{R}}(\mathbf{C})$. Then, the following 3 statements are equivalent:*

1. *For any Jacobian $\mathbf{\Psi}$ and initial state $\mathbf{C}$, spectral measure $\rho_{\mathbf{\Sigma}'}$ depends on $\mathbf{C}$ only through its spectral measure $\rho_{\mathbf{C}}$, and on $\mathbf{\Psi}$ only through non-zero eigenvalues λ_k of Hessian $\mathbf{H}$.*
2. *Tensor $\hat{R}$ belongs to SE family* (71).
3. *For any Jacobian $\mathbf{\Psi}$ and initial state $\mathbf{C}$, the operator $\mathbf{\Sigma}'$ does not change by the replacement of $\mathbf{\Psi}$ with $\widetilde{\mathbf{\Psi}} = \mathbf{\Psi}\mathbf{U}^T$ in* (73) *for any orthogonal matrix $\mathbf{U} \in \mathrm{O}(N)$.*

Proof. **From (2) to (3).** Substituting SE form (71) of tensor $\widehat{R}$ into (73) we get

$$\mathbf{\Sigma}' = \tau_1 \mathbf{H}\,\mathrm{Tr}[\mathbf{HC}] + \tau_2 \mathbf{HCH} \tag{74}$$

Then statement **(3)** follows from invariance of Hessian under orthogonal transformations of the Jacobian

$$\widetilde{\mathbf{H}} = \frac{1}{N}\widetilde{\mathbf{\Psi}}\widetilde{\mathbf{\Psi}}^T = \frac{1}{N}\mathbf{\Psi}\mathbf{U}^T\mathbf{U}\mathbf{\Psi}^T = \mathbf{H} \tag{75}$$

From (3) to (2). Denote $\widetilde{\mathbf{\Sigma}}'$ the result of (73) but rotated Jacobian $\widetilde{\mathbf{\Psi}} = \mathbf{\Psi}\mathbf{U}^T$ in the right-hand side. It is convenient to represent $\widetilde{\mathbf{\Sigma}}'$ as

$$\widetilde{\mathbf{\Sigma}}' = \frac{1}{N^2} \sum_{i_1,i_2,i_3,i_4=1}^{N} \widetilde{R}_{i_1 i_2 i_3 i_4} C^{\text{out}}_{i_3 i_4} \mathbf{\Psi}_{i_1} \otimes \mathbf{\Psi}_{i_2} \tag{76}$$

where $\widehat{\widetilde{R}}$ is the rotated tensor $\widehat{R}$ with coordinates $\widetilde{R}_{i_1 i_2 i_3 i_4} = U_{j_1 i_1} U_{j_2 i_2} U_{j_3 i_3} U_{j_4 i_4} R_{j_1 j_2 j_3 j_4}$. As statement **(3)** implies $\widetilde{\mathbf{\Sigma}}' = \mathbf{\Sigma}'$, we get

$$0 = \widetilde{\mathbf{\Sigma}}' - \mathbf{\Sigma}' = \frac{1}{N^2} \sum_{i_1,i_2,i_3,i_4=1}^{N} \delta R_{i_1 i_2 i_3 i_4} C^{\text{out}}_{i_3 i_4} \mathbf{\Psi}_{i_1} \otimes \mathbf{\Psi}_{i_2} \tag{77}$$

where the tensor $\delta\widehat{R} = \widehat{\widetilde{R}} - \widehat{R}$ is the difference between original and "rotated" tensor $\widehat{R}$. Note that $\widehat{R}$, $\widehat{\widetilde{R}}$ and therefore $\delta\widehat{R} = \widehat{\widetilde{R}} - \widehat{R}$ are symmetric w.r.t. permutation of the first two and the last two indices. Taking Jacobian with the full rank $\mathrm{rank}(\mathbf{\Psi}) = N$ and exploiting permutation symmetry $i_1 \leftrightarrow i_2$ of $\delta\widehat{R}$ we notice that (77) implies

$$\sum_{i_3,i_4=1}^{N} \delta R_{i_1 i_2 i_3 i_4} C^{\text{out}}_{i_3 i_4} = 0 \tag{78}$$

Next, we observe that for full rank $\mathbf{\Psi}$ we can choose $\mathbf{C}$ such that $\mathbf{C}^{\text{out}}$ is an arbitrary positive-definite matrix. Then, permutation symmetry $i_3 \leftrightarrow i_4$ and arbitrariness of $\mathbf{C}^{\text{out}}$ implies $\delta\widehat{R} = 0$. As $\mathbf{U}$ was arbitrary, we see that

$$U_{j_1 1_1} U_{j_2 i_2} U_{j_3 i_3} U_{j_4 i_4} R_{j_1 j_2 j_3 j_4} = R_{i_1 i_2 i_3 i_4}, \quad \forall \mathbf{U} \in \mathrm{O}(N). \tag{79}$$

According to Weyl Weyl (1939) (see Jeffreys (1973) for compact reference), the equality above implies that $\widehat{R}$ is isotropic - expressed as a sum of products of dirac deltas $\delta_{ii'}$ and possibly one

anti-symmetric symbol $\epsilon_{j_1 j_2 \dots j_N}$. In our case we can not have ϵ tensor: for $N > 4$ it has higher rank $N > \mathrm{rank}(\hat{R})$; for $N = 4$ it does not satisfy symmetries of $\widehat{R}$; for $N = 3$ it cannot be present because 3 is odd; for $N = 2$ all three possible combinations $\delta_{i_1 i_2}\epsilon_{i_3 i_4}, \delta_{i_3 i_4}\epsilon_{i_1 i_2}, \delta_{i_1 i_4}\epsilon_{i_2 i_3}$ again do not satisfy permutation symmetries $i_1 \leftrightarrow i_2$ and $i_3 \leftrightarrow i_4$. Then we left with 3 possible combinations of Dirac deltas

$$\mathbf{R}_{i_1 i_2 i_3 i_4} = c_1 \delta_{i_1 i_2}\delta_{i_3 i_4} + c_2 \delta_{i_1 i_3}\delta_{i_2 i_4} + c_3 \delta_{i_1 i_4}\delta_{i_2 i_3} \tag{80}$$

Again, from symmetry considerations we get $c_2 = c_3$, and finally recover (71) by setting $c_1 = \tau_1$ and $c_2 = c_3 = \frac{1}{2}\tau_2$.

From (2) to (1). We take arbitrary test function $\phi(\lambda)$ and calculate its average with spectral measure $\rho_{\mathbf{\Sigma}'}$

$$\begin{aligned}
\int \phi(\lambda)\rho_{\mathbf{\Sigma}'}(d\lambda) &= \mathrm{Tr}[\phi(\mathbf{H})(\tau_1 \mathbf{H}\,\mathrm{Tr}[\mathbf{HC}] + \tau_2 \mathbf{HCH})] \\
&= \tau_2\,\mathrm{Tr}[\phi(\mathbf{H})\mathbf{H}^2\mathbf{C}] + \tau_1\,\mathrm{Tr}[\mathbf{CH}]\,\mathrm{Tr}[\phi(\mathbf{H})\mathbf{H}] \\
&= \int \phi(\lambda)\tau_2\lambda^2\rho_{\mathbf{C}}(d\lambda) + \tau_1\,\mathrm{Tr}[\phi(\mathbf{H})\mathbf{H}]\int \lambda\rho_{\mathbf{C}}(d\lambda)
\end{aligned} \tag{81}$$

As $\phi(\lambda)$ is arbitrary, we obtain

$$\rho_{\mathbf{\Sigma}'}(d\lambda) = \tau_2\lambda^2\rho_{\mathbf{C}}(d\lambda) + \Big(\tau_1 \int \lambda'\rho_{\mathbf{C}}(d\lambda')\Big)\rho_{\mathbf{H}}(d\lambda) \tag{82}$$

As the right-hand side depends only on the initial spectral measure $\rho_{\mathbf{C}}(d\lambda)$ and the Hessian spectral measure $\rho_{\mathbf{H}}(d\lambda)$, statement **(1)** follows from the fact that $\rho_{\mathbf{H}}(d\lambda)$ depends only on non-zero eigenvalues of $\mathbf{H}$.

From (1) to (2). We take arbitrary $\mathbf{U} \in \mathrm{O}(N)$ and follow the notation of the **From (3) to (2)** proof. As we already noted, original and rotated Jacobians have the same hessian $\widetilde{\mathbf{H}} = \mathbf{H}$, and therefore eigenvalues λ_k. Then, as we do not change $\mathbf{C}$ with the rotation of Jacobian, statement **(3)** implies that spectral measure of $\mathbf{\Sigma}'$ and $\widetilde{\mathbf{\Sigma}}'$ are the same: $\rho_{\mathbf{\Sigma}'} = \rho_{\widetilde{\mathbf{\Sigma}}'}$. For an arbitrary test function $\phi(\lambda)$ we have

$$\begin{aligned}
0 &= \int \phi(\lambda)(\rho_{\widetilde{\mathbf{\Sigma}}'}(d\lambda) - \rho_{\mathbf{\Sigma}'}(d\lambda)) \\
&= \frac{1}{N^2}\,\mathrm{Tr}[\phi(\mathbf{H})\sum_{i_1 i_2}(\delta\widehat{R}\mathbf{C}^{\text{out}})_{i_1 i_2}\mathbf{\Psi}_{i_1}\otimes\mathbf{\Psi}_{i_2}] \\
&= \frac{1}{N^2}\sum_{i_1 i_2}\mathbf{\Psi}_{i_1}^T\phi(\mathbf{H})\mathbf{\Psi}_{i_2}(\delta\widehat{R}\mathbf{C}^{\text{out}})_{i_1 i_2} \\
&= \mathrm{Tr}\left[\phi(\mathbf{K})\mathbf{K}(\delta\widehat{R}\mathbf{C}^{\text{out}})\right] = 0
\end{aligned} \tag{83}$$

Let us now fix some arbitrary output space matrix $\mathbf{C}_*^{\text{out}}$. As $\mathbf{\Psi}$ and $\mathbf{C}$ were arbitrary, we can choose them so that we get arbitrary $\mathbf{K}$ but $\mathbf{C}^{\text{out}} = \mathbf{C}_*^{\text{out}}$. It means that in the last line in (83) we can consider $\mathbf{C}^{\text{out}}$ to be fixed, but $\mathbf{K}, \phi$ to be arbitrary. Then (83) implies that $\delta\widehat{R}\mathbf{C}^{\text{out}} = 0$ and we can repeat respective steps of **From (3) to (2)** proof to get the statement **(2)**. □

With proposition 6 as a basis, we are ready to state and prove the full version of theorem 1.

Theorem 3. *Consider SGD dynamics* (4) *with noise term* (8)*, initial condition* $\mathbf{v}_0 = 0$ *and* $\gamma, \alpha_0 > 0$*. If tensor* $\widehat{R}$ *is fixed, the following statements are equivalent:*

1. *For any Jacobian* $\mathbf{\Psi}$ *and initial state* $\mathbf{C}_0$*, spectral measures* $\rho_{\mathbf{C}_t}$ *occurring during SGD are uniquely determined by the initial spectral measure* $\rho_{\mathbf{C}_0}$*, the eigenvalues* λ_k *of Hessian* $\mathbf{H}$*, and SGD parameters* $\{\alpha_t, \beta_t\}_{t=0}^{\infty}$*.*

2. *Tensor* $\widehat{R}$ *is given by* $R_{i_1 i_2 i_3 i_4} = \tau_1\delta_{i_1 i_2}\delta_{i_3 i_4} - \frac{\tau_2 - 1}{2}(\delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3})$ *for some* τ_1, τ_2*, and Eq.* (7) *holds.*

3. *For any Jacobian $\mathbf{\Psi}$ and initial state $\mathbf{C}_0$, the trajectory $\{\mathbf{C}_t\}_{t=0}^{\infty}$ is invariant under transformations $\widetilde{\mathbf{\Psi}} = \mathbf{\Psi}\mathbf{U}^T$ of the feature matrix $\mathbf{\Psi}$ by any $\mathbf{U} \in \mathrm{O}(N)$.*

Proof. **From (1) to (2).** Let us consider the first iteration of SGD. The initial condition $\mathbf{v}_0 = 0$ implies $\mathbf{J}_0 = \mathbf{V}_0 = 0$. Also, denote for simplicity $\mathbf{\Sigma}_t = \mathbf{\Sigma}_{\widetilde{R}}(\mathbf{C}_t)$. Then, according to (4) we have

$$\mathbf{C}_1 = (\mathbf{I} - \alpha_0\mathbf{H})\mathbf{C}_0(\mathbf{I} - \alpha_0\mathbf{H}) + \gamma\alpha_0^2\mathbf{\Sigma}_0 \tag{84}$$

Respective spectral measure is

$$\rho_{\mathbf{C}_1} = (1-\alpha_0\lambda)^2\rho_{\mathbf{C}_0} + \gamma\alpha_0^2\rho_{\mathbf{\Sigma}_0} = \Big((1-\alpha_0\lambda)^2 - \gamma\alpha_0^2\lambda^2\Big)\rho_{\mathbf{C}_0} + \gamma\alpha_0^2\rho_{\mathbf{\Sigma}_0'} \tag{85}$$

Then, according to statement **(1)**, we get that $\rho_{\mathbf{\Sigma}_0'}$ is uniquely determined by $\rho_{\mathbf{C}_0}$ and eigenvalues of $\mathbf{H}$, and therefore $\widehat{R}$ is given by (71) according to equivalence of statements **(1)** and **(2)** of proposition 6.

