FLATTER, FASTER: SCALING MOMENTUM FOR OPTIMAL SPEEDUP OF SGD

Anonymous authors
Paper under double-blind review

ABSTRACT

Commonly used optimization algorithms often show a trade-off between good generalization and fast training times. For instance, stochastic gradient descent (SGD) tends to have good generalization; however, adaptive gradient methods have superior training times. Momentum can help accelerate training with SGD, but so far there has been no principled way to select the momentum hyperparameter. Here we study implicit bias arising from the interplay between SGD with label noise and momentum in the training of overparametrized neural networks. We find that scaling the momentum hyperparameter $1-\beta$ with the learning rate to the power of $2/3$ maximally accelerates training, without sacrificing generalization. To analytically derive this result we develop an architecture-independent framework, where the main assumption is the existence of a degenerate manifold of global minimizers, as is natural in overparametrized models. Training dynamics display the emergence of two characteristic timescales that are well-separated for generic values of the hyperparameters. The maximum acceleration of training is reached when these two timescales meet, which in turn determines the scaling limit we propose. We perform experiments, including matrix sensing and ResNet on CIFAR10, which provide evidence for the robustness of these results.

1 INTRODUCTION

The modern paradigm for optimization of deep neural networks has practitioners working with vastly overparametrized models and training to near perfect accuracy (Zhang et al., 2017). In this setting, a model will typically have not just isolated minima in parameter space, but a continuous set of minimizers, not all of which generalize well. As demonstrated in Liu et al. (2020), depending on parameter initialization and hyperparameters, stochastic gradient descent (SGD) is capable of finding minima with wildly different test accuracies. Thus, the power of a particular optimization method lies in its ability to select a minima amongst this vast set that also generalizes. In other words, good generalization relies on the *implicit bias or regularization* of an optimization algorithm.

There is plenty of evidence that training deep nets with SGD leads to good generalization. Intuitively, SGD appears to prefer flatter minima (Keskar et al., 2017; Wu et al., 2018), and flatter minima are better for generalization (Hochreiter & Schmidhuber, 1997). More recently, a variant of SGD which introduces "algorithmic" label noise has been especially amenable to rigorous treatment. Indeed Blanc et al. (2020) were able to rigorously determine the implicit regularization of SGD with label noise in the phase of learning after training loss has converged to a minima. In the overparametrized setting, they show that SGD converges not just to any minimum, but to those minima that lead to the smallest trace norm of the Hessian. However, Li et al. (2022) show that the dynamics of this regularization happen on a timescale proportional to the inverse square of the learning rate η – much slower than the time to converge to the zero loss manifold.

With the recent explosion in size of both models and datasets, training time becomes a question nearly as important as asymptotic generalization error. In this context, adaptive gradient methods such as Adam (Kingma & Ba, 2015) are unilaterally preferred over variants of SGD, even though they are often found to give worse generalization error in practical settings (Keskar & Socher, 2017; Wilson et al., 2017) (though extensive hyperparameter tuning Choi et al. (2019) or scheduling Xie et al. (2022) can potentially obviate this problem).

This motivates a search for methods to accelerate SGD. Classic work on acceleration methods, which we refer to generally as momentum, have found a provable benefit in the deterministic setting, where gradient updates have no error. However, rigorous guarantees have been harder to find in the stochastic setting, and when they do exist they tend to require strict conditions on the noise (Polyak, 1987; Kidambi et al., 2018).

In this work, we show that there exists a scaling limit for SGD with momentum (SGDM) which provably increases the rate of convergence.

1.1 Heuristic Explanation for Optimal Momentum-Based Speedup

Deep neural networks typically posses a manifold of parametrizations with zero training error. Because the gradients of the loss function vanish along this manifold, the dynamics of the weights is completely frozen under gradient descent. However, as appreciated by Wei & Schwab (2019) and Blanc et al. (2020), the effect of SGD noise can generate an average drift of the weights along the manifold, driving them to a lower-curvature region which, heuristically, explains the good generalization properties of SGD. Separately, it is well-known that adding momentum typically leads to acceleration in training Sutskever et al. (2013). We are interested in understanding the nature of this acceleration, and providing principled guidance for momentum hyperparameter tuning. In this section we will illustrate the main idea of this paper, showing how speedup towards flatter minima arises from the interaction of noise and momentum, and how this acceleration can be maximized by an appropriate hyperparameter scaling.

We model momentum SGD using the following equations

$$\pi_{k+1} = \beta\pi_k - \nabla L(w_k) + \sigma(w_k)\xi_k, \qquad w_{k+1} = w_k + \eta\pi_{k+1}, \tag{1}$$

where $\sigma(w) \in \mathbb{R}^{D\times r}$ is the noise function evaluated at $w \in \mathbb{R}^D$, and $\xi_k \in \mathbb{R}^r$ is a noise vector drawn i.i.d. at every timestep k with zero mean and unit variance. We will now form a heuristic understanding of the dynamics that happens along a manifold of minimizers $\Gamma = \{w : L(w) = 0\}$. Consider a point $w \in \Gamma$. The first non-zero order in the Taylor expansion of L at w will have a quadratic term given by the Hessian, $H(w)$, which produces dynamics only in the space transverse to Γ. This justifies assuming that the timescale associated to tangent displacements to Γ, which we shall call τ_2, is much longer than that associated to perpendicular displacements, τ_1. More explicitly, denote the evolution of the weights near the zero-loss manifold as $w_k = w_0 + v_k^T + v_k^L$ where $w_0 \in \Gamma$ is the initial value, and v_k^T and v_k^L are perpendicular and longitudinal displacements. Because the motion along Γ is slow, we can approximately draw v^T from a stationary distribution. We find the slow dynamics along Γ by noting that

$$v_{k+1}^L - v_k^L = \eta P_L \pi_{k+1} \approx \eta P_L \langle \pi_{k+1} \rangle \tag{2}$$

where P_L is the projector onto the tangent space of Γ at w_0, and we replaced the momentum by its expectation over noise $\langle \pi_{k+1} \rangle$. We can solve for π_k and use our approximation to write it in terms of the second moments of the transverse displacements:

$$\begin{aligned} P_L\langle \pi_{k+1}\rangle &= P_L \sum_{j=0}^{k} \beta^{k-j} \left\langle \sigma(w_j)\xi_j - \nabla L(w_j)\right\rangle \\ &\approx P_L \sum_{j=0}^{k} \beta^{k-j} \left\langle \partial(\nabla L)[v_k^T] + \tfrac{1}{2}\partial^2(\nabla L)[v_j^T, v_j^T]\right\rangle \approx \frac{1}{2(1-\beta)} P_L \partial^2(\nabla L)\left[\langle v_k^T, v_k^T\rangle\right] \end{aligned} \tag{3}$$

where square brackets denote evaluation of derivatives in the directions given by the vectors or matrices (see end of Sec.3.1 for precise definition). In going from the first to the second line, we assumed that $k \ll \tau_2$ so that the longitudinal displacement of the weights is negligible, and that the perpendicular displacements are small. In the last line, we used stationarity of the distribution of v_k^T. One can show that $\langle (v_k^T)^T v_k^T\rangle = \Theta\left(\frac{\eta}{1-\beta}\right)$ by computing the second moments and imposing that they do not change in time, and thus $v_{k+1}^L - v_k^L = \Theta(\frac{\eta^2}{(1-\beta)^2})$. By imposing the scaling

$$\beta = 1 - C\eta^\gamma \tag{4}$$

we find $v_{k+1}^L - v_k^L = \Theta(\eta^{2(1-\gamma)})$, and hence the timescale for the weights to move longitudinally an $\Theta(1)$ amount is $\tau_2 = \Theta(\eta^{2(\gamma-1)})$. We will derive this result rigorously in Sec. 3.2. On the other hand, the timescale τ_1 controls the convergence to the valley Γ, and is dominated by the noiseless dynamics. As shown in Sec. 3.3, this is $\tau_1 = \Theta(\eta^{\min\{\gamma,1-\gamma\}})$. The key observation now is that the total training time is essentially controlled by the largest timescale among τ_1 and τ_2, therefore the optimal speedup of training is achieved when these two timescales approach each other. Comparing the exponents, we thus find that $\gamma = 2/3$ is the optimal scaling in Eq. (4).

The **main contributions** of this paper are:

1. We develop a general formalism to study SGD with (heavy-ball) momentum in Sec.(3), extending the framework of Li et al. (2022) to study convergence rates and generalization with momentum.
2. We find a novel scaling regime of the momentum hyperparameter $1-\beta \sim \eta^\gamma$, and demonstrate a qualitative change in the noise-induced training dynamics as γ is varied.
3. We identify a special scaling limit, $1-\beta \sim \eta^{2/3}$, where training achieves a *maximal speedup at fixed learning rate* η.
4. In Sec.(4), we demonstrate the relevance of our theory with experiments on toy models (2-layer neural networks, and matrix sensing) as well as realistic models and datasets (ResNet-18 on CIFAR10).

2 Related Works

Loss Landscape in Overparametrized Networks The geometry of the loss landscape is very hard to understand for real-world models. Choromanska et al. (2015) conjectured, based on empirical observations and on an idealized model, that most local minima have similar loss function values. Subsequent literature has shown in wider generality the existence of a manifold connecting degenerate minima of the loss function, particularly in overparametrized models. This was supported by work on mode connectivity (Freeman & Bruna, 2017; Garipov et al., 2018; Draxler et al., 2018; Kuditipudi et al., 2019), as well as on empirical observations that the loss Hessian possesses a large set of (nearly) vanishing eigenvalues (Sagun et al., 2016; 2017). In particular, Nguyen (2019) showed that for overparametrized networks with piecewise linear activations, all global minima are connected within a unique valley.

The Implicit Regularization of SGD Wei & Schwab (2019), assuming the existence of a zero-loss valley, observed that SGD noise leads to a decrease in the trace of the Hessian. Blanc et al. (2020) demonstrated that SGD with label noise in the overparametrized regime induces a regularized loss that accounts for the decrease in the trace of the Hessian. Damian et al. (2021) extend this analysis to finite learning rate. HaoChen et al. (2021) study the effect of non-isotropic label noise in SGD and find a theoretical advantage in a quadratic overparametrized model. Wu et al. (2022) show that only minima with small enough Hessians (in Frobenius norm) are stable under SGD. The specific regularization induced by SGD was found in quadratic models (Pillaud-Vivien et al., 2022), 2-layer Relu networks (Blanc et al., 2020), linear models (Li et al., 2022), diagonal networks (Pesme et al., 2021). Additionally, Kunin et al. (2021) studied the limiting diffusive dynamics induced by SGD both empirically and in a simple theoretical model.

Momentum in SGD and Adaptive Algorithms Momentum is a general term applied to techniques introduced to speed up gradient descent. Popular implementations include Nesterov (Nesterov, 1983) and Heavy Ball (HB) or Polyak (Polyak, 1964). We focus on the latter in this paper, which we refer to simply as momentum. Momentum provably improves convergence time in the deterministic setting, but less is known rigorously when stochastic gradient updates are used. Indeed, Polyak (1987) suggests the benefits of acceleration with momentum disappear with stochastic optimization unless certain conditions are placed on the properties of the noise. See also (Jain et al., 2018; Kidambi et al., 2018) for more discussion and background on this issue. Nevertheless, in practice it is widely appreciated that momentum is important for convergence and generalization (Sutskever et al., 2013), and widely used in modern adaptive gradient algorithms Kingma & Ba (2015). Some limited results have been obtained showing speedup in the mean-field approximation

(Mannelli & Urbani, 2021) and linear regression Jain et al. (2018). Modifications to Nesterov momentum to make it more amenable to stochasticity (Liu & Belkin, 2020; Allen-Zhu, 2017), and near saddle points (Xie et al., 2022) have also been considered.

3 Theoretical Results

3.1 General setup

In this and the following section, we will rigorously derive the limiting drift-diffusion equation for the weights on the zero-loss manifold Γ, and extract the timescale τ_2 associated to this noise-induced motion. In Sec. 3.3 we will then compare this to the timescale τ_1 associated to the noiseless dynamics and evaluate the optimal value of γ discussed around Eq. (4). We will use Eq. (1) to model momentum SGD. The parameter $\epsilon > 0$ controls the size of the noise, and the results that follow will be proved for $\epsilon \to 0$. As illustrated around Eq. (3), the drift is controlled by the second moment of fluctuations, and we thus expect the drift timescale to be $\Theta(\epsilon^2)$. We will then rescale time $k = t/\epsilon^2$, so that the motion in the units of t is $O(1)$ as $\epsilon \to 0$. More explicitly, take ϵ_n to be a positive sequence such that $\epsilon_n \to 0$ as $n \to \infty$. For each n we consider the stochastic process that solves Eq. (1):

$$X_n(t) = X_n(0) + \int_0^t \tilde{\sigma}(X_n) dZ_n + \int_0^t F(X_n) dA_n \,, \tag{5}$$

with

$$A_n(t) = \left\lfloor \frac{t}{\epsilon_n^2} \right\rfloor, \qquad Z_n(t) = \epsilon_n \sum_{k=1}^{A_n(t)} \xi_k \tag{6}$$

and where $X(t = \epsilon^2 k) = (\pi_k, w_k)$, $\tilde{\sigma}(X) = (\sigma, \eta\sigma)$ and $F(X) = ((\beta - 1)\pi - \nabla L(w), \eta(\beta\pi - \nabla L(w)))$. $\lfloor x \rfloor$ denotes the integer part of a real number x.

