Journal of Machine Learning Research 27 (2026) 1-56 Submitted 4/24; Revised 1/25; Published 7/26

Singular-limit analysis of gradient descent with noise injection

Anna Shalova A.SHALOVA@UVA.NL
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
P.O. Box 94248, 1090 GE Amsterdam, The Netherlands

André Schlichting ANDRE.SCHLICHTING@UNI-ULM.DE
Institute of Applied Analysis
University Ulm
Helmholtzstraße 18, 89081 Ulm, Germany

Mark Peletier M.A.PELETIER@TUE.NL
Department of Mathematics and Computer Science
Eindhoven University of Technology
PO Box 513, 5600 MB Eindhoven, The Netherlands

Editor: Martin Jaggi

Abstract

We study the limiting dynamics of a large class of noisy gradient descent systems in the overparameterized regime. In this regime the *zero-loss set* of global minimizers of the loss is large, and when initialized in a neighbourhood of this zero-loss set a noisy gradient descent algorithm slowly evolves along this set. In some cases this slow evolution has been related to better generalization properties. We characterize this evolution for the broad class of noisy gradient descent systems in the limit of small learning rate.

Our results show that the structure of the noise affects not just the form of the limiting process, but also the time scale at which the evolution takes place. We apply the theory to Dropout, label noise and classical SGD (minibatching) noise, and show that these evolve on two different time scales. Classical SGD even yields a trivial evolution on both time scales, implying that additional noise is required for regularization.

The results are inspired by the training of neural networks, but the theorems apply to noisy gradient descent of any loss that has a non-trivial zero-loss set.

Keywords: Noise injection, stochastic optimization, stochastic gradient descent, zero-loss set, overparameterization, regularization, dropout.

1. Introduction

1.1 Noise injection

Modern machine learning and especially the training of neural networks rely on gradient descent and its many variants. Many of those variants introduce *noise* (randomness) into the algorithm, for instance through minibatching, Dropout, or random corruption of the labels. This noise may be motivated by practical considerations, as in the case of minibatching, but in many cases it is observed that the noise also improves the quality of the resulting parameter point. In particular, the noise often leads to parameter points that generalize better. It

would be of great practical value to understand this *implicit bias* of noisy algorithms, and this currently is an active area of research.

In a seminal paper, Li, Wang, and Arora (2022a) focused on a specific class of noisy gradient-descent algorithms, and showed how the particular form of the noise leads to improved generalization. They focused on overparameterized systems, in which the *zero-loss set* $\Gamma := \{w : L(w) = 0\}$ is a high-dimensional manifold. Their key observation is that gradient descent behaves differently with and without noise: in a neighbourhood of Γ, deterministic gradient descent converges to Γ and then stops, while noisy gradient descent may continue to evolve after reaching Γ. Figure 1 below shows an example of this. Li, Wang, and Arora characterized this continuing evolution in a small-learning-rate limit, and showed its relation to generalization.

The aim of this paper is to generalize the observations of Li et al. (2022a) to a much wider class of systems, and characterize the behaviour of these systems when they evolve in the neighbourhood of the zero-loss set Γ. This generalization was inspired by the case of Dropout, but the resulting setup covers many more types of noise. This generality also allows us to identify a hierarchy in scaling of different types of noise, such as minibatch noise, label noise, Dropout, and others. As it turns out, the case studied by Li et al. (2022a) is not the first but the second level in this hierarchy, with for instance Dropout occupying the first and more dominant level.

While this work is inspired by the training of neural networks, the main characterizations do not use any neural-network structure, and therefore apply to noisy gradient descent in other application areas as well.

We now describe the main results of this paper, and we start by fixing some notation. *Gradient descent* for a loss function $L : \mathbb{R}^m \to [0, \infty)$ with parameter dimension $m \in \mathbb{N}$ and learning rate $\alpha > 0$ is the iterative algorithm

$$w_{k+1} = w_k - \alpha \nabla_w L(w_k), \qquad \text{for } k \geq 0\,.$$

In this paper we consider a class of random perturbations of gradient descent, that we call for short *noisy gradient descent*:

$$w_{k+1} = w_k - \alpha \nabla_w \hat{L}(w_k, \eta_k), \qquad \text{for } k \geq 0\,. \tag{1a}$$

Here $\hat{L} : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ is an extension of L, each η_k is a random vector in $\mathbb{R}^d$ satisfying