From (3) to (2). Considering first iteration of SGD as in the previous step, we get that $\mathbf{\Sigma}_0'$ is invariant under orthogonal transformations of Jacobian $\mathbf{\Psi}$. Then, statement **(2)** follows from equivalence of statements **(3)** and **(2)** of proposition 6.

From (2) to (3). We proceed by induction. First, notice that $\mathbf{C}_0, \mathbf{J}_0, \mathbf{V}_0$ are (trivially) invariant under orthogonal transformations of Jacobian $\widetilde{\mathbf{\Psi}} = \mathbf{\Psi}$.

To make an induction step, assume that $\mathbf{C}_t, \mathbf{J}_t, \mathbf{V}_t$ are invariant under orthogonal transformations of Jacobian. Then, according to equivalence of statements **(3)** and **(2)** of proposition 6 and invariance of Hessian $\mathbf{H}$ under orthogonal transformations of Jacobian $\mathbf{\Psi}$ we get that $\mathbf{\Sigma}_t$ is also invariant. As first term in (4) depend on $\mathbf{\Psi}$ only through $\mathbf{H}$ we get that the whole right-hand side of (4) is invariant under orthogonal transformations of Jacobian, and so are $\mathbf{C}_{t+1}, \mathbf{J}_{t+1}, \mathbf{V}_{t+1}$.

From (2) to (1) As noted in the statement **(2)**, SE approximation (7) is satisfied. It means that we can apply proposition 5 with initial conditions $\rho_{\mathbf{J}_0} = \rho_{\mathbf{V}_0} = 0$. Then, according to representation (65) and update rules (66),(67), spectral measures $\rho_{\mathbf{C}_t}$ are uniquely determined by the initial spectral measure $\rho_{\mathbf{C}_0}$, the eigenvalues λ_k of Hessian $\mathbf{H}$, SGD parameters $\{\alpha_t, \beta_t\}_{t=0}^{\infty}$, and parameters γ, τ_1, τ_2. Recalling that γ, τ_1, τ_2 are considered to be fixed in statement **(1)** finishes the proof. □

D.3 Translation invariant models

The purpose of this section is to prove Proposition 2.

Proposition 2. *Consider a problem with regular grid dataset $\mathcal{D} = \{\mathbf{x_i} = (\frac{2\pi i_1}{N_1}, \dots, \frac{2\pi i_d}{N_d})\}, \quad \mathbf{i} \in (\mathbb{Z}/N_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/N_d\mathbb{Z})$ on the d-dimensional torus $\mathbb{T}^d = (\mathbb{R}/2\pi\mathbb{Z})^d$ and with translation invariant kernel $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x} - \mathbf{x}')$. Then (7) holds exactly with $\tau_1 = 1, \tau_2 = 1$.*

Proof. Statement (7) with $\tau_1 = 1, \tau_2 = 1$ is equivalent to

$$\mathrm{Tr}(\phi(\mathbf{H})\mathbf{\Sigma}) = \mathrm{Tr}(\mathbf{HC})\,\mathrm{Tr}(\mathbf{H}\phi(\mathbf{H})) - \mathrm{Tr}(\mathbf{H}^2\phi(\mathbf{H})\mathbf{C}), \quad \forall\phi. \tag{86}$$

We will be proving this latter statement. Denote the grid-indexing set $(\mathbb{Z}/N_1\mathbb{Z}) \times \cdots \times (\mathbb{Z}/N_d\mathbb{Z})$ by $\mathbb{Z}/\mathbf{N}\mathbb{Z}$. Thanks to translation invariance of the data set and kernel, the eigenvectors $\mathbf{u}_k$ of $\mathbf{H}$ have an explicit representation in the $\mathbf{x}$-domain in terms of Fourier modes. Specifically, let $\mathbf{K}$ be the $N \times N$ kernel matrix with matrix elements

$$K_{\mathbf{ij}} = K_{\mathbf{i-j}} = K_{\mathbf{j-i}} = \frac{1}{N}\langle\psi(\mathbf{x_i}), \psi(\mathbf{x_j})\rangle, \quad \mathbf{i}, \mathbf{j} \in \mathbb{Z}/\mathbf{N}\mathbb{Z}. \tag{87}$$

Consider the functions $\phi_{\mathbf{k}} : \mathcal{D} \to \mathbb{C}$ given by

$$\phi_{\mathbf{k}}(\mathbf{x_i}) = N^{-1/2}e^{\sqrt{-1}\mathbf{k}\cdot\mathbf{x_i}} \equiv \phi_{\mathbf{ki}}, \quad \mathbf{k} \in \mathbb{Z}/\mathbf{N}\mathbb{Z}. \tag{88}$$

These functions form an orthonormal basis in $\mathbb{C}^{\mathcal{D}}$ and diagonalize the matrix $\mathbf{K}$:

$$K_{\mathbf{ij}} = \sum_{\mathbf{k}\in\mathbb{Z}/\mathbf{N}\mathbb{Z}} \lambda_{\mathbf{k}}\phi_{\mathbf{ki}}\overline{\phi}_{\mathbf{kj}} \tag{89}$$

with

$$\lambda_{\mathbf{k}} = \sum_{\mathbf{i}\in\mathbb{Z}/\mathbf{N}\mathbb{Z}} K_{\mathbf{i}}\phi_{\mathbf{ki}}. \tag{90}$$

Operator $\mathbf{H}$ is unitarily equivalent to $\mathbf{K}$ up to its nullspace. Assuming that the Hilbert space of $\mathbf{H}$ is complexified, the respective normalized eigenvectors $\mathbf{u_k}$ of $\mathbf{H}$ can be written as

$$\mathbf{u_k} = (N\lambda_{\mathbf{k}})^{-1/2} \sum_{\mathbf{i}\in\mathbb{Z}/\mathbf{N}\mathbb{Z}} \psi(\mathbf{x_i})\phi_{\mathbf{ki}}, \quad \mathbf{H}\mathbf{u_k} = \lambda_{\mathbf{k}}\mathbf{u_k}. \tag{91}$$

Assume for the moment that the problem is nondegenerate in the sense that all $\lambda_{\mathbf{k}} > 0$ so that all the vectors $\mathbf{u_k}$ are well-defined. It follows in particular that the indices k of the eigenvalues λ_k can be identified with $\mathbf{k} \in \mathbb{Z}/\mathbf{N}\mathbb{Z}$. By inverting relation (91),

$$\psi(\mathbf{x_i}) = \sum_{\mathbf{k}\in\mathbb{Z}/\mathbf{N}\mathbb{Z}} (N\lambda_{\mathbf{k}})^{1/2}\mathbf{u_k}\overline{\phi}_{\mathbf{ki}}. \tag{92}$$

We can then find the diagonal elements of $\mathbf{\Sigma}$ using Eq. (5) and Eqs. (88), (92):

$$\begin{aligned}
\Sigma_{\mathbf{kk}} &= \Big(\frac{1}{N}\sum_{i=1}^{N}\langle\boldsymbol{\psi}(\mathbf{x}_i), \mathbf{C}\boldsymbol{\psi}(\mathbf{x}_i)\rangle\boldsymbol{\psi}(\mathbf{x}_i)\otimes\boldsymbol{\psi}(\mathbf{x}_i) - \mathbf{HCH}\Big)_{\mathbf{kk}} && (93)\\
&= N\sum_{\mathbf{i}}\lambda_{\mathbf{k}}\phi_{\mathbf{ki}}\overline{\phi}_{\mathbf{ki}}\sum_{\mathbf{k}',\mathbf{l}'}\lambda_{\mathbf{k}'}^{1/2}C_{\mathbf{k}'\mathbf{l}'}\lambda_{\mathbf{l}'}^{1/2}\overline{\phi}_{\mathbf{k}'\mathbf{i}}\phi_{\mathbf{l}'\mathbf{i}} - \lambda_{\mathbf{k}}^2 C_{\mathbf{kk}} && (94)\\
&= \lambda_{\mathbf{k}}\sum_{\mathbf{k}',\mathbf{l}'}\lambda_{\mathbf{k}'}^{1/2}\lambda_{\mathbf{l}'}^{1/2}C_{\mathbf{k}'\mathbf{l}'}\sum_{\mathbf{i}}\overline{\phi}_{\mathbf{k}'\mathbf{i}}\phi_{\mathbf{l}'\mathbf{i}} - \lambda_{\mathbf{k}}^2 C_{\mathbf{kk}} && (95)\\
&= \lambda_{\mathbf{k}}\sum_{\mathbf{l}}\lambda_{\mathbf{l}}C_{\mathbf{ll}} - \lambda_{\mathbf{k}}^2 C_{\mathbf{kk}}. && (96)
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{Tr}(\phi(\mathbf{H})\mathbf{\Sigma}) &= \sum_{\mathbf{k}}\phi(\lambda_{\mathbf{k}})\Sigma_{\mathbf{kk}} && (97)\\
&= \sum_{\mathbf{k}}\phi(\lambda_{\mathbf{k}})\lambda_{\mathbf{k}}\sum_{\mathbf{l}}\lambda_{\mathbf{l}}C_{\mathbf{ll}} - \sum_{\mathbf{k}}\phi(\lambda_{\mathbf{k}})\lambda_{\mathbf{k}}^2 C_{\mathbf{kk}} && (98)\\
&= \mathrm{Tr}(\mathbf{H}\phi(\mathbf{H}))\,\mathrm{Tr}(\mathbf{HC}) - \mathrm{Tr}(\mathbf{H}^2\phi(\mathbf{H})\mathbf{C}), && (99)
\end{aligned}$$

which is the desired Eq. (86).

Now we comment on the degenerate case, when $\lambda_{\mathbf{k}} = 0$ for some $\mathbf{k}$. This occurs if there is linear dependence between some $\psi(\mathbf{x_i})$, i.e. $\dim\mathrm{Ran}(\mathbf{H}) < N$. The vectors $\mathbf{u_k}$ given by (91) with $\lambda_{\mathbf{k}} \neq 0$ form a basis in $\mathrm{Ran}(\mathbf{H})$. Inverse relation (92) remains valid with arbitrary vectors $\mathbf{u_k}$ for $\mathbf{k}$ such that $\lambda_{\mathbf{k}} = 0$ (this can be seen, for example, by lifting the degeneracy of the problem with a regularization ε and then letting $\varepsilon \to 0$). As a result, Eq. (96) remains valid for $\mathbf{k}$ such that $\lambda_{\mathbf{k}} > 0$. On the other hand, if $\mathbf{u}_k$ is an eigenvector corresponding to the eigenvalue $\lambda_k = 0$, i.e. $\mathbf{u}_k \in \ker\mathbf{H}$, then $\mathbf{u}_k$ is orthogonal to $\psi(\mathbf{x_i})$ for all $\mathbf{i}$, and then $\Sigma_{\mathbf{kk}} = 0$ for such k. Thus, formula (96) remains valid in this case too, and so Eq. (99) holds true. □

E GENERATING FUNCTIONS AND THEIR APPLICATIONS

E.1 REDUCTION TO GENERATING FUNCTIONS

We clarify some details of derivations in Section 3. Let us first discuss SE condition (7). Recall that we assumed $\mathbf{H}$ to be diagonalized by the eigenbasis $\mathbf{u}_k$ with eigenvalues λ_k. If the eigenvalues of $\mathbf{H}$ are non-degenerate, SE condition (7) can be equivalently written as a relation between the diagonal matrix elements Σ_{kk}, C_{kk} of the matrices $\mathbf{\Sigma}, \mathbf{C}$:

$$\Sigma_{kk} = \lambda_k\sum_{l}\lambda_l C_{ll} - \tau\lambda_k^2 C_{kk} \quad \forall k \tag{100}$$

(where we have incorporated the convention $\tau_1 = 1, \tau_2 = \tau$). More generally, if several λ_k may be equal to some λ and so the choice of eigenvectors $\mathbf{u}_k$ is not unique, conditions (100) should be replaced by weaker conditions (invariant w.r.t. this choice):

$$\sum_{k:\lambda_k=\lambda} \Sigma_{kk} = \sum_{k:\lambda_k=\lambda} \lambda \sum_l \lambda_l C_{ll} - \tau\lambda^2 \sum_{k:\lambda_k=\lambda} C_{kk}, \quad \forall \lambda \in \mathrm{spec}(\mathbf{H}). \tag{101}$$

Note, however, that representation (100) can be used without any restriction of generality if we are interested in the observables $\mathrm{Tr}[\phi(\mathbf{H})\mathbf{C}_t]$ (in particular, $L(t)$) along the optimization trajectory (5): by Proposition 5 such quantities are uniquely determined by initial conditions and the learning rates of SGD once the SE approximation is assumed.