Assumption 3.1. *The loss function $L : \mathbb{R}^D \to \mathbb{R}$ is a C^3 function whose first 3 derivatives are locally Lipschitz, σ is continuous, and $\Gamma = \{w \in \mathbb{R}^D : L(w) = 0\}$ is a C^2-submanifold of $\mathbb{R}^D$ of dimension M, with $0 \leq M \leq D$. Additionally, for $w \in \Gamma$, rank$(\nabla^2 L(w)) = D - M$.*

Assumption 3.2. *There exists an open neighborhood U of $\{0\} \times \Gamma \subseteq \mathbb{R}^D \times \mathbb{R}^D$ such that the gradient descent starting in U converges to a point $x = (\pi, w) \in \{0\} \times \Gamma$. More explicitly, for $x \in U$, let $\psi(x, 0) = x$ and $\psi(x, k+1) = \psi(x, k) + F(\psi(x, k))$, i.e. $\psi(x, k)$ is the kth iteration of $x + F(x)$. Then $\Phi(x) \equiv \lim_{k\to\infty} \psi(x, k)$ exists and is in Γ. As a consequence, $\Phi \in C^2$ on U (Falconer, 1983).*

In what follows, we denote by C^n, for $n = 0, 1, \ldots$ the set of functions with continuous n^{th} derivatives. For any function f, $\partial f[u]$ and $\partial^2 f[u, v]$ will denote directional first and second derivatives along directions defined by vectors $u, v \in \mathbb{R}^D$. We may occasionally also write $\partial^2 f[\Sigma] = \sum_{i,j=1}^D \partial^2 f[e_i, e_j]\Sigma_{ij}$. Given a submanifold $\Gamma \subset \mathbb{R}^D$ and $w \in \Gamma$, we denote by $T_w\Gamma$ the tangent space to Γ in w, and by $P_L(w)$ the projector onto $T_w\Gamma$ (we will often omit the dependence on w and simply write P_L). Given a matrix $H \in \mathbb{R}^D \times \mathbb{R}^D$, we will denote by $H^\top$ the transpose, and $H^\dagger$ the pseudoinverse of H.

3.2 Limiting drift-diffusion in momentum SGD

In this section we shall obtain the explicit expression for the limiting drift-diffusion. The general framework is based on Katzenberger (1991) (reviewed in Appendix B). Before stating the result, we will need to introduce a few objects.

Definition 3.3. For a symmetric matrix $H \in \mathbb{R}^D \times \mathbb{R}^D$, and $W_H = \{\Sigma \in \mathbb{R}^D \times \mathbb{R}^D : \Sigma = \Sigma^\top, HH^\dagger\Sigma = H^\dagger H\Sigma = \sigma\}$, we define the operator $\tilde{\mathcal{L}}_H : W_H \to W_H$ with $\tilde{\mathcal{L}}_H S \equiv \{H, S\} + \frac{1}{2}C^{-2}\eta^{1-2\gamma}[[S, H], H]$, with $[S, H] = SH - HS$. It can be shown that the operator $\tilde{\mathcal{L}}_H$ is invertible (see Lemma C.3).

Consider the process in Eq. (5). Note that, while at initialization we can have $X_n(0) \notin \Gamma$, the solution $X_n(t) \to \Gamma$ as $n \to \infty$, i.e. it becomes discontinuous. This is an effect of the speed-up of time introduced around Eqs. (5),(6). To overcome this issue, it is convenient to introduce $Y_n(t) \equiv X_n(t) - \psi(X_n(0), A_n(t)) + \Phi(X_n(0))$, so that $Y_n(0) \in \Gamma$ is initialized on the manifold.

Theorem 3.4 (Informal). *Suppose the loss function L, the noise function σ, the manifold of minimizers Γ and the neighborhood U satisfy assumptions (3.1) and (3.2), and that $X_n(0) = X(0) \in U$. Then, as $\epsilon_n \to 0$, $Y_n(t)$ converges to $Y(t)$, where the latter satisfies the limiting drift-diffusion equation*

$$\begin{aligned} dY =& (\tfrac{1}{C}\eta^{1-\gamma} + \eta) P_L \sigma dW - \tfrac{1}{2C^2}\eta^{2-2\gamma}(\nabla^2 L)^\dagger \partial^2(\nabla L)[\Sigma_{LL}]dt \\ &- \tfrac{1}{C^2}\eta^{2-2\gamma} P_L \partial^2(\nabla L)[(\nabla^2 L)^\dagger \Sigma_{TL}]dt - \tfrac{1}{2C^2}\eta^{2-2\gamma} P_L \partial^2(\nabla L)[\tilde{\mathcal{L}}^{-1}_{\nabla^2 L}\Sigma_{TT}]dt\,, \end{aligned} \tag{7}$$

where $W(t)$ is a Wiener process.

The first term in Eq. (7) induces diffusion in the longitudinal direction. The second term is of geometrical nature, and is necessary to guarantee that $Y(t)$ remains on Γ. The second line describes the drift induced by the transverse fluctuations: it precisely corresponds the right-hand side of the heuristic Eq. (3).

Eq. (7) resembles in form that found in Li et al. (2022), although there are two crucial differences. First, time has been rescaled using the strength of the noise ϵ, rather than the learning rate. The different rescaling was necessary as the forcing term F in Eq. (5) depends non-homogeneously on η, and thus the theory of Katzenberger (1991) would not be directly applied had we taken the small learning rate limit. Second, and more crucially, the drift terms in Eq. (7) are proportional to $\eta^{2-2\gamma}$, which is a key ingredient leading to the change in hierarchy of the timescales discussed in Sec. 1.1. One final difference, is that the last term involves the operator $\tilde{\mathcal{L}}_H$ instead of the Lyapunov operator. For $\gamma < \frac{1}{2}$, $\tilde{\mathcal{L}}_H$ reduces to the Lyapunov operator $\mathcal{L}_H$ at leading order in η, with $\mathcal{L}_H S \equiv \{H, S\}$. For $\gamma > \frac{1}{2}$, however, we cannot neglect the η-dependent term in $\tilde{\mathcal{L}}_H$ (see discussion at the end of Appendix C).

Corollary 3.5. *In the case of label noise, i.e. when, for $w \in \Gamma$, $\Sigma = c\nabla^2 L$, for some constant $c > 0$, Eq. (7) reduces to*

$$dY = -\frac{\epsilon^2 \eta^{2-2\gamma}}{4C^2} P_L \nabla\, Tr(c\nabla^2 L)dt\,, \tag{8}$$

where we have rescaled time back to $t = k$, i.e. we performed $t \to t\epsilon^2$.

3.3 Separation of timescales and optimal momentum scaling

The above results provide an estimate for the timescale of the drift along the zero-loss valley. In this section, we look for a better understanding of the rate of convergence towards the valley. In this regime, the deterministic gradient flow dominates, so we study the simplified dynamics without noise. Our arguments are informal, and utilize the linearization of the dynamics near a fixed point to determine the set of characteristic timescales. Working in the extended phase space $x_t = (\pi_t, w_t)$, and linearizing Eq. (1) around a fixed point $\delta x_{t+1} = J(x^*)\delta x_t$, where $J(x^*)$ is the Jacobian evaluated at the fixed point (with the explicit form given in Eq. (115)). Denote by q^i the eigenvector and λ_i the corresponding eigenvalue of the Hessian. We show in Appendix E that The Jacobian is diagonalized by the eigenvectors $k^i_\pm = \left(\mu^i_\pm q^i, q^i\right)$ with eigenvalues

$$\kappa^i_\pm = \frac{1}{2}\left(1 + \beta - \eta\lambda_i \pm \sqrt{(1+\beta-\eta\lambda_i)^2 - 4\beta}\right), \tag{9}$$

and $\mu^i_\pm = \beta\eta\kappa^i_\pm - 1 - \eta\lambda_i$. We proceed to study the decay rate of the different modes of the Jacobian to draw conclusions about the characteristic timescales of fluctuations around the valley.

Longitudinal motion: On the valley, the Hessian will have a number of "zero modes" with $\lambda_i = 1$. These lead to two distinct modes in the present setting with momentum, which we distinguish as pure and mixed. The first **pure longitudinal mode** is an exact zero mode which has $\kappa^i_+ = 1$ with $k^i_+ = (0, q^i)$, corresponding to translations of the parameters along the manifold, and keeping $\pi = 0$ at its fixed point value. The second mode is a **mixed longitudinal mode** with $\kappa^i_- = \beta$ with $k^i_- = (-(1-\beta)/(2\beta\eta)q^i, q^i)$. This mode has a component of π along the valley, which must subsequently decay because the equilibrium is a single point $\pi = 0$. Therefore, this mode decays at the characteristic rate β for π, gleamed directly from Eq. (1).

Transverse motion: When the w and π are perturbed along the transverse directions q^i with positive λ_i, the relaxation behavior exhibits a qualitative change depending on β. Using the scaling function

$\beta = 1 - C\eta^\gamma$, for small learning rate, the spectrum is purely real for $\gamma < 1/2$, and comes in complex conjugate pairs for $\gamma > 1/2$. This leads to two distinct scaling behaviors for the set of timescales. Defining a positive $c_1 \leq \min\{\lambda_i | \lambda_i > 0\}$, we find: 1) For $\gamma < 1/2$, transverse modes are purely decaying as

$$(1 - C\eta^\gamma)^t \leq |\delta x_t^T| \leq (1 - (c_1/C)\eta^{1-\gamma})^t \,, \tag{10}$$

with the lower bound set by the mixed longitudinal mode. For $\gamma > 1/2$, the transverse modes are oscillatory but with an envelope that decays like

$$|\delta x_t^{T,env}| \approx (1 - C\eta^\gamma)^{t/2} \,. \tag{11}$$

We leave the derivation of these results to Appendix (E).

Collecting these results, we can describe the hierarchy of timescales τ_1 in the deterministic regime as a function of γ (excluding the pure longitudinal zero mode):

τ_1^{-1}	$\gamma < 1/2$	$\gamma > 1/2$
Long.	η^γ	η^γ
Transv.	$\eta^{1-\gamma}, \eta^\gamma$	η^γ, η^γ

These are illustrated schematically in Fig. 1(a), where the finite timescales are shown as a function of γ. We compare these "equilibration" timescales τ_1, i.e. characteristic timescales associated with relaxation back to the zero-loss manifold, with the timescale $\tau_2 \sim \eta^{2(\gamma-1)}$ associated with drift-diffusion of the noise-driven motion along the zero-loss manifold Eq. (7). For small γ, the timescale associated with the drift-diffusion along the valley is much faster than that associated with the relaxation of the dynamics toward steady state. Transverse and mixed longitudinal fluctuations relax much faster than the motion along the valley, and produce an effective drift toward the minimizer of the implicit regularizer. However, the timescales collide at $\gamma = 2/3$, suggesting a transition to a qualitatively different transport above this value, where the transverse and the mixed longitudinal dynamics, having a long timescale, will disrupt the longitudinal drift Eq. (7). This leads us to propose $\gamma = \frac{2}{3}$ as the optimal choice for SGD training. We consistently find evidence for such a qualitative transition in our experiments below. In addition, we see that speedup of SGD with label noise is in fact maximal at this value where the timescales meet.

3.4 A SOLVABLE EXAMPLE

In this section we analyse a model that will allow us to determine, on top of the optimal exponent $\gamma = \frac{2}{3}$, also the prefactor C. We will specify to a 2-layer linear MLP, which is sufficient to describe the transition in the hierarchy of timescales described above, and is simple enough to exactly compute C. We will show in Sec. 4.1 that C depends only mildly on the activation function. We apply a simple matching principle to determine C, by asking that the deterministic timescale τ_1 is equal to the drift-diffusion timescale τ_2. In the previous section, we found the critical $\gamma = 2/3$ by requiring these timescales have the same scaling in η. In order to determine C, we need more details of the model architecture.

Definition 3.6 (UV model)**.** We define the UV model as a 2-layer linear network parametrized by $f(x) = \frac{1}{\sqrt{n}} UVx \in \mathbb{R}^m$, where $x \in \mathbb{R}^d$, $V \in \mathbb{R}^{n \times d}$, and $U \in \mathbb{R}^{m \times n}$. For $d = m = 1$, we refer to this as the **vector UV model** (Rennie & Srebro, 2005; Saxe et al., 2014; Lewkowycz et al., 2020).

For a training dataset $\mathcal{D} = \{(x^a, y^a) \big| a = 1, ..., P\}$, the dataset covariance matrix is $\Sigma_{ij} = \frac{1}{P} \sum_{a=1}^P x_i^a x_j^a$, and the dataset variance is $\mu_2 = \mathrm{tr}\Sigma$. For mean-squared error loss, it is possible to explicitly determine the trace of the Hessian (see Appendix D). SGD with label noise introduces $y^a \to y^a + \delta\xi_t$ where $\langle \xi_t^2 \rangle = 1$, from which we identify $\epsilon = \delta/\sqrt{P}$ in Eq. (8), and $\sigma_{\mu,ja} = P^{-1/2} \nabla_\mu f_j(x^a)$, where μ runs over all parameter indices, $j \in [m]$, and $a \in [P]$. With this choice, the SGD noise covariance satisfies $\sigma\sigma^T = \nabla^2 L$ precisely. Equipped with this, we may use Corollary 3.5 with $c = 1$ to determine the effective drift (presented in Appendix D). For the vector UV model, the expression simplifies to $dY = -\tau_2^{-1} Y dt$, with

$$\tau_2^{-1} = \frac{\eta^{2-2\gamma} \delta^2 \mu_2}{2nPC^2} \,. \tag{12}$$

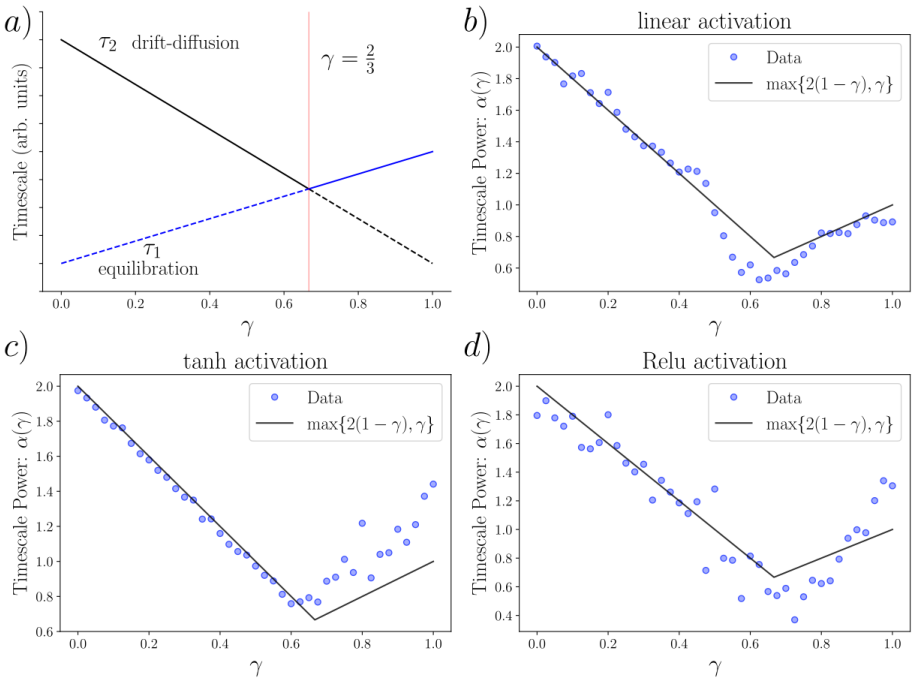

Figure 1: Timescale of training as a function of γ. $a)$: theoretical prediction with blue line representing the timescale for equilibration while the black line shows the timescale of the drift along Γ. The maximum of these two gives the overall timescale. In $b)$, $c)$, and $d)$ we demonstrate this result with the vector UV model with linear, $\tanh$ and Relu activations respectively.

The timescale of the fast initial phase $\tau_1^{-1} = (C/2)\eta^\gamma$ follows from the previous section. Then requiring $\tau_1 = \tau_2$ implies not only $\gamma = 2/3$, but

$$C = \left(\frac{\delta^2 \mu_2}{Pn}\right)^{1/3}. \tag{13}$$

One particular feature to note here is that C will be small for overparametrized models and/or training with large datasets.

4 Experimental Validation

4.1 2-layer MLP with Linear and Non-Linear Activations

The first experiment we consider is the vector UV model analyzed in Sec. 3.4. Our goal with this experiment is to analyze a simple model, and show quantitative agreement with theoretical expectations. Though simple, this model shows almost all of the important features in our analysis: the $2(1-\gamma)$ exponent below $\gamma = \frac{2}{3}$, the γ exponent above $\frac{2}{3}$, and the constant C theoretically evaluated in Sec. 3.4.

To this end, we extract the timescales at different values of γ and show them in Fig. 1. We train on an artificially generated dataset $\mathcal{D} = \{(x^a, y^a)\}_{a=1}^5$ with $x \sim \mathcal{N}(0,1)$ and $y = 0$. We use the full dataset to train. From Eq.(113) we know that the norm of the weights follows an approximately exponential trajectory as it approaches the widest minimum ($U = V = 0$). We therefore measure convergence timescale, T_c, by fitting an exponential ae^{-t/T_c} to the squared distance from the origin, $|\mathbf{U}|^2 + |\mathbf{V}|^2$. To extract the scaling of T_c with γ we perform SGD label noise with learning rates $\eta \in [10^{-3}, 10^{-1}]$ and corresponding momentum parameters $\beta = 1 - C\eta^\gamma$. We fit the timescale to a power-law in the learning rate $T_c(\eta, \gamma) = T_0 \eta^{-\alpha(\gamma)}$ (see Fig. 1(b)). Imposing that T_0 be independent of γ, as predicted by theory, we found the numerical value $C \approx 0.2$, which is consistent

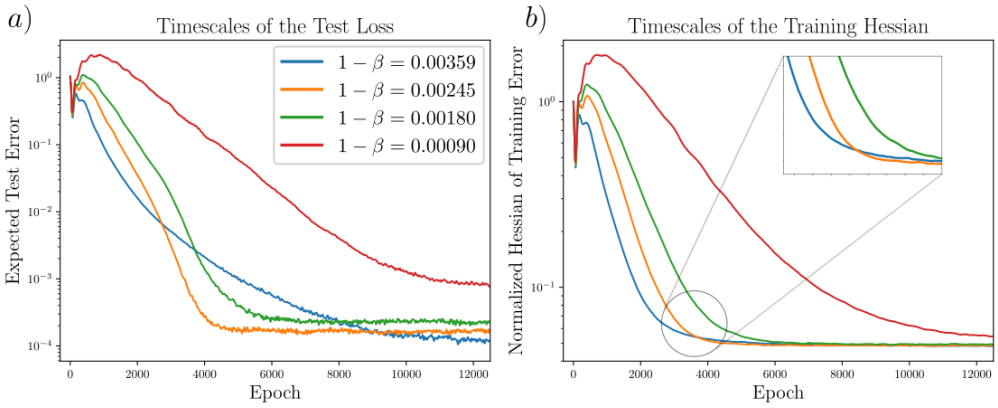

Figure 2: The expected test error, a), and Hessian of the training loss, b), in matrix sensing (with $d = 100, r = 5$, and $5rd = 2500$ samples) as a function of training epoch plotted for different values of β at $\eta = 0.1$. The label noise variance is 0.1. Each curve represents a different value of β. The inset shows that the orange curve crosses below the blue curve before convergence of the Hessian. Therefore, the same value of β is optimal for both the Hessian and the expected test error — increasing or decreasing β from this value slows down generalization.

with the theoretical estimate of $C = 0.17$ from Sec. 3.4. We find consistency with prediction across all the values of γ we simulated. Note that, for $\gamma > \frac{2}{3}$ the timescale estimate fluctuates more which is a consequence of having a slower timescale for the transverse modes. As discussed at the end of Sec. 3.3, such slowness disrupts the drift motion along the manifold. $\gamma = \frac{2}{3}$ is clearly the optimal scaling.

We repeated the same experiments using nonlinear activations, specifically we considered tanh and ReLU acting on the first layer. The timescales for tanh are shown in Fig. 1(c), and we refer the reader to Appendix A for the ReLU case. The optimal scaling value is still $\gamma = \frac{2}{3}$ for both tanh and ReLU and the optimal C remains close to 0.2.

4.2 Matrix Sensing

Next we explore speedup in a more complex setting with a higher-dimensional model, but for a well understood problem: matrix sensing. The goal is to find a low-rank matrix X^* given the measurements along random matrices A_i: $y_i = \mathrm{Tr}\, A_i X^*$. Here $X^* \in \mathbb{R}^{d\times d}$ is a matrix of rank r (Soudry et al., 2017; Li et al., 2018).

Blanc et al. (2020) analyze the problem of matrix sensing using SGD with label noise and show that if X^* is symmetric with the hypothesis $X = UU^\top$ for some matrix U, then gradient descent with label noise corresponds not only to satisfying the constraints $y_i = \mathrm{Tr}\, A_i X$, but also to minimizing an implicit regularizer, the Frobenius norm of U, which eventually leads to the ground truth.

In the analogous UV matrix model (with X^* is an asymmetric matrix of low rank r), we demonstrate a considerable learning speedup by adding momentum, and show that this speedup is not monotonic with increasing β; there is a value β^* at which the acceleration appears optimal. This non-monotonicity with an optimal β^* is observed for both the Hessian trace and the expected test error. If we assume that in this setting we also have $\gamma = 2/3$, we can extract $C^* = (1-\beta^*)/\eta^{2/3} \approx 0.24\, P^{-1/3}$, which compares favorably to the upper bound we may extract from (D) of $\approx 0.74\, (P)^{-1/3}$.

4.3 ResNet18 on CIFAR10

We now verify our predictions on a realistic problem, which will demonstrate the robustness of our analysis. We focus on ResNet18 (He et al., 2016), specifically implemented by Liu (2021), classifier trained on CIFAR10 Krizhevsky et al. (2009). We aim to extrapolate the theory by showing optimal acceleration with our hyperparameter choice once training reaches an interpolating solution. To this

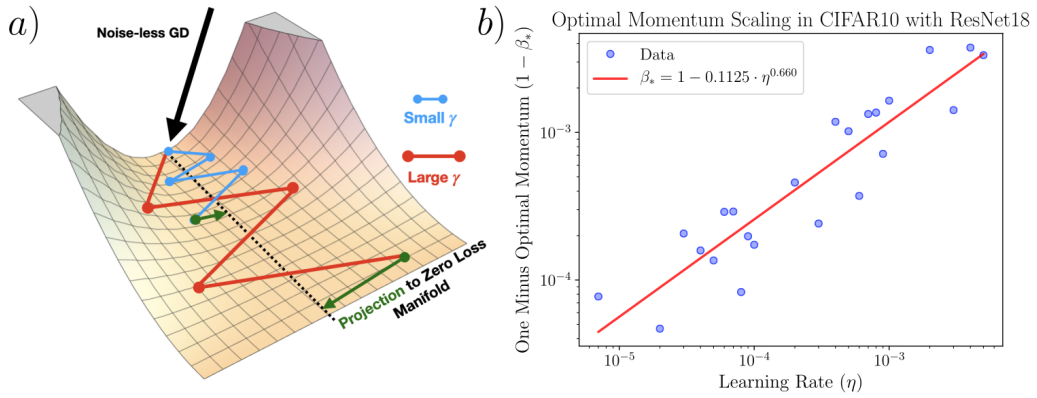

Figure 3: Classification of CIFAR10 using a ResNet18 model. Subfigure $a)$ shows our training protocol which is to first use noiseless gradient descent (black) to reach the zero-loss manifold, then to perform SGD label noise with various (blue, red) values of η and β, before finally projecting onto the valley (green) and measuring the test accuracy. Subfigure $b)$ shows the scaling of the optimal momentum, β_*, as a function of η. We perform a power-law fit whose exponent $\gamma_* = 0.660$ matches very closely to the value implied by the theory $\gamma = \frac{2}{3}$. Notice we also extract the constant $C = 0.11$.

purpose, we initialize the network on the valley. This initialization is obtained starting from a random weight values and training the network using full batch gradient descent without label noise and with a fixed value $\beta = 0.9$ until it reaches perfect training accuracy. With this initialization, we then train with SGD and label noise for a fixed number of epochs multiple times for various values the momentum hyperparameter β. Finally, we project the weights back onto the valley before recording the final test accuracy. This last step can be viewed as noise annealing and allows us to compare the performance of training the drift phase for the different values of β. From this procedure we extract the optimal momentum parameter $\beta_*(\eta)$ that maximizes the best test accuracy during training as a function of the learning rate, which we can then compare with the theoretical prediction.

As shown in Fig. 3(b), $1-\beta_*$ follows the power law we predicted almost exactly. The optimal choice for speedup does not have to coincide with the optimal choice for generalization. Strikingly, this optimal choice of scaling also leads to the best generalization in a realistic setting! This can be easily interpreted if we assume that the more we decrease the Hessian the better our model generalizes and by applying the fact that our scaling leads to the fastest transport along the manifold. The second important point is the value of the constant $C \approx 0.1$ found as the coefficient of the power-law fit. If we set $\eta = 1$ this corresponds to setting $\beta_* = 0.9$ which is the traditionally recommended value. The result here, can therefore be viewed a generalization of this common wisdom.

5 Conclusion

We studied the implicit regularization of SGD with label noise and momentum in the limit of small noise and learning rate. We found that there is an interplay between the speedup of momentum and the limiting diffusion generated by SGD noise. This gives rise to two characteristic timescales associated to the training dynamics, and the longest timescale essentially governs the training time. Maximum acceleration is thus reached when these two timescales coincide, which lead us to identifying an optimal scaling of the hyperparameters. This optimal scaling corresponds not only to faster training but also to superior generalization. More generally, we have shown how momentum can significantly enrich the dynamics of learning with SGD, modulating between qualitatively different phases of learning characaterized by different timescales and dynamical behavior. It would be interesting to explore our scaling limit in statistical mechanical theories of learning to uncover further nontrivial effects on feature extraction and generalization in the phases of learning (Jacot et al., 2018; Roberts et al., 2021; Yang et al., 2022). For future work, it will be very interesting to generalize this result to adaptive optimization algorithms such as Adam and its variants, to use this principle to design new adaptive algorithms, and to study the interplay between the scaling we found and the hyperparameter schedule.