Representations (9) – (11) now follow directly from the evolution law (5), by inspecting the evolution of diagonal elements in submatrices $\mathbf{C}_t, \mathbf{J}_t, \mathbf{V}_t$ using Eq. (100).

The remaining derivations in Section 3 are evident except for the expansions (18), (20) of the noise and signal generating functions $\widetilde{U}(z), \widetilde{V}(z)$ as series of scalar rational functions, which we now explain. Consider first expansion (18) for $\widetilde{U}(z)$. The terms of this expansion correspond to the eigenvalues λ_k of $\mathbf{H}$. For each eigenvalue, the respective term is, by Eq. (17),

$$\lambda^2 \langle \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right), (1 - zA_\lambda)^{-1} \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right) \rangle, \tag{102}$$

where A_λ is the linear operator acting on the space of 2×2 matrices and given by Eq. (11). The scalar product in (102) can be computed as the component x_{11} of the solution of the 4×4 linear system

$$(1 - zA_\lambda) \left(\begin{smallmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right). \tag{103}$$

Since all the components of the 4×4 matrix of the operator $1 - zA_\lambda$ are polynomial in the parameters $\alpha, \beta, \gamma, \tau, \lambda, z$, the resulting x_{11} is a rational function of these parameters. The explicit form of x_{11} is tedious to derive by hand, but it can be easily obtained with the help of a computer algebra system such as Sympy (Meurer et al., 2017). We provide the respective Jupyter notebook in the supplementary material.[2]

The terms of $\widetilde{V}$ are obtained similarly, but by solving the equation

$$(1 - zA_\lambda) \left(\begin{smallmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right) \tag{104}$$

instead of Eq. (103).

E.2 Convergence conditions

Proof of Proposition (3)**.** Note first that Proposition (3) is implied by the following lemma:

Lemma 1. *Assume (as in Proposition* (3)*) that* $\beta \in (-1, 1)$*,* $0 < \alpha < 2(\beta+1)/\lambda_{\max}$ *and* $\gamma \in [0, 1]$*. Then*

1. $S(\alpha, \beta, \gamma, \lambda, z) > 0$ *for all* $z \in (0, 1)$*;*
2. $\widetilde{U}(z), \widetilde{V}(z)$ *are analytic in an open neighborhood of the interval* $(0, 1)$ *and have no singularities there;*
3. $z\widetilde{U}(z)$ *is monotone increasing for all* $z \in (0, 1)$*.*

Indeed, by statement 2) and representation (16), a singularity of $\widetilde{L}(z) = \widetilde{V}(z)/(2(1 - z\widetilde{U}(z)))$ on the interval $(0, 1)$ can only be a pole resulting from a zero of the denominator $1 - z\widetilde{U}(z)$. By statement 3), such a zero is present if and only if $\widetilde{U}(z) > 1$, as claimed. It remains to prove the lemma.

Proof of Lemma 1.

[2] https://colab.research.google.com/drive/1eai1apCMCeLLbGIC7vfjbXNOKZ3CaIbw?usp=sharing

1) Our proof of statement 1 is computer-assisted. Note first that this statement can be written as the condition $W = \varnothing$ for a semi-algebraic set $W \subset \mathbb{R}^4$ defined by several polynomial inequalities with integer coefficients:

$$W = \{\exists(a, \beta, \gamma, z) \in \mathbb{R}^4 : (-1 < \beta < 1) \wedge (0 < a < 2(\beta + 1)) \tag{105}$$

$$\wedge\, (0 \leq \gamma \leq 1) \wedge (0 < z < 1) \wedge (S(a, \beta, \gamma, 1, z) \leq 0)\}, \tag{106}$$

where we have introduced the variable $a = \alpha\lambda$. By Tarski-Seidenberg theorem, the emptyness of a semi-algebraic set in $\mathbb{R}^n$ is an algorithmically decidable problem (solvable by sequential elimination of existential quantifiers, via repeated standard operations such as differentiation, evaluation of polynomials, division of polynomials with remainder, etc., see e.g. (Basu, 2014)). This procedure is implemented in some computer algebra systems, e.g. in REDUCE/Redlog (Dolzmann & Sturm, 1997). We used this system to verify that $W = \varnothing$:

```
1: rlset reals$
2: S := a**2*b*g*z**2 + a**2*b*z**2 + a**2*g*z - a**2*z - 2*a*b**2*z**2 -
2*a*b*z**2 + 2*a*b*z + 2*a*z - b**3*z**3 + b**3*z**2 + b**2*z**2 - b**2*z
+ b*z**2 - b*z - z + 1$
3: W := ex({a,b,g,z}, -1<b and b<1 and 0<a and a<2*(b+1) and 0<=g and g<=1
and 0<z and z<1 and S<=0)$
4: rlqe W;
false
```

2) Statement 2 follows from statement 1 and the fact that $S(\alpha, \beta, \gamma, \lambda, z) \to (1-z)(1-\beta z)(1-\beta^2 z)$ as $\lambda \to 0$. Indeed, since the series $\sum_k \lambda_k^2$ and $\sum_k \lambda_k C_{kk,0}$ converge, singularities in $\widetilde{U}, \widetilde{V}$ outside the points $z = 1, \beta^{-1}, \beta^{-2}$ can only be poles of some of the terms of expansions (18), (20) associated with zeros of S, which are excluded by statement 1.

3) Our proof of statement 3 is also computer-assisted and similar to the proof of statement 1. First we compute

$$\tfrac{d}{dz}(z\widetilde{U}(z)) = \gamma\alpha^2 \sum_k \lambda_k^2 \frac{R(\alpha, \beta, \lambda_k, z)}{S^2(\alpha, \beta, \gamma\tau, \lambda_k, z)}, \tag{107}$$

where

$$R(\alpha, \beta, \lambda_k, z) = -2\alpha^2\beta\lambda_k^2 z^2 + 4\alpha\beta^2\lambda_k z^2 + 4\alpha\beta\lambda_k z^2 + \beta^4 z^4 + 2\beta^3 z^3 - 2\beta^3 z^2 - 2\beta^2 z^2 - 2\beta z^2 + 2\beta z + 1. \tag{108}$$

By statement 1, it is sufficient to check that $R \geq 0$ on the domain of interest. This is done by verifying the emptyness of the set

$$W_1 = \{\exists(a, \beta, z) \in \mathbb{R}^3 : (-1 < \beta < 1) \wedge (0 < a < 2(\beta + 1)) \tag{109}$$

$$\wedge\, (0 < z < 1) \wedge (R(a, \beta, 1, z) < 0)\} \tag{110}$$

in REDUCE/Redlog:

```
1: rlset reals$
2: R1 := -2*a**2*b*z**2 + 4*a*b**2*z**2 + 4*a*b*z**2 + b**4*z**4
+ 2*b**3*z**3 - 2*b**3*z**2 - 2*b**2*z**2 - 2*b*z**2 + 2*b*z + 1$
3: W1 := ex({a,b,z}, -1<b and b<1 and 0<a and a<2*(b+1) and 0<z and z<1
and R1<0)$
4: rlqe W1;
false
```

□

We provide experimental verifications of statements 1 and 2 in the accompanying Jupyter notebook.

E.3 LOSS ASYMPTOTIC UNDER POWER-LAWS

The main purpose of this section is to prove theorem 2. We first formulate a lemma allowing to calculate the $\varepsilon \to +0,\ \varepsilon = 1 - z$ asymptotics of the generating functions (18) and (20).

Lemma 2. *Suppose that S_k and C_k are two sequences such that $S_k = k^{-\nu}(1 + o(1))$ and $F_k = \sum_{l=k}^{\infty} C_l = k^{-\varkappa}(1 + o(1))$ as $k \to \infty$, with some constants $\nu, \varkappa > 0$. Also denote $\zeta = \frac{\varkappa}{\nu}$. Then*

1. $$\sum_{l=k}^{\infty} C_l S_l = \frac{\zeta}{\zeta + 1} k^{-\varkappa-\nu}(1 + o(1)), \quad k \to \infty. \tag{111}$$

2. *Assume $\zeta < n$ for some $n > 0$, then*
$$\sum_{k=1}^{\infty} \frac{C_k}{(\varepsilon + S_k)^n} = \frac{\Gamma(\zeta + 1)\Gamma(n - \zeta)}{\Gamma(n)} \varepsilon^{\zeta - n}(1 + o(1)), \quad \varepsilon \to 0 + . \tag{112}$$

Proof.

Part 1. Fix k and consider a discrete measure μ_k with atoms located at points $x_{l,k} = (l/k)^{-\nu}$:

$$\mu_k = k^{\varkappa} \sum_{l=k}^{\infty} C_l \delta_{x_{l,k}} \tag{113}$$

By assumption on F_k as $k \to \infty$, the cumulative distribution function of μ_k converges to cumulative distribution function of measure $\mu(dx) = \mathbf{1}_{[0,1]} dx^{\zeta}$, i.e. the measure μ_k weakly converges to μ.

Next, consider the function $f_k(x)$ defined at points $x_{l,k}$ by $f_k(x_{l,k}) = k^{\nu} S_l$, with $f_k(0) = 0$, linearly interpolated between the points $x_{l,k}$, and with $f(x > 1) = f(x_{k,k})$. Then $f_k(x)$ converges uniformly on $[0, 1]$ to $f(x) = x$ as $k \to \infty$.

The above two observations on μ_k and f_k give

$$\lim_{k\to\infty} k^{\varkappa+\nu} \sum_{l=k}^{\infty} C_l S_l = \lim_{k\to\infty} \int_0^1 f_k(x)\mu_k(dx) = \int_0^1 x\zeta x^{\zeta-1} dx = \frac{\zeta}{\zeta + 1} \tag{114}$$

Part 2. For any $a > 0$ let $k_{a,\varepsilon} = \lfloor (\varepsilon a)^{-1/\nu} \rfloor$. Rescale the sum of interest and divide it into two parts separated by index $k_{a,\varepsilon}$:

$$\varepsilon^{n-\zeta} \sum_{k=1}^{\infty} \frac{C_k}{(\varepsilon + S_k)^n} = \varepsilon^{n-\zeta} \sum_{k=1}^{k_{a,\varepsilon}} + \varepsilon^{n-\zeta} \sum_{k=k_{a,\varepsilon}+1}^{\infty} =: I_1(a, \varepsilon) + I_2(a, \varepsilon). \tag{115}$$

We will show that

$$\limsup_{\varepsilon\to 0+} I_1(a, \varepsilon) = O(a^{\zeta-n}), \tag{116}$$

$$\lim_{\varepsilon\to 0+} I_2(a, \varepsilon) = \int_0^a \frac{dx^{\zeta}}{(1 + x)^n} =: J(a). \tag{117}$$

These two fact will imply the statement of the lemma since $\zeta - n < 0$ and

$$\lim_{a\to+\infty} J(a) = \int_0^{\infty} \frac{\zeta x^{\zeta-1}}{(1 + x)^n} dx \tag{118}$$

$$\overset{1+x=1/t}{=} \zeta \int_0^1 t^{n-\zeta-1}(1 - t)^{\zeta-1} dt \tag{119}$$

$$= \zeta B(n - \zeta, \zeta). \tag{120}$$

To prove Eq. (116), use summation by parts and the hypotheses on S_k, F_k:

$$\varepsilon^{n-\zeta}\sum_{k=1}^{k_{a,\varepsilon}}\frac{C_k}{(\varepsilon+S_k)^n} = \varepsilon^{n-\zeta}O\Big(\sum_{k=1}^{k_{a,\varepsilon}}C_k k^{\nu n}\Big) \tag{121}$$

$$= \varepsilon^{n-\zeta}O\Big(\sum_{k=1}^{k_{a,\varepsilon}}(F_k-F_{k+1})k^{\nu n}\Big) \tag{122}$$

$$= \varepsilon^{n-\zeta}O\Big(F_1 - F_{k_{a,\varepsilon}+1}k_{a,\varepsilon}^{\nu n} + \sum_{k=2}^{k_{a,\varepsilon}}F_k(k^{\nu n}-(k-1)^{\nu n})\Big) \tag{123}$$

$$= \varepsilon^{n-\zeta}O(k_{a,\varepsilon}^{\nu n-\varkappa}) \tag{124}$$

$$= O(a^{\zeta-n}). \tag{125}$$

To prove Eq. (117), it is convenient to represent the sum $I_2(a,\varepsilon)$ as an integral over discrete measure $\mu_{a,\varepsilon}$ with atoms $x_{k,\varepsilon} = k^{-\nu}/\varepsilon$:

$$I_2(a,\varepsilon) = \int f_\varepsilon d\mu_{a,\varepsilon}(x), \tag{126}$$

$$\mu_{a,\varepsilon} = \varepsilon^{-\zeta}\sum_{k=k_{a,\varepsilon}+1}^{\infty}C_k\delta_{x_{k,\varepsilon}}, \tag{127}$$

$$f_\varepsilon(x_{k,\varepsilon}) = (1+S_k/\varepsilon)^{-n}. \tag{128}$$

Again, by assumption on F_k, as $\varepsilon \to 0+$, the cumulative distribution functions of $\mu_{a,\varepsilon}$ converge to the cumulative distribution function of the measure $\mu_a(dx) = \mathbf{1}_{[0,a]}dx^\zeta$, i.e., the measures $\mu_{a,\varepsilon}$ weakly converge to μ_a. At the same time, the functions $f_\varepsilon(x)$ converge to $(1+x)^{-n}$ uniformly on $[0,a]$. This implies desired convergence (117). □

Now we are ready to proceed with the proof of theorem 2.