REFERENCES

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. The Journal of Machine Learning Research, 18(1):8194–8244, 2017.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In Conference on learning theory, pp. 483–513. PMLR, 2020.

Dami Choi, Christopher J Shallue, Zachary Nado, Jaehoon Lee, Chris J Maddison, and George E Dahl. On empirical comparisons of optimizers for deep learning. arXiv preprint arXiv:1910.05446, 2019.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In Artificial intelligence and statistics, pp. 192–204. PMLR, 2015.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise sgd provably prefers flat global minimizers. Advances in Neural Information Processing Systems, 34:27449–27461, 2021.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In International conference on machine learning, pp. 1309–1318. PMLR, 2018.

KJ Falconer. Differentiation of the limit mapping in a dynamical system. Journal of the London Mathematical Society, 2(2):356–372, 1983.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. ICLR, 2017.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. Advances in neural information processing systems, 31, 2018.

Jeff Z. HaoChen, Colin Wei, Jason Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In Conference on Learning Theory, volume 134 of Proceedings of Machine Learning Research, pp. 2315–2357. PMLR, 15–19 Aug 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. Neural computation, 9(1):1–42, 1997.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. Advances in neural information processing systems, 31, 2018.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In Conference On Learning Theory, pp. 545–604. PMLR, 2018.

Gary Shon Katzenberger. Solutions of a stochastic differential equation forced onto a manifold by a large drift. The Annals of Probability, 19(4):1587–1628, 1991.

Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from Adam to SGD. arXiv preprint arXiv:1712.07628, 2017.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. ICLR, 2017.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. ICLR, 2015.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Rong Ge, and Sanjeev Arora. Explaining landscape connectivity of low-cost solutions for multilayer nets. Advances in neural information processing systems, 32, 2019.

Daniel Kunin, Javier Sagastuy-Brena, Lauren Gillespie, Eshed Margalit, Hidenori Tanaka, Surya Ganguli, and Daniel LK Yamins. Limiting dynamics of sgd: Modified loss, phase space oscillations, and anomalous diffusion. arXiv preprint arXiv:2107.09133, 2021.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. arXiv preprint arXiv:2003.02218, 2020.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In Conference On Learning Theory, pp. 2–47. PMLR, 2018.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss?–a mathematical framework. ICLR, 2022.

Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. ICLR, 2020.

Kuang Liu. Train cifar10 with pytorch. https://github.com/kuangliu/pytorch-cifar/, February 2021. commit/49b7aa97b0c12fe0d4054e670403a16b6b834ddd.

Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. Advances in Neural Information Processing Systems, 33:8543–8552, 2020.

Stefano Sarao Mannelli and Pierfrancesco Urbani. Just a momentum: Analytical study of momentum-based acceleration methods in paradigmatic high-dimensional non-convex problems. arXiv preprint arXiv:2102.11755, 2021.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. ICLR, 2020.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In Dokl. akad. nauk Sssr, volume 269, pp. 543–547, 1983.

Quynh Nguyen. On connected sublevel sets in deep learning. In International conference on machine learning, pp. 4790–4799. PMLR, 2019.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. Advances in Neural Information Processing Systems, 34:29218–29230, 2021.

Loucas Pillaud-Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. arXiv preprint arXiv:2206.09841, 2022.

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

Boris T Polyak. Introduction to optimization. optimization software. Inc., Publications Division, New York, 1:32, 1987.

Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In Proceedings of the 22nd international conference on Machine learning, pp. 713–719, 2005.

Daniel A Roberts, Sho Yaida, and Boris Hanin. The principles of deep learning theory. arXiv preprint arXiv:2106.10165, 2021.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. arXiv preprint arXiv:1611.07476, 2016.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. arXiv preprint arXiv:1706.04454, 2017.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. ICLR, 2014.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. arxiv e-prints, art. arXiv preprint arXiv:1710.10345, 2017.

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.

Mingwei Wei and David J Schwab. How noise affects the hessian spectrum in overparameterized neural networks. arXiv preprint arXiv:1910.00195, 2019.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. Advances in neural information processing systems, 30, 2017.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. Advances in Neural Information Processing Systems, 31, 2018.

Lei Wu, Mingze Wang, and Weijie Su. When does sgd favor flat minima? a quantitative characterization via linear stability. arXiv preprint arXiv:2207.02628, 2022.

Zeke Xie, Xinrui Wang, Huishuai Zhang, Issei Sato, and Masashi Sugiyama. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In International Conference on Machine Learning, pp. 24430–24459. PMLR, 2022.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. arXiv preprint arXiv:2203.03466, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. corr abs/1611.03530. ICLR, 2017.

A EXPERIMENTAL DESIGN

A.1 UV MODEL

In our experiments with the vector UV model we aim to extract how the timescale of motion scales with η for each γ, therefore we train the model over a sweep of both of these parameters. As a reminder, the loss of the UV model is

$$L = \frac{1}{2P} \sum_{i=1}^{P} \left(y_i - \frac{1}{\sqrt{n}} \boldsymbol{u} \cdot \boldsymbol{v} x_i \right)^2 \tag{14}$$

where P is the dataset size and $\boldsymbol{u}, \boldsymbol{v}$ are n-dimensional vectors. We set $y_i = 0$ for simplicity, and with label noise this becomes $y_i(t) = \delta \cdot \xi_i(t)$ for i.i.d standard Gaussian distributed ξ_i. We initalize x_i once (not online) as i.i.d. random standard Gaussian variables. For all experiments we choose $\delta = \frac{1}{2}$. We initialize with u_i, v_i i.i.d. standard Gaussian distributed and keep this initialization constant over all our experiments to reduce noise associated with the specific initialization.

For each value of η, β we train with label noise SGD and momentum until $|\boldsymbol{u}|^2 + |\boldsymbol{v}|^2 < \epsilon n$ with $\epsilon = 0.1$, thereby obtaining a time series for each of $\boldsymbol{u}(t)$ and $\boldsymbol{v}(t)$. We extract the timescale by fitting $\log\left(|\boldsymbol{u}|^2 + |\boldsymbol{v}|^2\right)$ to a linear function and taking the slope.

A.2 MATRIX SENSING

In the experiments with matrix sensing we aim to demonstrate the benefit of momentum in a popular setting. Matrix sensing corresponds to the following problem: Given a target matrix $X^* \in \mathbb{R}^{d \times d}$ of low rank $r \ll d$ and measurements $\{y_i = \operatorname{Tr} A_i X^*\}_{i=1}^m$ how can we reconstruct X^*? One way to solve this problem is to write our guess $X = UV$ the product of two other matrices, and do stochastic gradient descent on them, hoping that the implicit regularization induced by this parametrization and the learning algorithm will converge to a good low rank X.

In our experiments we study the $d = 100, r = 5, m = 5rd = 2500$ case. We draw $(A_i)_{ij} \sim \mathcal{N}(0,1)$ as standard Gaussians and choose X^* by drawing first $(X_0)_{ij} \sim \mathcal{N}(0,1)$ and then performing SVD and projecting onto the top r singular values by zeroing out the smaller singular values in the diagonal matrix. We intitalize $U = V = I_d$. We perform SGD with momentum on the time dependent loss (with label noise depending on time)

$$L(t) = \frac{1}{dm} \sum_{i=1}^{m} \left(\delta \cdot \xi_i(t) + y_i - \operatorname{Tr}(A_i U V) \right)^2 \tag{15}$$

where $\delta^2 = 0.1$, $\xi_i(t) \sim \mathcal{N}(0,1)$. We choose $\eta = 0.1$ for all of our experiments.

The hessian of the loss is defined to be the Hessian averaged over the noise. Equivalently we may just set $\xi_i(t) = 0$ when we calculate the Hessian because averaging over the noise decouples the noise. Similarly when we define the expected test loss we define it as an average over all A_i setting $\xi_i(t) = 0$ in order to decouple the noise. Averaging over $\xi_i(t)$ and A_i would simply lead to an additional term $\langle \xi_i(t)^2 \rangle$ which would simply contribute a constant. We remove this constant for clarity.

It is also interesting to note that we observe epoch-wise double descent Nakkiran et al. in this problem. In particular, we observe that the peak in the test error can be controlled by the momentum hyperparameter, and becomes especially pronounced for $\beta \to 1$.

A.3 RESNET18 CIFAR10

We train our model in three steps: full batch without label noise until 100% test accuracy, SGD with label noise and momentum, and then a final projection onto the interpolating manifold. The model we use is the ResNet18 and we train on the CIFAR10 training set.

The first step is full batch gradient descent on the full CIFAR10 training set of 50,000 samples. We train with a learning rate $\eta = 0.1$ and momentum $\beta = 0.9$ and a learning rate schedule with linearly increases from 0 to η over 600 epochs, after which it undergoes a cosine learning rate schedule for

1400 more epochs stopping the first time the network reaches 100% test accuracy which happened on epoch 1119 in our run. This model is saved and the same one is used for all future runs.

The loss function we use is cross cross entropy loss. Because we will choose a label noise level of $p = 0.2$ which corresponds to a uniformly wrong label 20% of the time, during this phase of training we train with the expected loss over this randomness. Notice that this loss is actually linear in the labels so taking the expectation is easy.

The second step involves starting from the initialization in step 1 and training with a different learning rate and momentum parameter. In this step we choose the same level of label noise $p = 0.2$ but take it to be stochastic. Additionally we use SGD instead of gradient descent with a batch size of 512. This necessitates decreasing the learning rate because noise is greatly increased as demonstrated in the main text. In this step we train for a fixed 200 epochs for any learning rate momentum combination. We only compare runs with the same learning rate value.

The final step is a projection onto the zero loss manifold. This step is necessary because the total width of the distribution around the zero loss manifold scales with $\frac{1}{1-\beta}$, and this will distort the results systematically at larger momentum, making them look worse than they are. We perform this projection to correct for this effect and put all momenta on an equal footing. This projection is done by training on the full batch with Adam and a learning rate of $\eta = 0.0001$ in order to accelerate training. We do not expect any significant systematic distortion by using Adam for the projection instead of gradient descent.

To determine the optimal value of β we sweep several values of β and observe the test accuracy after the previously described procedure. To get a more precise value of β instead of simply selecting the one with the highest test accuracy we fit the accuracy $A(\beta)$ to

$$A(\beta) = a_{max} + \begin{cases} a_1(\beta - \beta_*) & \text{if } \beta \leq \beta_* \\ a_2(\beta - \beta_*) & \text{if } \beta \geq \beta_* \end{cases} \tag{16}$$

to the parameters a_{max}, a_1, a_2, and β_*, thereby extracting β_* for each η.

B Review of relevant results from Katzenberger (1991)

In this Appendix we summarize the relevant conditions and theorems from Katzenberger (1991) that we use to prove our result on the limiting drift-diffusion. We refer to Katzenberger (1991) for part of the definitions and conditions cited throughout the below. In what follows, $(\Omega^n, \mathcal{F}^n, \{\mathcal{F}^n_t\}_{t\geq 0}, P)$ will denote a filtered probability space, Z_n an $\mathbb{R}^r$-valued cadlag $\{\mathcal{F}^n_t\}$-semimartingale with $Z_n(0) = 0$, A_n a real-valued cadlag $\{\mathcal{F}^n_t\}$-adapted nondecreasing process with $A_n(0) = 0$, and $\tilde{\sigma} : U \to \mathbb{R}^{D\times r}$ a continuous function, where U is a neighborhood of $\{\mathbf{0}\} \times \Gamma$ as defined in the main text. Also, X_n is an $\mathbb{R}^D$ valued cadlag $\{\mathcal{F}^n_t\}$-semimartingale satisfying

$$X_n(t) = X_n(0) + \int_0^t \sigma_n(X_n) dZ_n + \int_0^t F(X_n) dA_n \tag{17}$$

for all $t \leq \lambda_n(K)$ and all compact $K \subset U$, where

$$\lambda_n(K) = \inf\{t \geq 0 | X_n(t-) \neq \mathring{K} or X_n(t) \neq \mathring{K}\} \tag{18}$$

be the stopping time of $X_n(t)$ to leave $\mathring{K}$, the interior of K. For cadlag real-valued seimimartingales X, Y let $[X, Y](t)$ be defined as the limit of sums

$$\sum_{i=0}^{n-1} (X(t_{i+1}) - X(t_i))(Y(t_{i+1} - Y(t_i)) \tag{19}$$

where $0 = t_0 < t_1 < \cdots < t_n = t$ and the limit is in probability as the mesh size goes to zero. If X is an $\mathbb{R}^D$-valued semimartingale, we write

$$[X] = \sum_{i=1}^{D} [X_i, X_i]. \tag{20}$$

Condition B.1. *For every $T > \epsilon > 0$ and compact $K \subset U$*

$$\inf_{0 \le t \le T \wedge \lambda_n(K) - \epsilon)} (A_n(t+\epsilon) - A_n(t)) \to \infty \tag{21}$$

as $n \to \infty$ where the infimum of the empty set is taken to be ∞.

Condition B.2. *For every compact $K \subset U$ $\{Z_n^{\lambda_n(K)}\}$ satisfies the following: For $n \ge 1$ let Z_n be a $\{\mathcal{F}_t^n\}-$semimartingale with sample paths in $D_{\mathbb{R}^d}[0,\infty)$. Assume that for some $\delta > 0$ allowing $\delta = \infty$ and every $n \ge 1$, there exist stopping times $\{\tau_n^k | k \ge 1\}$ and a decomposition of $Z_n - J_\delta(Z_n)$ into a local martingale M_n plus a finite variation process F_n such that $P[\tau_n^k \le k] \le 1/k$, $\{[M_n](t \wedge \tau_n^k) + T_{t \wedge \tau_n^k}(F_n) | n \ge 1\}$ is uniformly integrable for every $t \ge 0$ and $k \ge 1$ and*

$$\lim_{\gamma \to 0} \limsup_{n \to \infty} P\left[\sup_{0 \le t \le T} (T_{t+\gamma}(F_n) - T_t(F_n)) > \epsilon\right] = 0 \tag{22}$$

for every $\epsilon > 0$ and $T > 0$. Also as $n \to \infty$ and for any $T > 0$

$$\sup_{0 < t \le T \wedge \lambda_n(K)} |\Delta Z_n(t)| \to 0 \tag{23}$$

Condition B.3. *The process*

$$\bar{Z}_n(t) = \sum_{0 < s \le t} \Delta Z_n(s) \Delta A_n(s) \tag{24}$$

exists, is an $\{\mathcal{F}_t^n\}-$semimartingale, and for every compact $K \subset U$, the sequence $\{\bar{Z}_n^{\lambda_n(K)}\}$ is relatively compact and satisfies Condition 4.1 in Katzenberger (1991).

Theorem B.4 (Theorem 7.3 in Katzenberger (1991))**.** *Assume that Γ is C^2 and for every $y \in \Gamma$, the matrix $\partial F(y)$ has $D - M$ eigenvalues in $D(1)$. Assume (B.1),(B.2) and (B.3) hold, Φ is C^2 (or F is LC^2) and $X_n(0) \Rightarrow X(0) \in U$. Let*

$$Y_n(t) = X_n(t) - \psi(X(0), A(t)) + \Phi(X(0)) \tag{25}$$

and, for a compact $K \subset U$, let

$$\mu_n(K) = \inf\{t \ge 0 | Y_n(t-) \notin \mathring{K} \text{ or } Y_n(t) \notin \mathring{K}\}. \tag{26}$$

Then for every compact $K \subset U$, the sequence $\{(Y_n^{\mu_n(K)}, Z_n^{\mu_n(K)}, \mu_n(K))\}$ is relatively compact in $D_{\mathbb{R}^{2D \times r}}[0,\infty) \times [0,\infty]$ (see Katzenberger (1991) for details about the topology). If (Y, Z, μ) is a limit of this sequence then (Y, Z) is a continuous semimartingale, $Y(t) \in \Gamma$ for every t almost surely, $\mu \ge \inf\{t \ge 0 | Y(t) \notin \mathring{K}\}$ almost surely, and

$$Y(t) = Y(0) + \int_0^{t \wedge \mu} \partial\Phi(Y)\tilde{\sigma}(Y) dZ + \frac{1}{2} \sum_{ijkl} \int_0^{t \wedge \mu} \partial_{ij}\Phi(Y)\tilde{\sigma}^{ik}(Y)\tilde{\sigma}^{jl}(Y) d[Z^k, Z^l]. \tag{27}$$

B.1 Applying Theorem B.4

Recall the equations of motion of stochastic gradient descent

$$\pi_{k+1}^{(n)} = \beta\pi_k^{(n)} - \nabla L(w_k^{(n)}) + \epsilon_n \sigma(w_k^{(n)})\xi_k, \qquad w_{k+1}^{(n)} = w_k^{(n)} + \eta\pi_{k+1}^{(n)}, \tag{28}$$

where $\sigma(w) \in \mathbb{R}^{D \times r}$ is the noise function evaluated at $w \in \mathbb{R}^D$, and $\xi_k \in \mathbb{R}^r$ is a noise vector drawn i.i.d. at every timestep k with zero mean and unit variance. We now show that this equation satisfies all the properties required by Theorem B.4.

The manifold Γ is the fixed point manifold of (non-stochastic) gradient descent. $\{\mathbf{0}\} \times \Gamma$ is a $\mathcal{C}^2$ manifold because Γ is $\mathcal{C}^2$, which follows from assumption 3.1. The flow $F(w, \pi) = (\eta(\beta\pi - \nabla L(w)), \beta\pi - \nabla L(w))$. As shown in Appendix E, dF has exactly M zero eigenvalues on $\Gamma \cap K$. F inherits the differentiable and locally Lipschitz properties from ∇L, and therefore satisfies the conditions of B.4.

Next, notice that the noise function $\tilde{\sigma} : R^{2D} \to \mathbb{R}^{2D \times r}$ is continuous because σ is.

Now we define A_n and Z_n (as in the main text) so that X_n reproduces the dynamics in equation (28), except with a new time parameter $t = k\epsilon_n^2$.

$$A_n(t) = \left\lfloor \frac{t}{\epsilon_n^2} \right\rfloor, \qquad Z_n(t) = \epsilon_n \sum_{k=1}^{A_n(t)} \xi_k \tag{29}$$

So that, with these choices, Eq. (17) precisely corresponds to (28), up to the rescaling $t = k\epsilon_n^2$.

Now we show that A_n, Z_n satisfy the conditions of B.4. Clearly $A_n(0) = Z_n(0)$ by definition. Then

$$A_n(t+\varepsilon) - A_n(t) = \left\lfloor \frac{t+\varepsilon}{\epsilon_n^2} \right\rfloor - \left\lfloor \frac{t}{\epsilon_n^2} \right\rfloor \geq \frac{t+\varepsilon-t}{\epsilon_n^2} - 2 = \frac{\varepsilon}{\epsilon_n^2} - 2 \to \infty \tag{30}$$

when we take $\epsilon_n \to 0$, thus recoverying condition B.1.

By definition Z_n is a martingale. Notice also by the definition of Z_n, because ξ_k is i.i.d. with variance 1, that $Z_n(t)$ has variance $A_n(t)\epsilon_n^2 \leq t$ which is uniformly bounded and hence $Z_n(t)$ is uniformly integrable for stopping times $\tau_n^k = 2k > k$. Also note that $\left|\Delta Z_n(k\epsilon_n^2)\right| = |\epsilon_n \xi_k|$ which goes to zero in probability as ϵ becomes small because ξ_k has bounded variance, and $\Delta Z_n(t)$ is zero otherwise. This shows that Z_n satisfies condition B.2. Because Z_n and A_n are discontinuous at the same time we automatically satisfy condition B.3 as pointed out by Katzenberger.

This shows that we satisfy the conditions of Theorem B.4, therefore we have the following

Lemma B.5. *The SGD equations formulated as in (28) satisfy all the conditions of Theorem B.4.*

C Explicit expression of limiting diffusion in momentum SGD

In this section we provide the proof of Theorem 3.4. Recall that Φ satisfies

$$\Phi(x + F(x)) = \Phi(x)\,, \tag{31}$$

where $x = (\pi, w)$, and F given in (114). To obtain the explicit expression of the limiting drift-diffusion, according to Theorem (B.4), and keeping into account Assumption (3.2), we need to determine Φ up to its second derivatives. To this aim, we shall expand Eq. (31) up to second order in the series expansion in F. In components, this reads:

$$\begin{aligned} &\eta\partial_{w_i}\Phi^j(\pi^i - g^i) + \partial_{\pi_i}\Phi^j(-C\eta^\gamma\pi^i - g^i) + \frac{1}{2}\eta^2\partial_{w_i}\partial_{w_k}\Phi^j(\pi^i - g^i)(\pi^k - g^k) \\ &+\frac{1}{2}\eta\partial_{w_i}\partial_{\pi_k}\Phi^j(\pi^i - g^i)(-C\eta^\gamma\pi^k - g^k) + \frac{1}{2}\eta\partial_{\pi_i}\partial_{w_k}\Phi^j(-C\eta^\gamma\pi^i - g^i)(\pi^k - g^k) \\ &+\frac{1}{2}\partial_{\pi_i}\partial_{\pi_k}\Phi^j(-C\eta^\gamma\pi^i - g^i)(-C\eta^\gamma\pi^k - g^k) = 0 \end{aligned} \tag{32}$$

subject to the boundary condition

$$\Phi(\pi, w)|_{w\in\Gamma,\pi=0} = w\,. \tag{33}$$

Here, $g^i = \partial_i L$. In Eq. (32), we already substituted $\beta = 1 - C\eta^\gamma$. In what follows, we shall find Φ to leading order in η as $\eta \to 0$. We will solve the above problem by performing a series expansion in Φ around a point $\bar{w} \in \Gamma$ and $\pi = 0$ up to second order:

$$\Phi(\pi, \bar{w} + \delta w) = \Phi_{00} + \Phi_{01}^i\delta w^i + \Phi_{10}^i\pi^i + \frac{1}{2}\Phi_{02}^{ij}\delta w^i\delta w^j + \Phi_{11}^{ij}\pi^i\delta w^j + \frac{1}{2}\Phi_{20}^{ij}\pi^i\pi^j + \cdots\,. \tag{34}$$

For example, we have

$$\Phi_{01}^i = \frac{\partial\Phi}{\partial w^i}, \qquad \Phi_{11}^{ij} = \frac{\partial^2\Phi}{\partial\pi^i\partial w^j}, \qquad \Phi_{20}^{ij} = \frac{\partial^2\Phi}{\partial\pi^i\partial\pi^j}\,. \tag{35}$$

More precisely, we regard this as an expansion in powers of δw and π: Φ_{00} is zeroth order, Φ_{10}, Φ_{01} are first order, and the remaining terms are second order. We will occasionally write explicitly the index of Φ, e.g. $\Phi_{02}^{k,ij} = \frac{\partial\Phi^k}{\partial w^i\partial w^j}$. It will be useful to introduce the longitudinal projector onto Γ,

$P_L(w) : \mathbb{R}^D \to T_w\Gamma$, defined such that for any vector $v \in T_w\Gamma$: $P_L v = v$. The transverse projector is then $P_T = \mathrm{Id} - P_L$. We will also decompose various tensors using these projectors, e.g.

$$\Phi_{11}^{ij} = \Phi_{11LL}^{ij} + \Phi_{11LT}^{ij} + \Phi_{11TL}^{ij} + \Phi_{11TT}^{ij}\,, \tag{36}$$

with $\Phi_{11LT}^{ij} = \Phi_{11}^{kl} P_L^{ik} P_T^{jl}$. Note that the Hessian $H = \nabla g \in \mathbb{R}^{D\times D}$ satisfies $HH^\dagger = P_T$, where $H^\dagger$ denotes the pseudoinverse.

At zeroth order, we obviously have $\Phi_{00}(w) = w$.

Lemma C.1. *The first order terms in the series expansion (34) are given by*

$$\Phi_{01T} = \Phi_{10T} = 0, \qquad \Phi_{01L} = P_L, \qquad \Phi_{10L} = C^{-1}\eta^{1-\gamma} P_L\,. \tag{37}$$

Proof. We now proceed to finding the first order terms. Suppose $\hat{w}(s) \in \mathbb{R}^D$ is a curve lying on Γ. Then due to the boundary condition (33), $\partial_s \Phi(\pi = 0, \hat{w}) = \Phi_{01}^i \partial_s \hat{w}^i = \partial_s \hat{w}^i$. This means that $\Phi_{01L} = P_L$. Now from (32),

$$\eta \Phi_{01}^i (\pi^i - \partial_j g^i \delta w^j) + \Phi_{10}^i (-C\eta^\gamma \pi^i - \partial_j g^i \delta w^j) = 0\,. \tag{38}$$

This condition should hold for any π and δw, therefore we arrive at

$$\Phi_{10} = C^{-1}\eta^{1-\gamma}\Phi_{01} \tag{39}$$

and

$$(\eta \Phi_{01}^i + \Phi_{10}^i)\partial_k g^i = 0\,. \tag{40}$$

Decomposing into longitudinal and transverse components, and noting that the Hessian satisfies $P_L H = 0$, the above equation becomes

$$\eta \Phi_{01T}^i + \Phi_{10T}^i = 0\,, \tag{41}$$

which together with (39) and the above discussion gives

$$\Phi_{01T} = \Phi_{10T} = 0, \qquad \Phi_{01L} = P_L, \qquad \Phi_{10L} = C^{-1}\eta^{1-\gamma} P_L\,, \tag{42}$$

which concludes the first order analysis. □

We now need a Lemma, which requires Definition (3.3) in the main text. We report it here for convenience:

Definition C.2. For a symmetric matrix $H \in \mathbb{R}^D \times \mathbb{R}^D$, and $W_H = \{\Sigma \in \mathbb{R}^D \times \mathbb{R}^D : \Sigma = \Sigma^\top, HH^\dagger\Sigma = H^\dagger H\Sigma = \sigma\}$, we define the operator $\tilde{\mathcal{L}}_H : W_H \to W_H$ with $\tilde{\mathcal{L}}_H S \equiv \{H, S\} + \frac{1}{2}C^{-2}\eta^{1-2\gamma}[[S, H], H]$, with $[S, H] = SH - HS$.