Theorem 2. *Assume spectral conditions* (24) *with* $\nu > 1$ *and that parameters* α,β,γ *are as in Proposition 3 and such that convergence condition* $\widetilde{U}(1) < 1$ *is satisfied. Then*

$$\sum_{t=1}^{T}tL(t) = (1+o(1))\begin{cases}\frac{K\Gamma(\zeta+1)}{2(1-\widetilde{U}(1))(2-\zeta)}(\frac{2\alpha\Lambda}{1-\beta})^{-\zeta}t^{2-\zeta}, & 2-\zeta > 1/\nu,\\ \frac{\gamma\widetilde{V}(1)\Gamma(2-1/\nu)}{8(1-\widetilde{U}(1))^2}(\frac{2\alpha\Lambda}{1-\beta})^{1/\nu}t^{1/\nu}, & 2-\zeta < 1/\nu.\end{cases} \tag{25}$$

Proof. As asymptotic of partial sums $\sum_{t=1}^{T}tL(t)$ are associated with asymptotic of $\widetilde{L}'(1-\varepsilon)$ at $\varepsilon \to +0$ we recall its expression (26)

$$\frac{d}{dz}\widetilde{L}(z) = \frac{\frac{d}{dz}\widetilde{V}(z)}{2(1-z\widetilde{U}(z))} + \frac{\widetilde{V}(z)\frac{d}{dz}(z\widetilde{U}(z))}{2(1-z\widetilde{U}(z))^2} \tag{129}$$

Full expressions for generating functions (18) and (20) can be significantly simplified in the limit $\varepsilon \to +0$ and $\lambda_k \to 0$

$$\widetilde{U}(1-\varepsilon) = \gamma\sum_{k=1}^{\infty}\frac{\left(\frac{\alpha\lambda_k}{1-\beta}\right)^2}{\varepsilon+2\frac{\alpha\lambda_k}{1-\beta}}(1+O(\lambda_k)+O(\varepsilon)) \tag{130}$$

$$\widetilde{V}(1-\varepsilon) = \sum_{k=1}^{\infty}\frac{\lambda_k C_{kk}}{\varepsilon+2\frac{\alpha\lambda_k}{1-\beta}}(1+O(\lambda_k)+O(\varepsilon)) \tag{131}$$

Differentiating we get

$$\widetilde{U}'(1-\varepsilon) = \gamma \sum_{k=1}^{\infty} \frac{\left(\frac{\alpha\lambda_k}{1-\beta}\right)^2}{\left(\varepsilon + 2\frac{\alpha\lambda_k}{1-\beta}\right)^2}(1 + O(\lambda_k) + O(\varepsilon)) \tag{132}$$

$$\widetilde{V}'(1-\varepsilon) = \sum_{k=1}^{\infty} \frac{\lambda_k C_{kk}}{\left(\varepsilon + 2\frac{\alpha\lambda_k}{1-\beta}\right)^2}(1 + O(\lambda_k) + O(\varepsilon)) \tag{133}$$

Let's make a couple of observations. From asymptotic expressions above one can show that

$$\widetilde{U}(1-\varepsilon) \sim \sum_{k=1}^{\infty} \frac{\lambda_k^2}{\varepsilon + \lambda_k} \sim \sum_{k=1}^{\varepsilon^{-\frac{1}{\nu}}} \lambda_k \sim 1 \tag{134}$$

$$\widetilde{U}'(1-\varepsilon) \sim \sum_{k=1}^{\infty} \frac{\lambda_k^2}{(\varepsilon + \lambda_k)^2} \sim \sum_{k=1}^{\varepsilon^{-\frac{1}{\nu}}} 1 \sim \varepsilon^{-\frac{1}{\nu}} \tag{135}$$

$$\widetilde{V}(1-\varepsilon) \sim \sum_{k=1}^{\infty} \frac{\lambda_k C_{kk}}{\varepsilon + \lambda_k} \sim \sum_{k=1}^{\varepsilon^{-\frac{1}{\nu}}} C_{kk} \sim \begin{cases} 1, & \zeta > 1 \\ \varepsilon^{\zeta-1}, & \zeta < 1 \end{cases} \tag{136}$$

$$\widetilde{V}'(1-\varepsilon) \sim \sum_{k=1}^{\infty} \frac{\lambda_k C_{kk}}{(\varepsilon + \lambda_k)^2} \sim \sum_{k=1}^{\varepsilon^{-\frac{1}{\nu}}} \frac{C_{kk}}{\lambda_k} \sim \begin{cases} 1, & \zeta > 2 \\ \varepsilon^{\zeta-2}, & \zeta < 2 \end{cases} \tag{137}$$

Here the sign $\sim$ means an asymptotic ($\varepsilon \to 0$) equality up to a multiplicative constant. Using these asymptotic relations we see that for $\widetilde{L}'(1-\varepsilon)$ leading asymptotic terms are given by either $\widetilde{U}'(1-\varepsilon)$ when $\zeta > 2 - \frac{1}{\nu}$, or by $\widetilde{V}'(1-\varepsilon)$ when $\zeta < 2 - \frac{1}{\nu}$.

Now we set for simplicity $\Lambda = K = 1$ and apply lemma 2 to (132) and (133) to get more refined asymptotic expressions. For $V'(1-\varepsilon)$ we assume $\zeta < 2$ because $V'(1-\varepsilon) = O(1)$ for $\zeta > 2$ and refined asymptotic is not needed.

$$\begin{aligned}
\widetilde{V}'(1-\varepsilon) &= \sum_{k=1}^{\infty} \frac{\lambda_k C_{kk}(1 + O(\lambda_k) + O(\varepsilon))}{\left(\varepsilon + 2\frac{\alpha\lambda_k}{1-\beta}\right)^2} \\
&= (1 + O(\varepsilon)) \sum_{k=1}^{\infty} \frac{\lambda_k C_{kk}}{\left(\varepsilon + 2\frac{\alpha\lambda_k}{1-\beta}\right)^2} + \sum_{k=1}^{\infty} \frac{O(C_{kk}\lambda_k^2)}{\left(\varepsilon + 2\frac{\alpha\lambda_k}{1-\beta}\right)^2} \\
&\overset{(1)}{=} \left(\frac{2\alpha}{1-\beta}\right)^{-2} \sum_{k=1}^{\infty} \frac{C_{kk}\lambda_k}{\left(\frac{\varepsilon(1-\beta)}{2\alpha} + \lambda_k\right)^2} + O(\varepsilon^{\min(0,\zeta-1)}) \\
&\overset{(1)}{=} \Gamma(\zeta+1)\Gamma(2-\zeta)\left(\frac{2\alpha}{1-\beta}\right)^{-\zeta} \varepsilon^{\zeta-2}(1 + o(1))
\end{aligned} \tag{138}$$

Here in (1) we first used first part of lemma 2 to get $\sum_{l \geq k} C_{ll}\lambda_l^2 = \frac{\zeta}{\zeta+1} k^{-\varkappa-\nu}(1 + o(1))$. Then for $1 < \zeta < 2$ the second sum is finite (see (137)) and for $\zeta < 1$ second part of lemma 2 allows to bound the sum as $O(\varepsilon^{\zeta-1})$. In (2) we applied second part of lemma 2 to the first sum in (1).

For $U'(1-\varepsilon)$ we have

$$\begin{aligned}\widetilde{U}'(1-\varepsilon) &= \gamma \sum_{k=1}^{\infty} \frac{\left(\frac{\alpha\lambda_k}{1-\beta}\right)^2 (1+O(\lambda_k)+O(\varepsilon))}{\left(\varepsilon + 2\frac{\alpha\lambda_k}{1-\beta}\right)^2} \\ &\overset{(1)}{=} \frac{\gamma}{4}(1+O(\varepsilon)) \sum_{k=1}^{\infty} \frac{\lambda_k^2}{\left(\frac{\varepsilon(1-\beta)}{2\alpha}+\lambda_k\right)^2} + \frac{\gamma}{4}\sum_{k=1}^{\infty} \frac{O(\lambda_k^3)}{\left(\frac{\varepsilon(1-\beta)}{2\alpha}+\lambda_k\right)^2} \\ &\overset{(2)}{=} \frac{\gamma}{4\nu}\Gamma(2-\frac{1}{\nu})\Gamma(\frac{1}{\nu})\left(\frac{2\alpha}{1-\beta}\right)^{\frac{1}{\nu}} \varepsilon^{-\frac{1}{\nu}}(1+o(1))\end{aligned} \tag{139}$$

In (1) the second sum can be bounded as $O(1)$ after recalling that $\nu > 1$. In (2) we again applied second part of lemma 2 with $C_k = \lambda_k^2$ and $n = 2$ since $\sum_{l\geq k} \lambda_k^2 = \frac{1}{2\nu-1}k^{-2\nu+1}(1+o(1))$. To get the coefficient in (2) we used $\frac{\Gamma(3-1/\nu)}{2\nu-1} = \frac{\Gamma(2-1/\nu)}{\nu}$.

Substituting (139) and (138) into (129) we get

$$\widetilde{L}'(1-\varepsilon) = \frac{C_{V'}}{2(1-\widetilde{U}(1))}\varepsilon^{-\max(0,2-\zeta)}(1+o(1)) + \frac{\widetilde{V}(1)C_{U'}}{2(1-\widetilde{U}(1))^2}\varepsilon^{-\frac{1}{\nu}}(1+o(1)) \tag{140}$$

where $C_{V'}$ and $C_{U'}$ are constants in the asymptotics (138) and (139). Picking the leading term depending on the sign of $\zeta - (2-\frac{1}{\nu})$ we obtain final expressions

$$\widetilde{L}'(1-\varepsilon) = \begin{cases} \frac{\Gamma(\zeta+1)\Gamma(2-\zeta)}{2(1-\widetilde{U}(1))}\left(\frac{2\alpha}{1-\beta}\right)^{-\zeta}\varepsilon^{\zeta-2}(1+o(1)), & \zeta < 2-\frac{1}{\nu} \\ \gamma\frac{\widetilde{V}(1)\Gamma(2-\frac{1}{\nu})\Gamma(\frac{1}{\nu})}{8\nu(1-\widetilde{U}(1))^2}\left(\frac{2\alpha}{1-\beta}\right)^{\frac{1}{\nu}}\varepsilon^{-\frac{1}{\nu}}(1+o(1)), & \zeta < 2-\frac{1}{\nu} \end{cases} \tag{141}$$

The last step is to apply Tauberian theorem (Feller (1971), p. 445) which states that if generating function $\widetilde{F}(z)$ of a sequence $F(t)$ has asymptotic $\widetilde{F}(1-\varepsilon) = C\varepsilon^{-\rho}(1+o(1))$, then the partial sums have the asymptotic $\sum_{t=1}^{T} F(t) = \frac{C}{\Gamma(\rho+1)}T^{\rho}(1+o(1))$. Applying this theorem to $F(t) = tL(t)$ and its generating function given by (141) we get precisely (25) with $\Lambda = K = 1$. To restore the values of constants Λ, K recall that the part of the loss coming from signal ($\widetilde{V}(z)$) is simply proportional to K, and Λ always goes in combination $\alpha\Lambda$. □

Finally, let us show how (25) can be used to obtain asymptotics (27a),(27b) for the loss values $L(t)$ themselves. Assuming that $L(t) = Ct^{-\xi}, \xi < 2$, we get

$$\sum_{t=1}^{T} tL(t) = C\sum_{t=1}^{T} t^{-\xi+1}(1+o(1)) \approx C\int_{1}^{T} t^{-\xi+1}(1+o(1)) = \frac{C}{2-\xi}T^{2-\xi}(1+o(1)) \tag{142}$$

Comparing this with (25), we find the constants C appearing in (27a),(27b).