Lemma C.3. *The inverse of the operator $\tilde{\mathcal{L}}_H$ is unique.*

Proof. Let us go to a basis where H is diagonal, with eigenvalues λ_i. In components, the equation $\tilde{\mathcal{L}}_H S = M$ reads

$$(\lambda_i + \lambda_j)S_{ij} + \frac{1}{2}C^{-2}\eta^{1-2\gamma}(\lambda_i - \lambda_j)^2 S_{ij} + C^{-1}\eta^{1-\gamma}M_{ij} = 0 \tag{43}$$

which has a unique solution, with

$$S_{ij} = -C^{-1}\eta^{1-\gamma}\left(\lambda_i + \lambda_j + \frac{1}{2}C^{-2}\eta^{1-2\gamma}(\lambda_i - \lambda_j)^2\right)^{-1} M_{ij}\,. \tag{44}$$

□

Lemma C.4. *The second order terms in the series expansion (34) are given by*

$$\Phi_{02LL}^{j,ik} = -(H^\dagger)^{jl}\partial_i^L H^{ln} P_L^{nk}, \quad \Phi_{02LT}^{j,ik} = -P_L^{jl}\partial_i^L H^{ln}(H^\dagger)^{nk}, \quad \Phi_{02TT}^{j,ik} = O\left(\eta^{min\{0,1-2\gamma\}}\right) \tag{45}$$

$$\Phi_{11LL}^{j,ik} = -C^{-1}\eta^{1-\gamma}(H^\dagger)^{jl}\partial_i^L H^{ln} P_L^{nk}, \quad \Phi_{11TL}^{j,ik} = -C^{-1}\eta^{1-\gamma}P_L^{jl}\partial_k^L H^{ln}(H^\dagger)^{ni} \tag{46}$$

$$\Phi_{11TT}^{j,ik} = -C^{-1}\eta^{1-\gamma}\tilde{\mathcal{L}}_H^{-1}(M^{(j)})_{ik} - \frac{1}{2}C^{-3}\eta^{2-3\gamma}[H, \tilde{\mathcal{L}}_H^{-1}M^{(j)}]_{ik} \tag{47}$$

$$\Phi_{20LL}^{j,ik} = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}\left((H^\dagger)^{jl}\partial_i^L H^{ln} P_L^{nk} + (H^\dagger)^{jl}\partial_k^L H^{ln} P_L^{ni}\right) \tag{48}$$

$$\Phi_{20TL}^{j,ik} = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}\left(P_L^{jl}\partial_k^L H^{ln}(H^\dagger)^{ni} + P_L^{jl}\partial_i^L H^{ln}(H^\dagger)^{nk}\right) \tag{49}$$

$$\Phi_{20TT}^{j,ik} = -C^{-2}\eta^{2-2\gamma}\tilde{\mathcal{L}}_H^{-1}(P_T\partial_j^L H P_T)_{ik} \tag{50}$$

where

$$(M^{(j)})_{kl} = P_T^{ki} \partial_j^L H_{ni} P_T^{nl} \tag{51}$$

Proof. Consider a path $\hat{w}(s)$ lying on Γ. From $P_T H = H$ we have

$$\frac{dP_T}{ds} H = (\mathrm{Id} - P_T) \frac{dH}{ds} = P_L \frac{dH}{ds} \,, \tag{52}$$

and thus, using $HH^\dagger = P_T$, we find

$$\frac{dP_T}{ds} P_T = P_L \frac{dH}{ds} H^\dagger, \qquad P_T \frac{dP_T}{ds} = H^\dagger \frac{dH}{ds} P_L \,, \tag{53}$$

where the second equation is obtained from the first by taking the transpose. Putting the last two relations together, we find

$$\frac{dP_L}{ds} = -\frac{dP_T}{ds} = -P_T \frac{dP_T}{ds} - \frac{dP_T}{ds} P_T = -H^\dagger \frac{dH}{ds} P_L - P_L \frac{dH}{ds} H^\dagger \,. \tag{54}$$

From (37) we can then write

$$\partial_i^L \Phi_{01}^{j,k} = -(H^\dagger)^{jl} \partial_i^L H^{ln} P_L^{nk} - P_L^{jl} \partial_i^L H^{ln} (H^\dagger)^{nk} \,, \tag{55}$$

where $\partial_i^L = P_L^{ij} \partial_j$. This leads to

$$\Phi_{02LL}^{j,ik} = P_L^{kl} \partial_i^L \Phi_{01}^{j,l} = -(H^\dagger)^{jl} \partial_i^L H^{ln} P_L^{nk} \tag{56}$$

$$\Phi_{02LT}^{j,ik} = P_T^{kl} \partial_i^L \Phi_{01}^{j,l} = -P_L^{jl} \partial_i^L H^{ln} (H^\dagger)^{nk} \,. \tag{57}$$

Also note that $\Phi_{02LT}^{j,ik} = \Phi_{02TL}^{j,ki}$, so the only component still to be determined in Φ_{02} is Φ_{02TT}.

The next step is to expand Eq. (32) to second order:

$$\begin{aligned}
&\pi^i \pi^k \left(\eta(1 + C\eta^\gamma) \Phi_{11}^{ki} - C\eta^\gamma (1 - \tfrac{1}{2} C\eta^\gamma) \Phi_{20}^{ik} + \tfrac{1}{2} \eta^2 \Phi_{02}^{ik} \right) \\
&+ \delta w^k \pi^i \left(\eta \Phi_{02}^{ik} - \eta^2 H_{kj} \Phi_{02}^{ji} - \eta \Phi_{11}^{il} H_{kl} - C\eta^\gamma \Phi_{11}^{ik} - \eta(1 - C\eta^\gamma) \Phi_{11}^{(ij)} H_{kj} - (1 - C\eta^\gamma) \Phi_{20}^{li} H_{kl} \right) \\
&- \delta w^k \delta w^l \Big(\eta \Phi_{02}^{ik} H_{li} + \tfrac{1}{2} \eta^2 \Phi_{02}^{ji} H_{lj} H_{ki} + \Phi_{11}^{ik} H_{li} + \eta \Phi_{11}^{ji} H_{kj} H_{li} \\
&\qquad\qquad + \tfrac{1}{2} (\eta \Phi_{01}^i + \Phi_{10}^i) \partial_k H_{li} + \tfrac{1}{2} \Phi_{20}^{ji} H_{lj} H_{ki} \Big) = 0 \,,
\end{aligned} \tag{58}$$

where $A^{(ij)} = \frac{1}{2}(A^{ij} + A^{ji})$ denotes the symmetric part. Neglecting various terms that are subleading at small η, gives

$$\begin{aligned}
&\pi^i \pi^k \left(\eta \Phi_{11}^{ki} - C\eta^\gamma \Phi_{20}^{ik} + \tfrac{1}{2} \eta^2 \Phi_{02}^{ik} \right) \\
&+ \delta w^k \pi^i \left(\eta \Phi_{02}^{ik} - C\eta^\gamma \Phi_{11}^{ik} - \Phi_{20}^{li} H_{kl} \right) \\
&- \delta w^k \delta w^l \left(\eta \Phi_{02}^{ik} H_{li} + \Phi_{11}^{ik} H_{li} + \tfrac{1}{2} (\eta \Phi_{01}^i + \Phi_{10}^i) \partial_k H_{li} + \tfrac{1}{2} \Phi_{20}^{ji} H_{lj} H_{ki} \right) = 0 \,,
\end{aligned} \tag{59}$$

The first line immediately gives

$$\Phi_{20}^{ij} = C^{-1} \eta^{1-\gamma} \Phi_{11}^{(ij)} + \frac{1}{2} C^{-1} \eta^{2-\gamma} \Phi_{02}^{ij} \,. \tag{60}$$

Taking the second line of (59) and projecting onto the longitudinal part the index k gives

$$\eta\beta \Phi_{02LL}^{ik} - C\eta^\gamma \Phi_{11LL}^{ik} = 0 \tag{61}$$

$$\eta\beta \Phi_{02TL}^{ik} - C\eta^\gamma \Phi_{11TL}^{ik} = 0 \tag{62}$$

and using (56),(57),

$$\Phi_{11LL}^{j,ik} = C^{-1} \eta^{1-\gamma} \Phi_{02LL}^{j,ik} = -C^{-1} \eta^{1-\gamma} (H^\dagger)^{jl} \partial_i^L H^{ln} P_L^{nk} \tag{63}$$

$$\Phi_{11TL}^{j,ik} = C^{-1} \eta^{1-\gamma} \Phi_{02TL}^{j,ik} = -C^{-1} \eta^{1-\gamma} P_L^{jl} \partial_k^L H^{ln} (H^\dagger)^{ni} \tag{64}$$

Projecting the second line of (59) on the transverse part of the index k we have

$$\eta\Phi^{j,ik}_{02LT} - C\eta^{\gamma}\Phi^{j,ik}_{11LT} - \Phi^{j,li}_{20TL}H_{kl} = 0 \tag{65}$$

$$\eta\Phi^{j,ik}_{02TT} - C\eta^{\gamma}\Phi^{j,ik}_{11TT} - \Phi^{j,li}_{20TT}H_{kl} = 0 \tag{66}$$

Using (60) and keeping (57) into account, Eq. (65) becomes, neglecting subleading terms in η,

$$-\eta P^{jl}_{L}\partial^{L}_{i}H^{ln}(H^{\dagger})^{nk} - C\eta^{\gamma}\Phi^{ik}_{11LT} - \frac{1}{2}C^{-1}\eta^{1-\gamma}\Phi^{li}_{11TL}H_{kl} - \frac{1}{2}C^{-1}\eta^{1-\gamma}\Phi^{il}_{11LT}H_{kl} = 0\,. \tag{67}$$

Using (64) in (67),

$$\begin{aligned} -\eta P^{jl}_{L}\partial^{L}_{i}H^{ln}(H^{\dagger})^{nk} - C\eta^{\gamma}\Phi^{ik}_{11LT} \\ +\frac{1}{2}C^{-2}\eta^{2-2\gamma}P^{jp}_{L}\partial^{L}_{i}H^{pn}(H^{\dagger})^{nl}H_{kl} - \frac{1}{2}C^{-1}\eta^{1-\gamma}\Phi^{il}_{11LT}H_{kl} = 0\,, \end{aligned} \tag{68}$$

and simplifying,

$$\begin{aligned} -\eta P^{jl}_{L}\partial^{L}_{i}H^{ln}(H^{\dagger})^{nk} - C\eta^{\gamma}\Phi^{ik}_{11LT} \\ +\frac{1}{2}C^{-2}\eta^{2-2\gamma}P^{jl}_{L}\partial^{L}_{i}H^{ln}P^{nk}_{T} - \frac{1}{2}C^{-1}\eta^{1-\gamma}\Phi^{il}_{11LT}H_{kl} = 0\,. \end{aligned} \tag{69}$$

which determines Φ_{11LT} in close form. Indeed the above has the form, in matrix notation

$$-\frac{1}{2}C^{-1}\eta^{1-\gamma}\Phi_{11LT}H - C\eta^{\gamma}\Phi_{11LT} = M\,, \tag{70}$$

and can be immediately inverted to solve for Φ_{11LT}. The only two undetermined components now are Φ_{11TT} and Φ_{02TT}. One condition is obtained from (66) which gives, keeping (60) into account,

$$\eta\Phi^{j,ik}_{02TT} - C\eta^{\gamma}\Phi^{j,ik}_{11TT} - C^{-1}\eta^{1-\gamma}\Phi^{j,(li)}_{11TT}H_{kl} = 0\,, \tag{71}$$

and thus

$$\Phi^{j,ik}_{02TT} = C\eta^{\gamma-1}\Phi^{j,ik}_{11TT} + C^{-1}\eta^{-\gamma}\Phi^{j,(ni)}_{11TT}H_{kn}\,. \tag{72}$$

Further taking symmetric and antisymmetric part in ik of the above gives

$$\Phi^{j,ik}_{02TT} = C\eta^{\gamma-1}\Phi^{j,(ik)}_{11TT} + \frac{1}{2}C^{-1}\eta^{-\gamma}\Phi^{j,(ni)}_{11TT}H_{kn} + \frac{1}{2}C^{-1}\eta^{-\gamma}\Phi^{j,(nk)}_{11TT}H_{in} \tag{73}$$

$$0 = C\eta^{\gamma-1}\Phi^{j,[ik]}_{11TT} + \frac{1}{2}C^{-1}\eta^{-\gamma}\Phi^{j,(ni)}_{11TT}H_{kn} - \frac{1}{2}C^{-1}\eta^{-\gamma}\Phi^{j,(nk)}_{11TT}H_{in}\,. \tag{74}$$

The other condition comes from the third line of (59) which, using (37) and (60), and neglecting subleading terms in η, gives

$$\eta\Phi^{j,ik}_{02}H_{li} + \eta\Phi^{j,il}_{02}H_{ki} + \Phi^{j,ik}_{11}H_{li} + \Phi^{j,il}_{11}H_{ki} + C^{-1}\eta^{1-\gamma}P^{ji}_{L}\partial_{k}H_{li} = 0 \tag{75}$$

Projecting the k and l indices on the longitudinal part, gives $P^{ji}_{L}\partial^{L}_{k}H_{li}P^{ln}_{L} = 0$, which is an identity that can be checked from (52). Projecting k on the longitudinal part and l on the transverse part gives

$$\eta\Phi^{j,ik}_{02TL}H_{li} + \Phi^{j,ik}_{11TL}H_{li} + C^{-1}\eta^{1-\gamma}P^{ji}_{L}\partial^{L}_{k}H_{ni}P^{nl}_{T} = 0\,, \tag{76}$$

which is implied by (64), indeed

$$\begin{aligned} (\eta\Phi^{j,ik}_{02TL} + \Phi^{j,ik}_{11TL})H_{li} &= C^{-1}\eta^{1-\gamma}\Phi^{j,ik}_{02TL}H_{li} \\ &= -C^{-1}\eta^{1-\gamma}P^{jl}_{L}\partial^{L}_{k}H^{ln}(H^{\dagger})^{ni}H_{li} = -C^{-1}\eta^{1-\gamma}P^{jl}_{L}\partial^{L}_{k}H^{ln}P^{nl}_{T}\,, \end{aligned} \tag{77}$$

therefore (76) does not give a new condition. The only new condition comes from projecting the k and l indices of (75) on the transverse direction, giving

$$\eta\Phi^{j,ik}_{02TT}H_{li} + \eta\Phi^{j,il}_{02TT}H_{ki} + \Phi^{j,ik}_{11TT}H_{li} + \Phi^{j,il}_{11TT}H_{ki} + C^{-1}\eta^{1-\gamma}P^{ji}_{L}\partial^{T}_{k}H_{ni}P^{nl}_{T} = 0\,. \tag{78}$$

Plugging in (72),

$$\begin{aligned} C\eta^{\gamma}\Phi^{j,ik}_{11TT}H_{li} + C^{-1}\eta^{1-\gamma}\Phi^{j,(in)}_{11TT}H_{kn}H_{li} + C\eta^{\gamma}\Phi^{j,il}_{11TT}H_{ki} + C^{-1}\eta^{1-\gamma}\Phi^{j,(in)}_{11TT}H_{ln}H_{ki} \\ +\Phi^{j,ik}_{11TT}H_{li} + \Phi^{j,il}_{11TT}H_{ki} + C^{-1}\eta^{1-\gamma}P^{ji}_{L}\partial^{T}_{k}H_{ni}P^{nl}_{T} = 0\,. \end{aligned} \tag{79}$$

Neglecting subleading terms in η this can be rewritten as

$$2C^{-1}\eta^{1-\gamma}\Phi^{j,(in)}_{11TT}H_{kn}H_{li} + \Phi^{j,ik}_{11TT}H_{li} + \Phi^{j,il}_{11TT}H_{ki} + C^{-1}\eta^{1-\gamma}\partial^L_j H_{ni}P^{ki}_T P^{nl}_T = 0\,. \tag{80}$$

where we recall that the last term is symmetric in k and l. In matrix notation this reads

$$C^{-1}\eta^{1-\gamma}H(\Phi^j_{11TT} + \Phi^{jt}_{11TT})H + H\Phi^j_{11TT} + \Phi^{jt}_{11TT}H + C^{-1}\eta^{1-\gamma}M^{(j)} = 0\,, \tag{81}$$

where

$$(M^{(j)})_{kl} = P^{ki}_T\partial^L_j H_{ni}P^{nl}_T \tag{82}$$

is the longitudinal derivative of the Hessian, projected on the transverse directions. Decomposing Φ^j_{11TT} into symmetric and antisymmetric parts $\Phi^j_{11TT} = S^{(j)} + A^{(j)}$, Eq. (81) reads (suppressing the index j)

$$2C^{-1}\eta^{1-\gamma}HSH + HS + SH + HA - AH + C^{-1}\eta^{1-\gamma}M = 0\,. \tag{83}$$

The first term is subleading in η, therefore we have

$$HS + SH + HA - AH + C^{-1}\eta^{1-\gamma}M = 0\,. \tag{84}$$

This equation together with (74), which in matrix form reads

$$C^{-1}\eta^{-\gamma}(SH - HS) + 2C\eta^{\gamma-1}A = 0\,, \tag{85}$$

determine S and A, and thus Φ_{11TT}. Note that M^j has only a longitudinal component in the index j, therefore the transverse parts of S and A vanish, i.e.

$$P^{pj}_T\Phi^{j,ik}_{11TT} = 0\,. \tag{86}$$

Eq. (86) is natural as the slow degrees of freedom are the longitudinal coordinates.

Note that these two equations admit a unique solution. Indeed, solving (85) for A gives

$$A = \frac{1}{2}C^{-2}\eta^{1-2\gamma}[H, S] \tag{87}$$

Plugging in (84), we find

$$\tilde{\mathcal{L}}_H S = -C^{-1}\eta^{1-\gamma}M\,, \tag{88}$$

where $\tilde{\mathcal{L}}_H S \equiv \{H, S\} + \frac{1}{2}C^{-2}\eta^{1-2\gamma}[[S, H], H]$ is introduced in definition C.2. By Lemma C.3, (88) admits a unique solution.

Then

$$\Phi^j_{11TT} = -C^{-1}\eta^{1-\gamma}\tilde{\mathcal{L}}_H^{-1}M^{(j)} - \frac{1}{2}C^{-3}\eta^{2-3\gamma}[H, \tilde{\mathcal{L}}_H^{-1}M^{(j)}]\,. \tag{89}$$

The other components are (see eqs. (56),(57),(63),(64),(72),(60))

$$\Phi^{j,ik}_{02LL} = -(H^\dagger)^{jl}\partial^L_i H^{ln}P^{nk}_L \tag{90}$$

$$\Phi^{j,ik}_{02LT} = -P^{jl}_L\partial^L_i H^{ln}(H^\dagger)^{nk} \tag{91}$$

$$\Phi^{j,ik}_{11LL} = -C^{-1}\eta^{1-\gamma}(H^\dagger)^{jl}\partial^L_i H^{ln}P^{nk}_L \tag{92}$$

$$\Phi^{j,ik}_{11TL} = -C^{-1}\eta^{1-\gamma}P^{jl}_L\partial^L_k H^{ln}(H^\dagger)^{ni} \tag{93}$$

$$\Phi^{j,ik}_{02TT} = C\eta^{\gamma-1}\Phi^{j,ik}_{11TT} + C^{-1}\eta^{-\gamma}\Phi^{j,(ni)}_{11TT}H_{kn} = O\left(\eta^{\min\{0,1-2\gamma\}}\right) \tag{94}$$

$$\Phi^{ij}_{20} = C^{-1}\eta^{1-\gamma}\Phi^{(ij)}_{11} + \frac{1}{2}C^{-1}\eta^{2-\gamma}\Phi^{ij}_{02} = C^{-1}\eta^{1-\gamma}\Phi^{(ij)}_{11}\,. \tag{95}$$

The only contributing term at leading order in η to the limiting diffusion equation is (95). Splitting it into longitudinal and transverse components, we find:

$$\Phi^{j,ik}_{20LL} = C^{-1}\eta^{1-\gamma}\Phi^{j,(ik)}_{11LL} = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}\left((H^\dagger)^{jl}\partial^L_i H^{ln}P^{nk}_L + (H^\dagger)^{jl}\partial^L_k H^{ln}P^{ni}_L\right) \tag{96}$$

$$\Phi^{j,ik}_{20TL} = C^{-1}\eta^{1-\gamma}\Phi^{j,(ik)}_{11TL} = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}\left(P^{jl}_L\partial^L_k H^{ln}(H^\dagger)^{ni} + P^{jl}_L\partial^L_i H^{ln}(H^\dagger)^{nk}\right) \tag{97}$$

and, using (89) and (82) we have, in matrix notation,

$$\Phi^j_{20TT} = \frac{1}{2}C^{-1}\eta^{1-\gamma}(\Phi^j_{11TT} + (\Phi^j_{11TT})^T) = -C^{-2}\eta^{2-2\gamma}\tilde{\mathcal{L}}_H^{-1}(P_T\partial^L_j HP_T)\,. \tag{98}$$

□

To write things more compactly, the following Lemma will be useful:

Lemma C.5. *For any transverse symmetric matrix T:*

$$\Phi^j_{20TT}[T] = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}M^{(j)}[\tilde{\mathcal{L}}_H^{-1}T]\,, \tag{99}$$

Proof. From (50), Φ_{20TT} is proportional to the symmetric part of Φ_{11TT} which was denoted by S below Eq. (82) and satisfies Eq. (88). Therefore Φ_{20TT} also satisfies Eq. (88), up to an overall factor:

$$\tilde{\mathcal{L}}_H\Phi^j_{20TT} = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}M^{(j)}\,. \tag{100}$$

Then, for any T:

$$(\tilde{\mathcal{L}}_H\Phi^j_{20TT})[T] = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}M^{(j)}[T]\,. \tag{101}$$

Moreover,

$$(\tilde{\mathcal{L}}_H\Phi^j_{20TT})[T] = \Phi^j_{20TT}[\tilde{\mathcal{L}}_H T]\,. \tag{102}$$

Since the two above equations hold for any T and $\tilde{\mathcal{L}}_H$ is invertible, this implies

$$\Phi^j_{20TT}[T] = -\frac{1}{2}C^{-2}\eta^{2-2\gamma}M^{(j)}[\tilde{\mathcal{L}}_H^{-1}T]\,, \tag{103}$$

which is the statement of the lemma. □

To leading order in η we then have, for a symmetric matrix V, $\partial^2\Phi[V] = \sum_{i,j=1}^D \tilde{\partial}_i\tilde{\partial}_j\Phi V_{ij}$, and thus

$$\begin{aligned}\partial^2\Phi[V] = &-\frac{1}{2C^2}\eta^{2-2\gamma}(\nabla^2 L)^\dagger\partial^2(\nabla L)[V_{LL}]dt - \frac{1}{C^2}\eta^{2-2\gamma}P_L\partial^2(\nabla L)[(H^\dagger)V_{TL}]dt\\ &-\frac{1}{2C^2}\eta^{2-2\gamma}P_L\partial^2(\nabla L)[\tilde{\mathcal{L}}_H^{-1}V_{TT}]\,,\end{aligned} \tag{104}$$

where $V_{LL} = P_L V P_L$, $V_{TL} = P_T V P_L$, and $V_{TT} = P_T V P_T$ are transverse and longitudinal projections of V.

We also have, using (37), and to leading order in η,

$$\partial\Phi\tilde{\sigma}dZ = \Phi_{10}\sigma dZ + \eta\Phi_{01}\sigma dZ = (C^{-1}\eta^{1-\gamma}+\eta)P_L\sigma dW\,, \tag{105}$$

where dW is a Wiener process. Applying Theorem B.4, and keeping into account that, to leading order in dt, $d[Z^i, Z^j] = \delta^{ij}dt$, we find

$$\begin{aligned}dY =&(C^{-1}\eta^{1-\gamma}+\eta)P_L\sigma dZ - \frac{1}{2C^2}\eta^{2-2\gamma}(\nabla^2 L)^\dagger\partial^2(\nabla L)[\Sigma_{LL}]dt\\ &-\frac{1}{C^2}\eta^{2-2\gamma}P_L\partial^2(\nabla L)[(H^\dagger)\Sigma_{TL}]dt - \frac{1}{2C^2}\eta^{2-2\gamma}P_L\partial^2(\nabla L)[\tilde{\mathcal{L}}_H^{-1}\Sigma_{TT}]dt\,,\end{aligned} \tag{106}$$

where $\Sigma = \sigma\sigma^T$.

For $\gamma < \frac{1}{2}$, $\tilde{\mathcal{L}}_H$ reduces to the Lyapunov operator at leading order in η, i.e. $\tilde{\mathcal{L}}_H S = \{H, S\}$. For $\gamma > \frac{1}{2}$, from (44), it is easy to see that the role of the divergent term proportional to $\eta^{1-2\gamma}$, when acting $\tilde{\mathcal{L}}_H^{-1}$ on S, is to set to zero the off-diagonal entries of S_{ij} at $O(\eta^{1-\gamma})$, i.e.

$$S_{ii} = -2C^{-1}\eta^{1-\gamma}\lambda_i M_{ii}, \qquad S_{i\neq j} = 0,. \tag{107}$$

Using Lemma B.5, we finally conclude the following Corollary, which is the formal version of Theorem 3.4 in the main text:

Corollary C.6. *Consider the stochastic process defined in Eq. (28) parametrized by ϵ_n, with initial conditions $(\pi_0, w_0) \in U$, under assumptions 3.1 and 3.2. Fix a compact $K \subset U$. Then the conclusions of Theorem B.4 apply, and $Y(t)$ satisfies the limiting diffusion equation*

$$\begin{aligned}dY =&(\tfrac{1}{C}\eta^{1-\gamma}+\eta)P_L\sigma dW - \tfrac{1}{2C^2}\eta^{2-2\gamma}(\nabla^2 L)^\dagger\partial^2(\nabla L)[\Sigma_{LL}]dt\\ &-\tfrac{1}{C^2}\eta^{2-2\gamma}P_L\partial^2(\nabla L)[(\nabla^2 L)^\dagger\Sigma_{TL}]dt - \tfrac{1}{2C^2}\eta^{2-2\gamma}P_L\partial^2(\nabla L)[\tilde{\mathcal{L}}^{-1}_{\nabla^2 L}\Sigma_{TT}]dt\,,\end{aligned} \tag{108}$$

where $W(t)$ is a Wiener process.