E.4 Exponential divergence of the loss and convergence-divergence transition

Convergence radius r_L and exponential divergence rate. By Proposition 3, the convergence radius $r_L < 1$ if and only if $\widetilde{U}(1) > 1$, and in the case $\widetilde{U}(1) > 1$ it is determined by the condition

$$1 = r_l\widetilde{U}(r_L). \tag{143}$$

If $r_L < 1$, loss exponentially diverges as $t \to \infty$:

Proposition 7. *Under assumption of Proposition 3, suppose that $r_L < 1$. Then*

$$\sum_{t=0}^{T} L(t)r_L^t = \frac{\widetilde{V}(r_L)}{2(1+r_L^2\widetilde{U}'(r_L))}T(1+o(1)), \quad (T\to\infty). \tag{144}$$

Proof. Consider the modified generating function $\widetilde{L}_1(z) = \widetilde{L}(r_L z)$, which is the generating function for the sequence $L(t)r_L^t$:

$$\widetilde{L}_1(z) = \sum_{t=0}^{\infty} z^t L(t) r_L^t. \tag{145}$$

The convergence radius for $\widetilde{L}_1$ is 1. We can also write

$$\widetilde{L}_1(1-\epsilon) = \frac{\widetilde{V}(r_L(1-\epsilon))}{2(1-r_L(1-\epsilon)\widetilde{U}(r_L(1-\epsilon)))} \tag{146}$$

$$= \frac{\widetilde{V}(r_L)}{2(1+r_L^2\widetilde{U}'(r_L))\epsilon}(1+o(1)), \quad (\epsilon \to 0). \tag{147}$$

The statement of the proposition then follows by the Tauberian theorem. □

If we assume that

$$L(t) \sim C r_L^{-t} \tag{148}$$

with some constant C, then this constant is fixed by Eq. (144):

$$L(t) \sim \frac{\widetilde{V}(r_L)}{2(1+r_L^2\widetilde{U}'(r_L))} r_L^{-t}. \tag{149}$$

"Eventual divergence" phase with small α. Recall that we characterized the "eventual divergence" phase by the condition $\widetilde{U}(1) = \infty$. In this phase we necessarily have $r_L < 1$. However, r_L can be very close to 1 (and thus "divergence delayed") – for example if the learning rate α is very small. In this scenario we can derive an asymptotic expression for the difference $\epsilon_* = 1 - r_L$.

In the sequel, we assume for simplicity that $\beta = 0$ and $\tau = \gamma = 1$, and that the eigenvalues and initial covariance coefficients $C_{kk,0}$ satisfy power law relations (24) with $\frac{1}{2} < \nu < 1$ (corresponding to the "eventual divergence" phase).

Proposition 8. *As* $\alpha \to 0$,

$$\epsilon_* \equiv 1 - r_L = \Big(\frac{1}{4\nu}\Gamma(2-1/\nu)\Gamma(1/\nu-1)\Big)^{\nu/(1-\nu)}(2\alpha\Lambda)^{1/(1-\nu)}(1+o(1)). \tag{150}$$

Proof. With $\beta = 0$ and $\tau = \gamma = 1$, we have by Eq. (18)

$$\widetilde{U}(1-\epsilon) = \alpha^2 \sum_k \frac{\lambda_k^2}{2\alpha\lambda_k(1-\epsilon)+\epsilon} \tag{151}$$

$$= \frac{\alpha}{2\Lambda}\sum_k \frac{\lambda_k^2/(1-2\alpha\lambda_k)}{\lambda_k/(\Lambda(1-2\alpha\lambda_k))+\epsilon/(2\alpha\Lambda)}. \tag{152}$$

For any α, we can now apply Lemma 2(2) with $\varkappa = 2\nu - 1, \zeta = 2 - 1/\nu$ and $n = 1$ and get

$$\widetilde{U}(1-\epsilon) = \frac{\alpha\Lambda}{2(2\nu-1)}\Gamma(3-1/\nu)\Gamma(1/\nu-1)\Big(\frac{\epsilon}{2\alpha\Lambda}\Big)^{1-1/\nu}(1+o(1)) \tag{153}$$

$$= \frac{1}{4\nu}\Gamma(2-1/\nu)\Gamma(1/\nu-1)(2\alpha\Lambda)^{1/\nu}\epsilon^{1-1/\nu}(1+o(1)), \quad (\epsilon \to 0). \tag{154}$$

It is easy to see that $o(1)$ here is uniform for small α. The desired $\epsilon_* = \epsilon_*(\alpha)$ is determined from the equation $1 = (1-\epsilon_*)\widetilde{U}(1-\epsilon_*)$. As $\alpha \to 0$, we clearly have $\epsilon_* = \epsilon_*(\alpha) \to 0$. It follows that

$$1 = \frac{1}{4\nu}\Gamma(2-1/\nu)\Gamma(1/\nu-1)(2\alpha\Lambda)^{1/\nu}\epsilon_*(\alpha)^{1-1/\nu}(1+o(1)), \quad (\alpha \to 0). \tag{155}$$

This implies desired expression (150). □

We can similarly use Lemma 2 to also estimate the derivative $\widetilde{U}'(1-\epsilon)$:

$$\widetilde{U}'(1-\epsilon) = \alpha^2 \sum_k \frac{\lambda_k^2(1-2\alpha\lambda_k)}{(2\alpha\lambda_k(1-\epsilon_*)+\epsilon)^2} \tag{156}$$

$$= \frac{1}{4\Lambda^2} \sum_k \frac{\lambda_k^2/(1-2\alpha\lambda_k)}{(\lambda_k/(\Lambda(1-2\alpha\lambda_k))+\epsilon/(2\alpha\Lambda))^2} \tag{157}$$

$$= \frac{1}{4(2\nu-1)}\Gamma(3-1/\nu)\Gamma(1/\nu)\Big(\frac{\epsilon}{2\alpha\Lambda}\Big)^{-1/\nu}(1+o(1)) \tag{158}$$

$$= \frac{1}{4\nu}\Gamma(2-1/\nu)\Gamma(1/\nu)(2\alpha\Lambda)^{1/\nu}\epsilon^{-1/\nu}(1+o(1)) \tag{159}$$

$$= \frac{1/\nu-1}{\epsilon}\widetilde{U}(1-\epsilon)(1+o(1)), \quad (\epsilon \to 0) \tag{160}$$

(this result can also formally be obtained by differentiating Eq. (153) in ϵ). In particular, for $\epsilon = \epsilon_*$ we get

$$\widetilde{U}'(1-\epsilon_*) = \frac{1/\nu-1}{\epsilon_* r_L}(1+o(1)), \quad (\alpha \to 0). \tag{161}$$

"Early convergent" regime. If $r_L < 1$, then, as discussed above, loss eventually diverges exponentially. However, in general, we expect this divergence to occur only after some interval of convergence (see Figure 3(right)). We will give the loss asymptotic for this "early convergent" phase and give a rough estimate of the time scale at which the transition between the two regimes occurs. We retain the assumptions $\beta = 0$ and $\gamma = \tau = 1$ from the previous paragraph, and assume power laws (24) with $\frac{1}{2} < \nu < 1$ and $\zeta < 1$.

We start with finding the loss asymptotic in the early convergent phase. First we argue that this phase can be approximately described by the loss generating function

$$\widetilde{L}(z) \approx \frac{1}{2}\widetilde{V}(z), \tag{162}$$

i.e. generating function (16) in which we approximate the denominator by 1:

$$1 - z\widetilde{U}(z) \approx 1. \tag{163}$$

This approximation can be justified for $z < r_L$ by noting that, by Eq. (153),

$$\widetilde{U}(1-\epsilon) \approx \Big(\frac{\epsilon}{\epsilon_*}\Big)^{1-1/\nu}\widetilde{U}(1-\epsilon_*) = \Big(\frac{\epsilon}{\epsilon_*}\Big)^{1-1/\nu} \tag{164}$$

with $1 - 1/\nu < 0$. We remark that approximations (163) and (162) essentially mean that we ignore terms containing factors $U_{t,s}$ associated with noise in expansion (12).

If relation (162) held for all $z \in (0,1)$, then we would have an asymptotic convergence of the loss for all t. Indeed, under our assumptions,

$$\widetilde{V}(1-\epsilon) = \sum_{k=1}^{\infty} \frac{\lambda_k C_{kk,0}}{2\alpha\lambda_k(1-\epsilon)+\epsilon} \tag{165}$$

and $\widetilde{V}(1-\epsilon) \to \infty$ as $\epsilon \to 0$ (since $\zeta < 1$). Applying again Lemma 2, we find

$$\widetilde{V}(1-\epsilon) = \frac{1}{2\alpha\Lambda}\sum_k \frac{\lambda_k C_{kk,0}/(1-2\alpha\lambda_k)}{\lambda_k/(\Lambda(1-2\alpha\lambda_k))+\epsilon/(2\alpha\Lambda)} \tag{166}$$

$$= \frac{K}{2\alpha\Lambda}\Gamma(\zeta+1)\Gamma(1-\zeta)\Big(\frac{\epsilon}{2\alpha\Lambda}\Big)^{\zeta-1}(1+o(1)) \tag{167}$$

$$= \frac{K}{2}\Gamma(\zeta+1)\Gamma(1-\zeta)(2\alpha\Lambda)^{-\zeta}\epsilon^{\zeta-1}(1+o(1)), \quad (\epsilon \to 0). \tag{168}$$

The Tauberian theorem then implies that

$$\sum_{t=0}^{T} L(t) = \frac{K}{2\Gamma(2-\zeta)}\Gamma(\zeta+1)\Gamma(1-\zeta)(2\alpha\Lambda)^{-\zeta}T^{1-\zeta}(1+o(1)) \tag{169}$$

$$= \frac{K}{2(1-\zeta)}\Gamma(\zeta+1)(2\alpha\Lambda)^{-\zeta}T^{1-\zeta}(1+o(1)), \quad (\epsilon \to 0). \tag{170}$$

Assuming a power-law form

$$L(t) \sim Ct^{-\zeta}, \tag{171}$$

we then find

$$L(t) \sim \frac{K}{2}\Gamma(\zeta+1)(2\alpha\Lambda)^{-\zeta}t^{-\zeta}. \tag{172}$$

This expression agrees with "signal-dominated" loss asymptotic (27a) up to the factor $(1-\widetilde{U}(1))^{-1}$ (missing because of our approximation (163)).

Transition from convergence to divergence. Of course, the obtained asymptotic (172) does not hold for very large t, since representation (162) breaks down at $z \gtrsim r_L$, and the loss trajectory switches to exponential divergence. We can roughly estimate the switching time t_{blowup} by equating the asymptotic expressions for $L(t)$ in the convergent and divergent phases:

$$\frac{K}{2}\Gamma(\zeta+1)(2\alpha\Lambda)^{-\zeta}t_{\text{blowup}}^{-\zeta} \approx \frac{\widetilde{V}(r_L)}{2(1+r_L^2\widetilde{U}'(r_L))}r_L^{-t_{\text{blowup}}}. \tag{173}$$

This is a transcendental equation not explicitly solvable in t_{blowup}. We simplify it using a number of approximations. Using approximation (168) for $\widetilde{V}(r_L)$ and the fact that ϵ_* is small, we can simplify this equation to

$$\frac{1+r_L^2\widetilde{U}'(1-\epsilon_*)}{\Gamma(1-\zeta)}\epsilon_*^{1-\zeta}t_{\text{blowup}}^{-\zeta} \approx (1-\epsilon_*)^{-t_{\text{blowup}}} \approx e^{\epsilon_* t_{\text{blowup}}}. \tag{174}$$

Next, using Eq. (161), we can approximate the l.h.s. by

$$\frac{1+r_L^2\widetilde{U}'(1-\epsilon_*)}{\Gamma(1-\zeta)}\epsilon_*^{1-\zeta}t_{\text{blowup}}^{-\zeta} \approx \frac{1/\nu-1}{\Gamma(1-\zeta)}\epsilon_*^{-\zeta}t_{\text{blowup}}^{-\zeta}. \tag{175}$$

It follows that Eq. (174) can be approximated as

$$\frac{1/\nu-1}{\Gamma(1-\zeta)}(\epsilon_* t_{\text{blowup}})^{-\zeta} = e^{\epsilon_* t_{\text{blowup}}}. \tag{176}$$

Thus, we can write

$$t_{\text{blowup}} \approx \frac{a_*}{\epsilon_*}, \tag{177}$$

where a_* is the solution of the equation

$$\frac{1/\nu-1}{\Gamma(1-\zeta)}a_*^{-\zeta} = e^{a_*}. \tag{178}$$

If $t_{\text{div}} = -1/\log r_L$ denotes the characteristic time scale of exponential divergence, then at small α we have $t_{\text{div}} \approx 1/\epsilon_*$, so the characteristic transition time t_{blowup} is related to t_{div} by the simple relation

$$t_{\text{blowup}} \approx a_* t_{\text{div}}. \tag{179}$$

E.5 Budget optimization

Figure 5 shows that the loss is, approximately, a function of total computation budget given by the product bt, regardless of specific values of b and t. See (Ma et al., 2018; Shallue et al., 2018) for general discussions of batch scalability.