Let us now see the special case of label noise. In this case $\Sigma = cH$, so that Σ is only transverse. Moreover, using (44),

$$\tilde{\mathcal{L}}_H^{-1} H = \frac{1}{2} P_T \,, \tag{109}$$

and

$$dY = -\frac{1}{4C^2}\eta^{2-2\gamma} P_L \partial^2(\nabla L)[cP_T]dt = -\frac{1}{4C^2}\eta^{2-2\gamma} P_L \nabla \mathrm{Tr}(c\partial^2 L)dt \,. \tag{110}$$

This proves Corollary 3.5 in the main text.

D Effective Drift in UV model

We start with mean-square loss

$$L = \frac{1}{2P}\sum_{a=1}^{P} ||f(x^a) - y^a||^2, \tag{111}$$

with the data covariance matrix Σ as defined in the text. The trace of the Hessian on the zero loss manifold ($L(w^*) = 0$) is given explicitly by

$$\mathrm{tr} H = \frac{1}{n}\left(m\mathrm{Tr}\left(\Sigma V^T V\right) + \mathrm{Tr}\,\Sigma\,\mathrm{Tr}\left(UU^T\right)\right). \tag{112}$$

Taking gradients of this and plugging into Corrollary 3.5, repeated in Eq. (110) leads to an explicit expression for the drift-diffusion along the manifold

$$dY = -\frac{\delta^2\eta^{2-2\gamma}}{4PC^2}\frac{2}{n}\hat{S}_L Y dt, \quad \hat{S} = \begin{pmatrix} \mathrm{tr}\Sigma & 0 \\ 0 & m\Sigma \end{pmatrix}, \quad \hat{S}_L = P_L \hat{S} P_L, \tag{113}$$

The simplification we cite in the main text in Sec. 3.4 is due to the fact that for input and output dimension $d = m = 1$, we have that $\Sigma = \mathrm{Tr}\Sigma = \mu_2$, and $\hat{S}$ is proportional to the identity.

E Linearization Analysis of Momentum Gradient Descent in the Scaling Limit

Here we elaborate on the discussion in Sec. 3.3, providing derivations of various results. We take a straightforward linearization of the deterministic (noise-free) gradient descent with momentum. Working in the extended phase space $x = (\pi, w)$, the dynamical updates are of the form

$$x_{t+1} = x_t + F(x_t), \quad F(x_t) = \begin{pmatrix} (\beta - 1)\pi_t - \nabla L(w_t) \\ \eta(\beta\pi_t - \nabla L(w_t)) \end{pmatrix}. \tag{114}$$

The fixed point of the evolution $x^* = (0, w^*)$ will have the momentum variable $\pi = 0$, and the coordinate satisfying $\nabla L(w^*) = 0$. Linearizing the update (114) around this point

$$\delta x_{t+1} = J(x^*)\delta x_t, \quad J(x^*) = \begin{pmatrix} \beta & -\nabla^2 L(w^*) \\ \eta\beta & 1 - \eta\nabla^2 L(w^*) \end{pmatrix}. \tag{115}$$

Note $\nabla^2 L(w^*)$ is the Hessian of the loss function at the fixed point. The spectrum of the Jacobian can be written in terms of the eigenvalues of the Hessian λ_i. This is accomplished by using a straightforward ansatz for the (unnormalized) eigenvectors of the Jacobian $k^i = (\mu^i q^i, q^i)$, where q^i are eigenvectors of the Hessian with eigenvalue λ_i. Solving the resulting coupled eigenvalue equations for eigenvalue κ^i:

$$1 - \eta\lambda_i + \eta\beta\mu^i = \kappa^i, \quad -\lambda_i + \mu^i\beta = \mu^i\kappa^i. \tag{116}$$

For a fixed λ_i, there will be two solutions given by

$$\kappa^i_\pm = \frac{1}{2}\left(1+\beta-\eta\lambda_i \pm \sqrt{(1+\beta-\eta\lambda_i)^2-4\beta}\right), \quad i=1,...,D \tag{117}$$

$$\mu^i_\pm = \frac{1}{2\beta\eta}\left(\beta-1+\eta\lambda_i \pm \sqrt{(1+\beta-\eta\lambda_i)^2-4\beta}\right). \tag{118}$$

For the set of zero modes $\lambda_i = 0$, we get the following modes: $\kappa_+ = 1$, corresponding to motion only along w, with eigenvector $k^i = (0, q^i)$. In addition, there is a **mixed longitudinal mode** which includes a component of π along the zero manifold $k^i = (\mu_- q^i, q^i)$, and has an eigenvalue $\kappa_- = \beta$.

On the zero loss manifold, we can assume the Hessian is positive semi-definite, and that the positive eigenvalues satisfy

$$0 < c_1 \leq \lambda_i \leq c_2. \tag{119}$$

for constants c_1, c_2 independent of η, β.

We now analyze the spectrum of the Jacobian one eigenvalue at a time, and then use these results to informally control the relaxation rate of off-manifold perturbations. It is useful first to consider the conditions for stability, i.e. $|\kappa^i| < 1$, which are stated in (120,121) below:

$$\text{Case 1}: \text{ If } \eta\lambda_i < (1-\sqrt{\beta})^2, \text{ then } \kappa^i_\pm \in \mathbb{R} \text{ and } |\kappa^i_\pm| < 1 \text{ iff } 0 < \eta\lambda_i < 2(1+\beta). \tag{120}$$

$$\text{Case 2}: \text{ If } \eta\lambda_i > (1-\sqrt{\beta})^2, \text{ then } \kappa^i_\pm \in \mathbb{C} \text{ and } |\kappa^i_\pm| = \sqrt{\beta} < 1. \tag{121}$$

Proof of Case 1: The condition $\eta\lambda_i < (1-\sqrt{\beta})^2$ implies $A^2 > 4\beta$, where $A = 1+\beta-\eta\lambda_i$. The condition for stability then requires $-1 < \kappa_i < +1$. We satisfy both sides of this inequality:

If $1+\beta-\eta\lambda_i = A > 0$, then $|\kappa^i_-| < 1$,and $\kappa^i_+ > 0$, so we simply require $\kappa^i_+ < 1$, i.e.

$$\kappa_i < +1 \tag{122}$$

$$\frac{1}{2}(A+\sqrt{A^2-4\beta}) < 1 \tag{123}$$

$$+\sqrt{A^2-4\beta} < 2-A \tag{124}$$

$$A^2-4\beta < 4-4A+A^2 \tag{125}$$

$$A = 1+\beta-\eta\lambda_i < 1+\beta \quad \Rightarrow \eta\lambda_i > 0 \tag{126}$$

If $1+\beta-\eta\lambda_i < 0$, then $|\kappa^i_+| < 1$, and $\kappa^i_- < 1$, so we only require $\kappa^i_- > -1$

$$-1 < \kappa^i_- \tag{127}$$

$$-1 < \frac{1}{2}(-|A|-\sqrt{A^2-4\beta}) \tag{128}$$

$$|A|-2 < -\sqrt{A^2-4\beta} \tag{129}$$

$$2-|A| > \sqrt{A^2-4\beta} \tag{130}$$

$$A^2-4|A|+4 > A^2-4\beta \tag{131}$$

$$|A| = -1-\beta+\eta\lambda_i < 1+\beta \quad \Rightarrow \eta\lambda_i < 2(1+\beta) \tag{132}$$

Proof of Case 2: the condition $\eta\lambda_i > (1-\sqrt{\beta})^2$ implies $A^2 < 4\beta$, where $A = (1+\beta-\eta\lambda_i)$. The corresponding eigenvalues of the Jacobian can be written $\kappa^\pm_i = (1/2)(A \pm \sqrt{A^2-4\beta}) = (1/2)(A \pm i\sqrt{4\beta-A^2})$. Computing the absolute value then gives $|\kappa^\pm_i|^2 = (1/4)(A^2+4\beta-A^2) = \beta$ □

These results show us when the GD+momentum is stable. Next, assuming stability, we want to estimate the rate of convergence to the fixed point. More precisely, we would like to determine the

fastest mode as well as the slowest mode. To this end, we define two quantities

$$\rho_1 = \max\{|\kappa^i_\pm|\big|\ i = 1, ..., D,\ |\kappa^i_\pm| < 1\}. \tag{133}$$

$$\rho_2 = \min\{|\kappa^i_\pm|\big|\ i = 1, ..., D,\ |\kappa^i_\pm| < 1\}. \tag{134}$$

Using the explicit scaling for momentum, and in the limit of small learning rate, we prove the following for ρ_1:

Lemma E.1. *Let $\beta = 1 - C\eta^\gamma$, and η sufficiently small:*

For $\gamma < 1/2$, the condition for Case 1 (120) holds,

$$\rho_1 \approx 1 - (c_1/C)\eta^{1-\gamma}, \tag{135}$$

$$\rho_2 \approx 1 - C\eta^\gamma. \tag{136}$$

.

For $\gamma > 1/2$, the condition for Case 2 (121), and

$$\rho_1 = \sqrt{\beta} \approx 1 - (C/2)\eta^\gamma, \tag{137}$$

$$\rho_2 = \beta = 1 - C\eta^\gamma. \tag{138}$$

.

Proof: For small η, we find that

$$\eta^{-1}(1 - \sqrt{\beta})^2 = \left(1 - \sqrt{1 - C\eta^\gamma}\right)^2 \approx C^2\eta^{2\gamma-1}/4. \tag{139}$$

When $\gamma > 1/2$, this expression tends to zero as $\eta \to 0$. Therefore, for sufficiently small η, the condition for Case 2 (121) will be satisfied and $|\kappa^i| = \sqrt{\beta}$ for all $\lambda_i > 0$. Since $\beta < \sqrt{\beta}$, this implies $\rho_1 = \sqrt{\beta}$ and $\rho_2 = \beta$. The scaling behavior in the Lemma follows by substitution.

For $\gamma < 1/2$, (139) diverges as $\eta \to 0$, which means Case 1 (120) will obtain for all λ_i. Next, since for small η, $1 + \beta - \eta\lambda_i > 1 + \beta - \eta c_2 = 2 - C\eta^\gamma - \eta c_2 > 0$, since C and c_2 are order one constants, and $\eta \to 0$. In this case, the largest contribution from the nonzero eigenvalues λ_i will come from κ^i_+. In particular, we find

$$|\kappa^i_+| \leq \frac{1}{2}\left(1 + \beta - \eta c_1 + \sqrt{(1 + \beta - \eta c_1)^2 - 4\beta}\right), \tag{140}$$

$$= \frac{1}{2}\left(2 - C\eta^\gamma - \eta c_1 + \sqrt{C^2\eta^{2\gamma} - 2(2 - C\eta^\gamma)\eta c_1 + \eta^2 c_1^2}\right). \tag{141}$$

For $\gamma < 1/2$, we have the hierarchy $\eta^\gamma > \eta^{2\gamma} > \eta > \eta^{\gamma+1} > \eta^2$. This allows us to simplify the upper bound

$$\rho_1 = \max|\kappa^i_+| \leq 1 - \frac{c_1}{C}\eta^{1-\gamma} + O(\eta). \tag{142}$$

Next, we find a lower bound for ρ_2. This will be controlled by κ^i_-. We may use then that

$$\rho_2 = \min\{|\kappa^i_-|\} \geq \frac{1}{2}\left(1 + \beta - \eta c_2 - \sqrt{(1 + \beta - \eta c_1)^2 - 4\beta}\right), \tag{143}$$

$$= \frac{1}{2}\left(2 - C\eta^\gamma - \eta c_2 - \sqrt{C^2\eta^{2\gamma} - 2(2 - C\eta^\gamma)\eta c_1 + \eta^2 c_1^2}\right) \approx 1 - C\eta^\gamma + O(\eta^{1-\gamma}) \tag{144}$$

Finally, note that since $\eta^{1-\gamma} < \eta^\gamma$, we have that $1 - \frac{c_1}{C}\eta^{1-\gamma} > 1 - C\eta^\gamma = \beta$, so indeed $\rho_2 < \rho_1$.

Equipped with the upper and lower bounds on the spectrum leads naturally to bounds on the relaxation rate. For the purely decaying modes at $\gamma < 1/2$, we use $\rho_2^t \leq |\delta x_t^T| \leq \rho_1^t$, δx_t^T represents the

projection of the fluctuations δx_t onto the transverse and mixed longitudinal modes. After applying Eqs. (135) and (136), we arrive at the result quoted in the main text in Sec. 3.3. For $\gamma > 1/2$, the modes are oscillatory. However, the eigenvalues within the unit circle have norm either $\sqrt{\beta}$, β, as reflected in Eqs. (137) and (138). This implies that we can estimate the decay rate of the envelope of the transverse and mixed longitudinal modes in this regime, thereby arriving at second expression quoted in the main text in Sec. 3.3.