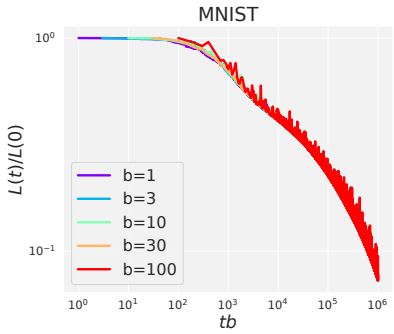

Figure 5: Representation of loss $L(t)$ obtained using different batch sizes b as (approximately) a function of the computation budged bt (regardless of how it is factorized over b and t) for MNIST. Momentum is fixed at $\beta = 0$ and the ratio α/b is kept the same for all curves.

E.6 Negative momenta

Our analysis of the effect of momentum is based on the approximations (27a), (27b) for $L(t)$ at large t. Let us denote these approximations by $L_{\text{approx}}(t)$:

$$L(t) \approx L_{\text{approx}}(t) := \begin{cases} \frac{K\Gamma(\zeta+1)}{2(1-\widetilde{U}(1))}\left(\frac{2\alpha\Lambda}{1-\beta}\right)^{-\zeta} t^{-\zeta}, & \zeta < 2 - 1/\nu, \quad (180a) \\ \frac{\gamma\widetilde{V}(1)\Gamma(2-1/\nu)}{8\nu(1-\widetilde{U}(1))^2}\left(\frac{2\alpha\Lambda}{1-\beta}\right)^{1/\nu} t^{1/\nu-2}, & \zeta > 2 - 1/\nu. \quad (180b) \end{cases}$$

We can now formulate a rigorous result on how $L_{\text{approx}}(t)$ is affected by changing the momentum from $\beta = 0$ at the optimal α. Note that t enters expression $L_{\text{approx}}(t)$ only through the factor $t^{-\zeta}$ or $t^{1/\nu-2}$; as a result, the statements below hold in the same form for all $t > 0$.

Proposition 9. *Suppose that the eigenvalues λ_k and initial moments $C_{kk,0}$ are subject to large-k power law asymptotics* (24) *with some exponents $\nu, \varkappa > 0, \zeta = \varkappa/\nu$ and coefficients Λ, K. Let $\tau = \gamma = 1$. Let $\alpha_{\text{opt}} = \arg\min_\alpha L_{\text{approx}}(t)|_{\beta=0}$.*

*I. (**signal-dominated phase**) If $\zeta < 2 - 1/\nu$, then*

1. $\alpha_{\text{opt}} = \frac{2\zeta}{\zeta+1}(\sum_k \lambda_k)^{-1}$;
2. $\partial_\beta L_{\text{approx}}(t)|_{\beta=0,\alpha=\alpha_{\text{opt}}} < 0$.

*II. (**noise-dominated phase**) If $\zeta > 2 - 1/\nu$, then*

1. $\alpha_{\text{opt}} = \frac{2(\nu-1)}{3\nu-1}(\sum_k \lambda_k)^{-1}$;
2. *The sign of $\partial_\beta L_{\text{approx}}(t)|_{\beta=0,\alpha=\alpha_{\text{opt}}}$ equals the sign of the expression*

$$\nu \sum_k \lambda_k \sum_k \lambda_k C_{kk,0} - (\nu - 1) \sum_k \lambda_k^2 \sum_k C_{kk,0} \quad (181)$$

(as in Eq. (29)*).*

Before giving the proof, note that if $\beta = 0$ then, by Eq. (21), our main stability condition $\widetilde{U}(1) \leq 1$ (see Prop. 3) becomes simply

$$\alpha \leq \alpha_{\max} := 2\Big(\sum_k \lambda_k\Big)^{-1}. \quad (182)$$

It follows that α_{opt} described in statements I1 and I2 belong to the allowed interval $(0, \alpha_{\max})$.

Proof of Proposition 9.

I. (**signal-dominated phase**) Differentiating Eq. (180a) and denoting by ϕ various (in general, different) positive values, we get

$$\partial_\alpha L_{\text{approx}}(t)|_{\beta=0} = \phi[(1-\widetilde{U}(z=1,\beta=0))\partial_\alpha \alpha^{-\zeta} + \alpha^{-\zeta}\partial_\alpha \widetilde{U}(z=1,\beta=0)] \tag{183}$$

$$= \phi\alpha^{-\zeta}\Big[(1-\frac{\alpha}{2}\sum_k \lambda_k)\frac{-\zeta}{\alpha} + \frac{1}{2}\sum_k \lambda_k\Big] \tag{184}$$

$$= \phi\Big[\alpha\frac{\zeta+1}{2}\sum_k \lambda_k - \zeta\Big], \tag{185}$$

implying statement 1:

$$\alpha_{\text{opt}} = \frac{2\zeta}{\zeta+1}\Big(\sum_k \lambda_k\Big)^{-1}. \tag{186}$$

Next,

$$\partial_\beta L_{\text{approx}}(t)|_{\beta=0} = \phi[(1-\widetilde{U}(z=1,\beta=0))\partial_\beta (1-\beta)^\zeta + (1-\beta)^\zeta\partial_\beta \widetilde{U}(z=1,\beta=0)] \tag{187}$$

$$= \phi(1-\beta)^\zeta\Big[(1-\frac{\alpha}{2}\sum_k \lambda_k)\frac{-\zeta}{1-\beta} + \frac{\alpha}{2}\sum_k \lambda_k(1-\alpha\lambda_k)\Big]\Big|_{\beta=0} \tag{188}$$

$$= \phi\Big[(1-\frac{\alpha}{2}\sum_k \lambda_k)(-\zeta) + \frac{\alpha}{2}\sum_k \lambda_k(1-\alpha\lambda_k)\Big]. \tag{189}$$

Substituting $\alpha = \alpha_{\text{opt}}$ we get

$$\partial_\beta L_{\text{approx}}(t)|_{\beta=0,\alpha=\alpha_{\text{opt}}} = \phi\Big[\big(1-\tfrac{\zeta}{\zeta+1}\big)(-\zeta) + \tfrac{\zeta}{\zeta+1}\big(1-\tfrac{\zeta}{\zeta+1}\big(\sum_k \lambda_k^2(\sum_k \lambda_k)^{-1}\big)\big)\Big] \tag{190}$$

$$= \phi\Big[-\big(\tfrac{\zeta}{\zeta+1}\big)^2(\sum_k \lambda_k)^{-1}\sum_k \lambda_k^2\Big] \tag{191}$$

$$< 0, \tag{192}$$

proving statement 2.

II. (**noise-dominated phase**) Differentiating Eq. (180b),

$$\partial_\alpha L_{\text{approx}}(t)|_{\beta=0} = \phi\Big[(1-\widetilde{U}(1))[\widetilde{V}(1)\partial_\alpha \alpha^{1/\nu} + \alpha^{1/\nu}\partial_\alpha \widetilde{V}(1)] + 2\alpha^{1/\nu}\widetilde{V}(1)\partial_\alpha \widetilde{U}(1)\Big]\Big|_{\beta=0}$$

$$= \phi\Big[\Big(1-\frac{\alpha}{2}\sum_k \lambda_k\Big)\Big(\frac{1}{\nu\alpha}(2\alpha)^{-1}\sum_k C_{kk,0} - (2\alpha^2)^{-1}\sum_k C_{kk,0}\Big) \tag{193}$$

$$+ 2(2\alpha)^{-1}\sum_k C_{kk,0}\frac{1}{2}\sum_k \lambda_k\Big] \tag{194}$$

$$= \phi\Big[\Big(1-\frac{\alpha}{2}\sum_k \lambda_k\Big)\Big(\frac{1}{\nu}-1\Big) + \alpha\sum_k \lambda_k\Big] \tag{195}$$

$$= \phi\Big[\Big(\frac{1}{\nu}-1\Big) + \alpha\Big(\frac{3}{2}-\frac{1}{2\nu}\Big)\sum_k \lambda_k\Big] \tag{196}$$

implying statement 1:

$$\alpha_{\text{opt}} = \frac{2(\nu-1)}{3\nu-1}\Big(\sum_k \lambda_k\Big)^{-1}. \tag{197}$$

Next,

$$\partial_\beta L_{\text{approx}}(t)|_{\beta=0} = \phi\Big[(1-\widetilde{U}(1)\big[\widetilde{V}(1)\partial_\beta(1-\beta)^{-1/\nu} + (1-\beta)^{-1/\nu}\partial_\beta\widetilde{V}(1)\big] \tag{198}$$

$$+ 2(1-\beta)^{-1/\nu}\widetilde{V}(1)\partial_\beta\widetilde{U}(1)\Big]\Big|_{\beta=0} \tag{199}$$

$$= \phi\Big[\Big(1-\frac{\alpha}{2}\sum_k\lambda_k\Big)\Big(\frac{(2\alpha)^{-1}}{\nu(1-\beta)}\sum_k C_{kk,0} - (2\alpha)^{-1}\sum_k C_{kk,0}(1-\alpha\lambda_k)\Big)$$

$$+ 2(2\alpha)^{-1}\sum_k C_{kk,0}\frac{\alpha}{2}\sum_k\lambda_k(1-\alpha_k\lambda_k)\Big]\Big|_{\beta=0} \tag{200}$$

$$= \phi\Big[\Big(1-\frac{\alpha}{2}\sum_k\lambda_k\Big)\Big((\tfrac{1}{\nu}-1)\sum_k C_{kk,0} \tag{201}$$

$$+ \alpha\sum_k C_{kk,0}\lambda_k\Big) + \alpha\sum_k C_{kk,0}\Big(\sum_k\lambda_k - \alpha\sum_k\lambda_k^2\Big)\Big]. \tag{202}$$

Substituting $\alpha = \alpha_{\text{opt}}$, we find

$$\partial_\beta L_{\text{approx}}(t)\Big|_{\substack{\beta=0\\ \alpha=\alpha_{\text{opt}}}} = \phi\Big[\frac{2\nu}{3\nu-1}\Big((\tfrac{1}{\nu}-1)\sum_k C_{kk,0} + (\sum_k\lambda_k)^{-1}\tfrac{2(\nu-1)}{3\nu-1}\sum_k C_{kk,0}\lambda_k\Big)$$

$$+ (\sum_k\lambda_k)^{-1}\tfrac{2(\nu-1)}{3\nu-1}\sum_k C_{kk,0}\Big(\sum_k\lambda_k - (\sum_k\lambda_k)^{-1}\tfrac{2(\nu-1)}{3\nu-1}\sum_k\lambda_k^2\Big)\Big]$$

$$= \phi\Big[\nu\sum_k\lambda_k\sum_k C_{kk,0}\lambda_k - (\nu-1)\sum_k C_{kk,0}\sum_k\lambda_k^2\Big], \tag{203}$$

proving statement 2. □

In Figure 4 we illustrate this result for a range of models with exact (not asymptotic) power laws (24) at different ν and $\varkappa$. In this case, in the noise-dominated regime, characteristic (29) is positive for sufficiently large $\varkappa$ but negative for smaller $\varkappa$. In agreement with Proposition 9, we observe experimentally found optimal β to be negative or positive, respectively.

F EXPERIMENTS

F.1 TRAINING REGIMES

We have considered four regimes of training with the increasing level of approximation for real network dynamics:

1. **Real neural network (*NN*).**
   We used fully-connected neural network with one hidden layer in the NTK parametrization:

   $$f(\mathbf{x},\mathbf{w}) = \frac{1}{\sqrt{N}}\sum_{l=1}^{d}\mathbf{w}_l^{(2)}\text{ReLU}(\mathbf{w}_l^{(1)} + b_l)$$

   Here $\mathbf{w}_l^{(1,2)}$ are the weight matrices, b_l are the biases for neuron l, and d is the number of units in the hidden layer. We have used $D = 1000$ in our experiments. For the sake of simplicity, the network outputs a single value for each input, i.e the classification task is treated as regression to the label with the MSE loss. The model is trained via SGD in the standard way: at each step a subset of size b is sampled without repetition from the training dataset and the weights are updated according to the equation (1).

2. **Linearized regime with sampled batches (*linearized* or *sampled*).**
   Model outputs $f(\mathbf{x})$ evolve on the dataset of interest according to the equation (46). At each step a batch of size b is sampled without replacement from the whole dataset of size N and SGD step is performed.

3. **All second matrix moments dynamics (*all sm*).**
   We simulate the dynamics of the whole second moment matrix that is obtained after taking expectation over all possible noise configurations, defined by (57) with the noise term given in the last line of (59).
4. **spectrally expressible regime (*SE*).**
   In the spectrally expressible regime, the diagonal of the second moment matrix is invariant under SGD dynamics, i.e diagonal elements evolve independently from off-diagonal terms. Therefore, one needs to keep track only of the diagonal terms. The dynamics in the output space follows Eq. (64).

F.2 DATASETS

In the experiments we have used synthetic data with the NTK spectrum and second moment matrix obeying a power-law decay rule $\lambda_k = k^{-\nu}$ and $\lambda_k C_{kk,0} = k^{-\varkappa-1}$ with specific ν and $\varkappa$ and the subset of size N of the MNIST dataset. The whole train dataset of size 60000 requires much memory for the computation of NTK and is too costly from computational side for running large series of experiments, therefore we have used only part of the data of size 1000 – 10000 depending on the experiment. The digits are distributed uniformly in the dataset, i.e there are $\frac{N}{10}$ samples of each class. We have observed that this subset of data is sufficient to approximate the power-law asymptotic for NTK and partial sums since the exponents do not change significantly with the increase of the amount of data.

F.3 EXPERIMENTAL DETAILS

For Fig. 1 one we have taken a subset of $N = 3000$ digits from MNIST and ran 10^4 SGD updates for the real neural network with one hidden layer, linearized dynamics, and in the spectrally expressible appoximation with $\tau = \pm 1$. One can see that $\tau = 1$ provides more accurate approximation for smaller batch size $b = 10$, whereas for $b = 100$ both $\tau = \pm 1$ are rather accurate.

For Fig. 2 we ran 10000 gradient descent updates for every point on the grid of learning rates α and β, where we have taken 100 values of α uniformly selected on $[0, 4]$ and 50 values for β uniformly on $[0, 1]$. The grey regions on the 2d plots correspond to the values of α and β for which the optimization procedure has diverged. The size of the dataset is 1000 both in the synthetic and MNIST case.

For Figs 5, 4 (left, center and right) we have taken the dataset of size $N = 10000$ both for synthetic and MNIST case and performed 10000 steps of gradient descent. In Fig. 4 (center) we have taken 100 points uniformly on the interval $\alpha \in [0, 0.5]$, and 100 points uniformly on the interval $\beta \in [-0.3, 0.3]$ for Fig. 4 (right). Concerning the correpspondence of optimal learning rates 4 with the analytical formula (186) we observe that is quite accurate for large ratio κ/ν since the ratio $\frac{C_{\text{noise}}}{C_{\text{signal}}} \simeq 100$ for $(\nu, \varkappa) = (1.5, 3.5), (2.0, 5.0)$, whereas for $\frac{C_{\text{noise}}}{C_{\text{signal}}} \simeq 1$ for $(\nu, \varkappa) = (1.5, 2.5), (2.0, 3.5)$, therefore both the signal and noise asymptotic are of same order for quite a long time, since the difference between the 'signal' exponent ξ and 'noisy' $2 - \frac{1}{\nu}$ is relatively small.

F.4 VALIDITY OF SPECTRALLY EXPRESSIBLE APPROXIMATION

In section 2 we introduced several scenarios where SE approximation is exact. However, more interesting and at the same time much more difficult question is how well SE approximation well suited to describe the practical problems, where we are given a particular Jacobian $\mathbf{\Psi}$, initial state $\mathbf{C}_0$ and then generate the sequence of states $\mathbf{C}_t$ according to (4) with true noise term (5).

We can approach this question in a different ways. First approach is to consider individually the true dynamics with (5) and approximated dynamics with noise term (7). Then we can look at certain quantities of practical interest and check whether they are close to each other. In figure 2 we compared stability regions in (α, β) plane for MNIST and found a good agreement between actual dynamics and SE approximation with $\tau_1 = 1, \tau_2 = 1$. In figure 1 we did it for the loss values on MNIST, but instead of noise-averaged loss values we compared with loss on a single stochastic trajectory. To make the comparison for loss values more thorough, in figure F.4 we add average of

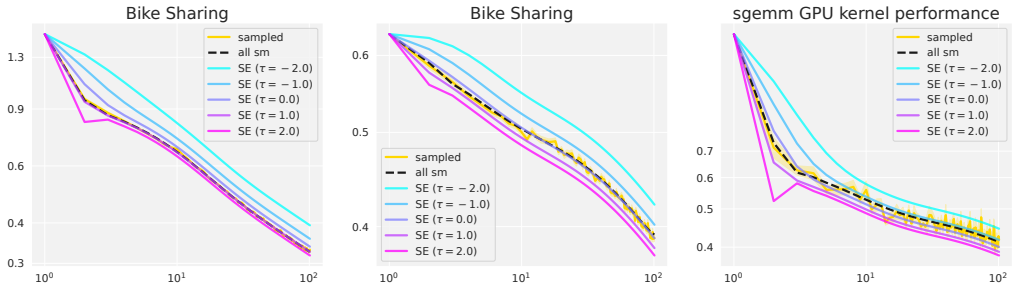

Figure 6: Loss trajectories for linearized stochastic loss dynamics (*solid yellow*), exact simulation of the loss dynamics with full second matrix (*dashed black*) (54), and the loss trajectories for several values of τ (shown with different colors). Linearized dynamics with stochastic sampling of batches is averaged over 1000 runs.

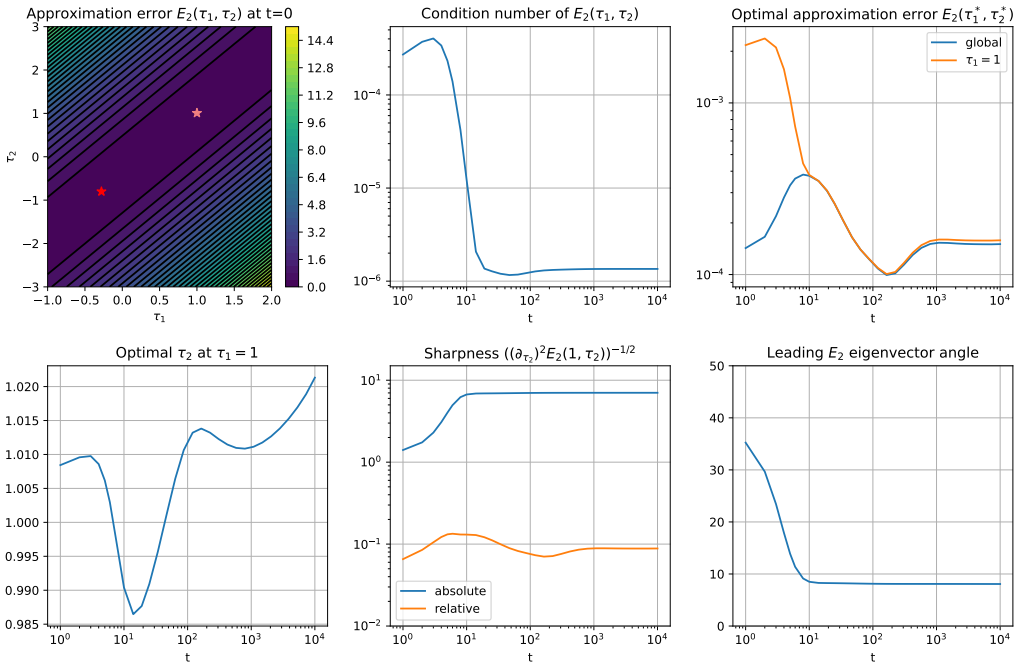

Figure 7: Various characteristics of relative approximation error $E_2(\tau_1, \tau_2)$ for SGD optimization on MNIST dataset. **Left top:** Dependence of $E_2(\tau_1, \tau_2)$ on τ_1, τ_2 at initialization. Red star denotes the global minimum, while light red star denotes the minimum on the line $\tau_1 = 1$. **Center top:** The ratio of smaller to bigger eigenvalue of the 2×2 matrix $\mathbf{E}$ which determines purely quadratic part of $E_2(\tau_1, \tau_2)$. **Right top:** Comparison of globally optimal approximation errors and approximation error optimal on the line $\tau_1 = 1$. **Left bottom:** Value τ_2 minimizing $E_2(\tau_1, \tau_2)$ on the line $\tau_1 = 1$. **Center bottom:** length scale on line $\tau_1 = 1$ on which either approximation error increases by 1 (*absolute*) when moving from optimum τ_2^*, or increases by optimal approximation error (*relative*) when moving from optimum τ_2^*. **Right bottom:** Angle between τ_2 axis and leading egenvector of $\mathbf{E}$.

several stochastic trajectories (*sampled*), true second moments dynamics (all sm), and several SE approximations with $\tau_1 = 1$ but different τ_2. For MNIST we find excellent agreement with $\tau_2 = 1$, for Bike Sharing dataset agreement is good with $\tau_2 = 0$, and for GPU performance dataset the agreement is worse with optimal τ_2 changing during optimization from 0 to -1.

The second approach is to run true dynamics of second moments, and on each step compare how true noise covariance $\mathbf{\Sigma}_t$ given by (5) with SE noise covariance $\mathbf{\Sigma}_t^{SE}(\tau_1, \tau_2)$, which we take to be given by

$$\mathbf{\Sigma}_t^{SE}(\tau_1, \tau_2) = \tau_1 \mathbf{H} \operatorname{Tr}[\mathbf{H}\mathbf{C}_t] - \tau_2 \mathbf{H}\mathbf{C}_t\mathbf{H} \tag{204}$$

We measure closeness of true and approximated noise covariances with relative quadratic trace distance

$$E_2(\tau_1, \tau_2) = \frac{\operatorname{Tr}\left[\left(\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^{SE}(\tau_1, \tau_2)\right)^T \left(\mathbf{\Sigma}_t - \mathbf{\Sigma}_t^{SE}(\tau_1, \tau_2)\right)\right]}{\operatorname{Tr}\left[\mathbf{\Sigma}_t^T \mathbf{\Sigma}_t\right]} \tag{205}$$

Note that linearity of $\mathbf{\Sigma}_t^{SE}(\tau_1, \tau_2)$ implies that $E_2(\tau_1, \tau_2)$ is quadratic in τ_1, τ_2. We analyze quadratic form $E_2(\tau_1, \tau_2)$ in Figure F.4. From the bottom left part we see that the ellipses corresponding to level sets of $E_2(\tau_1, \tau_2)$ are extremely stretched, meaning that there exist a direction along which approximation error changes very weekly. To check that conclusion for during the whole SGD optimization, we plot in center top part the ratio of smaller to bigger principal axes of level sets ellipses and see that during optimization the stretching only increases. One particular consequence is that there is a ambiguity in choosing the optimal τ_1, τ_2, which means that we can safely put $\tau_1 = 1$ and vary only τ_2. This conclusion is illustrated on right top part where both global and on $\tau_1 = 1$ optimal aprroximation error are very small $\lesssim 10^{-3}$ during the whole optimization. Finally, we observe left bottom part of F.4 that optimal τ_2 is extremely close to 1, which well agrees with the conclusion of the first approach based on predicting the evolution of loss values and stability regions.

F.5 Additional experiments

In addition to the experiments with MNIST and synthetic data we ran experiments with real neural networks and with the approximate dynamics for two more datasets from UCI Machine Learning Repository: Bike Sharing dataset [3] and SGEMM GPU kernel performance dataset [4]. The former is a regression problem, namely, prediction of the count of rental bikes per hour or per day, and we've selected the per hour dataset. There are 17389 instances in the dataset and 16 features per input sample. The latter is a regression problem as well, there are 14 features, describing the different properties of the computation hardware and software and the outputs are the execution times of 4 runs of matrix-matrix product for two matrices of size 2048×2048. We predict the average of four runs. There are 241600 instances in the dataset. See results in Figures 8 and 9.

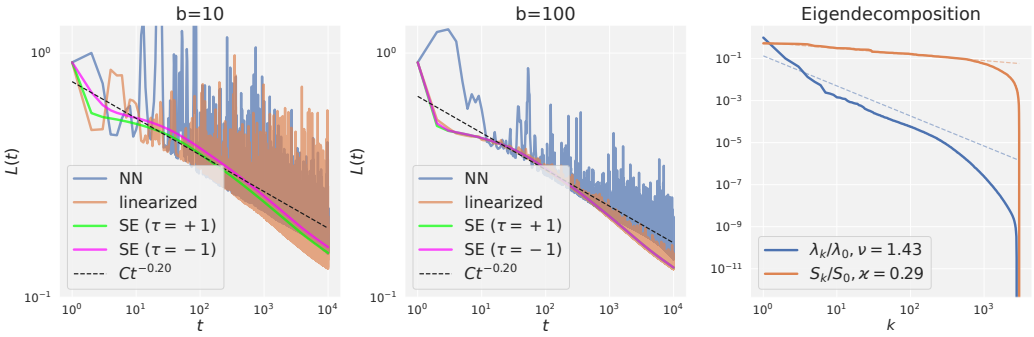

Figure 8: Loss trajectories and spectrum for Bike Sharing dataset. All notations follow Fig. 1. The power-law exponents are $\nu \simeq 1.43, \varkappa \simeq 0.29, \zeta \simeq 0.20$

In order to inverstigate the problem of the loss convergence and determination of the critical learning rate we performed experiments with different batch sizes $b = 1, 10, 1000, 1000$ on MNIST and synthetic data with $\nu = 1.5, \varkappa = 0.75$ (the 'signal' regime) and with h $\nu = 1.5, \varkappa = 3.0$ (the 'noisy' regime). For each value of momentum β we plot the critical value of learning rate α_{crit} such that the dynamics starts to diverge. See results in Figures 10, 11 and 12.

[3] https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset
[4] https://archive.ics.uci.edu/ml/datasets/SGEMM+GPU+kernel+performance

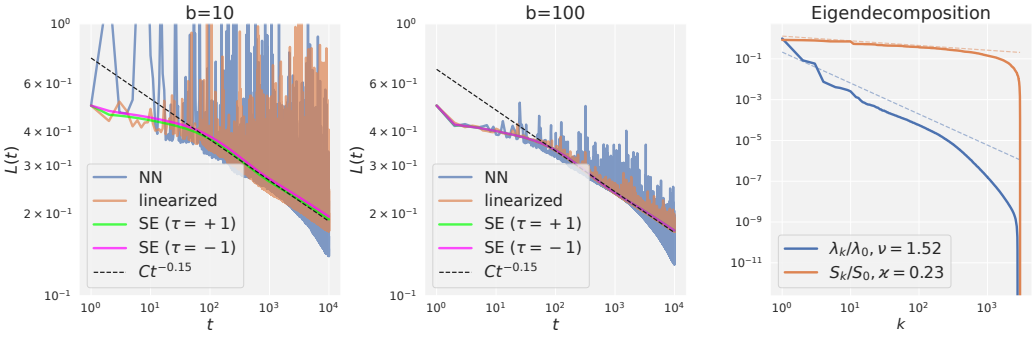

Figure 9: Loss trajectories and spectrum for Sgemm GPU kernel performance dataset. All notations follow Fig. 1. The power-law exponents are $\nu \simeq 1.52, \varkappa \simeq 0.23, \zeta \simeq 0.15$

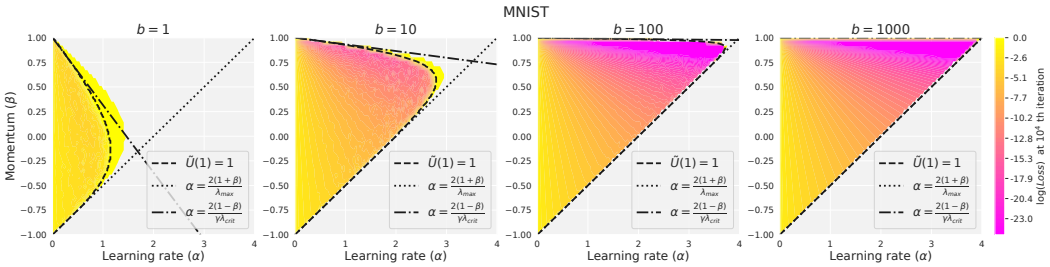

Figure 10: $L(t = 10^4)$ for different learning rates α, momenta β and batch sizes b on MNIST. Legend and notations follow Fig. 2

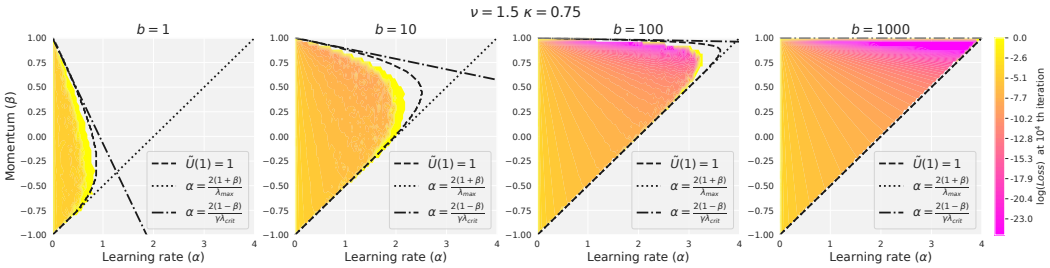

Figure 11: $L(t = 10^4)$ for different learning rates α, momenta β and batch sizes b on synthetic dataset with $\nu = 1.5$ and $\varkappa = 0.75$. Legend and notations follow Fig. 2

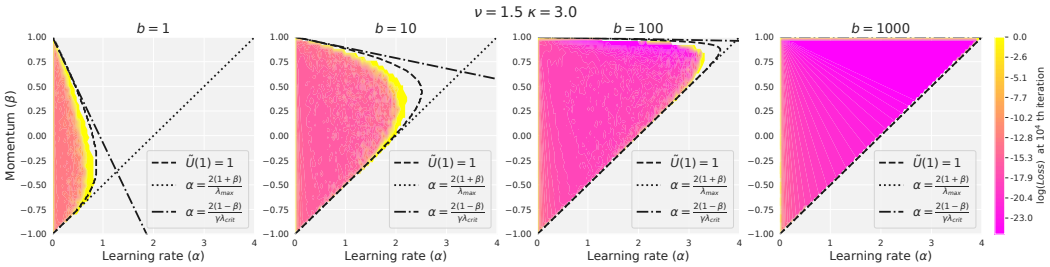

Figure 12: $L(t = 10^4)$ for different learning rates α, momenta β and batch sizes b on synthetic dataset with $\nu = 1.5$ and $\varkappa = 3.0$. Legend and notations follow Fig. 2

G MINI-BATCH SGD VS. SGD WITH ADDITIVE NOISE

In the paper we considered SGD with the natural noise structure induced by random sampling of batches. However, it is instructive to consider SGD with an additive noise model and observe how this surrogate noise model qualitatively changes optimization. Specifically, we will assume that on each iteration, a Gaussian vector with covariance $\mathbf{G}$ is added to the true loss gradient: $\nabla_{\mathbf{w}}^{(t)} L(\mathbf{w}) = \nabla_{\mathbf{w}} L(\mathbf{w}) + \mathbf{g}_t,\ \mathbf{g}_t \sim \mathcal{N}(0, \mathbf{G})$. For quadratic problems the SGD iteration becomes

$$\begin{pmatrix} \Delta \mathbf{w}_{t+1} \\ \mathbf{v}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \Delta \mathbf{w}_t \\ \mathbf{v}_t \end{pmatrix} - \alpha_t \begin{pmatrix} \mathbf{g}_t \\ \mathbf{g}_t \end{pmatrix} \tag{206}$$

Then, due to $\mathbb{E}[\mathbf{g}_t] = 0$, the dynamics of first moments is identical to that of noiseless GD, like in the case of sampling noise. However, dynamics of second moments is different:

$$\mathbf{M}_{t+1} = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix} \mathbf{M}_t \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix}^T + \alpha_t^2 \mathbf{G} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{207}$$

Considering again constant learning rates $\alpha_t = \alpha, \beta_t = \beta$, we see that by making a linear shift of second moments we get the noiseless dynamics:

$$\mathbf{M}_{t+1} - \mathbf{M}^\infty = \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix} \left(\mathbf{M}_t - \mathbf{M}^\infty\right) \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix}^T, \tag{208}$$

where $\mathbf{M}^\infty$ is the "residual uncertainty" determined from the linear equation

$$\mathbf{M}^\infty - \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix} \mathbf{M}^\infty \begin{pmatrix} \mathbf{I} - \alpha_t \mathbf{H} & \beta_t \mathbf{I} \\ -\alpha_t \mathbf{H} & \beta_t \mathbf{I} \end{pmatrix}^T = \alpha^2 \mathbf{G} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{209}$$

For simplicity, let's restrict ourselves to SGD without momentum ($\beta = 0$). Then in the eigenbasis of $\mathbf{H}$ we have

$$C_{kk,t+1} - C_{kk}^\infty = (1 - \alpha\lambda_k)^2 (C_{kk,t} - C_{kk}^\infty), \quad C_{kk}^\infty = \frac{\alpha}{\lambda_k(2 - \alpha\lambda_k)} G_{kk} \tag{210}$$

Thus, optimization converges with the speed of noiseless GD, but the limiting loss is nonzero:

$$\lim_{t\to\infty} L(\mathbf{w}_t) = L^\infty = \frac{1}{2} \sum_k \lambda_k C_{kk}^\infty > 0. \tag{211}$$

The key difference of SGD with sampling noise (4), (5) is the multiplicative character of the noise, in the sense that it vanishes as the optimization converges to the true solution and $\boldsymbol{M}_t$ becomes close to 0. To summarize, the two largest differences between SGD with additive and multiplicative noises are

1. Loss converges to a value $L^\infty > 0$ for additive noise and to 0 for multiplicative noise.
2. Convergence rate is independent of noise strength for additive noise, but significantly depends on it for multiplicative noise.

H NOISELESS CONVERGENCE CONDITION

The convergence condition $\alpha < \frac{2(1+\beta)}{\lambda_{\max}}$ for noiseless GD is known (see e.g. (Tugay & Tanik, 1989)). For completeness, below we present its derivation in our setting.

First observe that, in the noiseless GD, different eigenspaces of $\mathbf{H}$ evolve independently from each other. Thus we need to separately require convergence in each eigenspace of $\mathbf{H}$. Denoting the eigenvalue of the chosen subspace by λ, we write the GD iteration as

$$\begin{pmatrix} \Delta \mathbf{w}_{t+1} \\ \mathbf{v}_{t+1} \end{pmatrix} = \mathbf{S} \begin{pmatrix} \Delta \mathbf{w}_{t+1} \\ \mathbf{v}_{t+1} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 1 - \alpha\lambda & \beta \\ -\alpha\lambda & \beta \end{pmatrix} \tag{212}$$

The characteristic equation for eigenvalues s of matrix $\mathbf{S}$ is

$$s^2 - (1 - \alpha\lambda + \beta)s + \beta = 0, \tag{213}$$

yielding solutions

$$s_{1,2} = \frac{1 - \alpha\lambda + \beta \pm \sqrt{(1 - \alpha\lambda + \beta)^2 - 4\beta}}{2} \tag{214}$$

The GD dynamics converges iff $|s_1| < 1$ and $|s_2| < 1$. First note that $|s_1||s_2| = |\beta|$ and therefore for $|\beta| \geq 1$ at least one of $|s_{1,2}|$ will be ≥ 1 and GD will not converge. Therefore we only need to consider $\beta \in (-1, 1)$.

When $\beta \in \big((1 - \sqrt{\alpha\lambda})^2, (1 + \sqrt{\alpha\lambda})^2\big)$, the eigenvalues $s_{1,2}$ are complex and $|s_1| = |s_2| = \sqrt{|\beta|}$. The convergence condition in this region is $|\beta| < 1$ which is automatically satisfied.

Now consider the region where $s_{1,2}$ are real. As $s_1 > s_2$, the convergence conditions in this region are $s_1 < 1$ and $s_2 > -1$. Note that for $\beta \leq 0$ the eigenvalues $s_{1,2}$ are always real, while for $\beta > 0$ they are real for $\lambda \in [0, \infty) \setminus (\lambda_1, \lambda_2)$ with $\lambda_1 = (1 - \sqrt{\beta})^2/\alpha$ and $\lambda_2 = (1 + \sqrt{\beta})^2/\alpha$.

Let us first check the condition $s_1 < 1$. The equation $s_1(\lambda) = 1$ gives $\lambda = 0$. For $\beta > 0$ we have $s_1(\lambda = 0) = 1$, $s_1(\lambda = \lambda_1) = \sqrt{\beta} < 1$ and $s_1(\lambda = \lambda_2) = -\sqrt{\beta} < 1$ and therefore $s_1(\lambda) < 1$ on $(0, \infty) \setminus (\lambda_1, \lambda_2)$ due to continuity of $s_1(\lambda)$. For $\beta < 0$ the condition $s_1(\lambda) < 1$ is always satisfied because $s_1(\lambda)$ is decreasing with λ.

Proceeding to condition $s_2 > -1$ we find that equation $s_2(\lambda^\star) = -1$ gives $\alpha\lambda^\star = 2(\beta + 1)$. Note that $\lambda^\star > \lambda_2$. For $\beta > 0$ we have $s_2(\lambda = 0) = \beta > -1$, $s_2(\lambda = \lambda_1) = \sqrt{\beta} > -1$, $s_2(\lambda = \lambda_2) = -\sqrt{\beta} > -1$ and $s_2(\lambda \to \infty) \to -\infty$. Therefore, due to continuity of $s_2(\lambda)$, we have $s_2(\lambda) > -1$ on $(0, \lambda^\star)/(\lambda_1, \lambda_2)$ and $s_2(\lambda) < -1$ on $(\lambda^\star, \infty)$. Similarly, for $\beta < 0$ we again have $s_2(\lambda) > -1$ on $(0, \lambda^\star)$ and $s_2(\lambda) < -1$ on $(\lambda^\star, \infty)$

Combining all the observations above, we get a single convergence condition for non-stochastic GD:

$$\alpha < \frac{2(1 + \beta)}{\lambda_{\text{max}}}, \quad \beta \in (-1, 1) \tag{215}$$

where λ_{max} is the largest eigenvalues of $\mathbf{H}$.