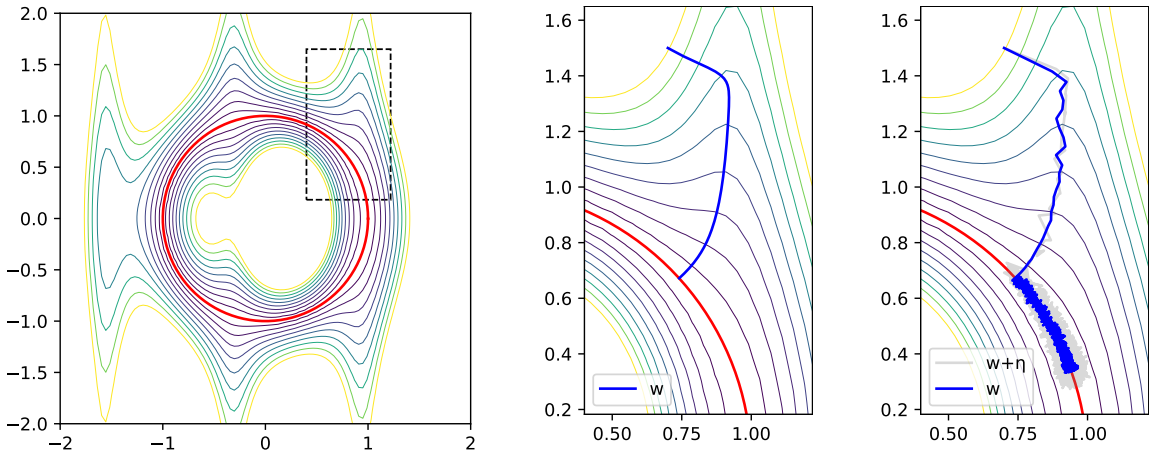
$$\mathbb{E}\,\eta_{k,i} = 0 \quad \text{and} \quad \operatorname{Var}\eta_{k,i} = \sigma^2, \qquad \text{for } i = 1, \dots, d\,, \tag{1b}$$

and $\eta_{k,i}$ is independent from $\eta_{\ell,j}$ for $(k,i) \neq (\ell,j)$. The connection between $\hat{L}$ and L is enforced by the following *consistency requirement* at $\eta = 0$:

$$\hat{L}(w, 0) = L(w) \qquad \text{for all } w \in \mathbb{R}^m\,. \tag{1c}$$

Apart from (1c) we only require a certain regularity of the function $(w, \eta) \mapsto \hat{L}(w, \eta)$ (for the full setup, see Section 3.2).

As mentioned above, many common forms of noise injection can be written in the form (1). In the training of neural networks, the most common form of noise arises from



(a) Level curves of the function L (b) Gradient descent (c) Noisy gradient descent

Figure 1: Noisy gradient descent may continue to move after reaching the zero-loss set Γ. The left-hand panel shows the level curves of a function $L : \mathbb{R}^2 \to [0, \infty)$, with the zero-level set Γ marked in red. The middle panel shows a gradient-descent evolution, starting at the top, and converging to Γ. The right-hand panel shows an evolution of the noisy gradient descent (1) with $\hat{L}(w, \eta) := L(w + \eta)$. See Appendix A for more details.

evaluating the loss on random minibatches; we discuss this in Section 5.2.1. Dropout (Hinton et al., 2012; Srivastava et al., 2014) is another example of (1), both in the more common 'Bernoulli' form and in the 'Gaussian' form. Other forms are 'label noise' (Blanc et al., 2020; Damian et al., 2021; Li et al., 2022a) and 'stochastic Langevin gradient descent' (Raginsky et al., 2017; Mei et al., 2018). We discuss all these in Section 5.

1.2 The non-degenerate case

As mentioned above, the focus of this paper lies on the behaviour of the noisy gradient descent (1) once it has entered a neighbourhood of the zero-loss set Γ. As illustrated by Figure 1c, the algorithm typically continues to evolve after reaching Γ, and in this paper we aim to characterize this continued evolution.

We motivate the main results by some heuristic calculations. Let a sequence $(w_k, \eta_k)_{k=0}^{\infty}$ be given by (1), and assume for simplicity that dynamics of the parameters w_k is slow compared to the noise variables η_k. It then becomes reasonable to average over the fast variables η_k:

$$w_{k+1} \approx w_k - \alpha \mathbb{E}_\eta[\nabla_w \hat{L}(w_k, \eta_k)].$$

If the variance σ^2 of the noise is small, then by Taylor expansion we find

$$\begin{aligned} w_{k+1} - w_k \approx -\alpha\, \mathbb{E}_\eta[\nabla_w \hat{L}(w_k, 0)] - \alpha\, \mathbb{E}_\eta[\partial_\eta \nabla_w \hat{L}(w_k, 0)[\eta_k]] \\ - \frac{\alpha}{2}\, \mathbb{E}_\eta[\partial_\eta^2 \nabla_w \hat{L}(w_k, 0)[\eta_k, \eta_k]]\,. \end{aligned} \tag{2}$$

(See Section 3.1 for our use of ∂.) Recall that $\hat{L}(w,0) = L(w)$, so that the first term in (2) is the gradient of the loss without noise. From the property (1b) we have $\mathbb{E}[\eta_k] = 0$ and $\mathbb{E}_\eta[\eta_k \otimes \eta_k] = \sigma^2 I$, leading to the resulting dynamics

$$w_{k+1} - w_k \approx -\alpha \nabla_w L(w_k) - \frac{\alpha}{2}\sigma^2 \nabla_w \Delta_\eta \hat{L}(w_k, 0) + o(\sigma^2) \qquad \text{as } \sigma^2 \to 0\,. \tag{3}$$

From the equation (3) we can recognize two phases in the evolution of w_k. In the first, faster phase, w_k has not yet reached Γ, and the first term $-\alpha\nabla L$ on the right-hand side dominates. This leads to an evolution at time scale $1/\alpha$.

When w_k reaches Γ, however, the term $-\alpha\nabla L$ becomes small, since $\nabla L = 0$ on Γ, and the second term on the right-hand side of (3) becomes important. This term leads to a slower evolution, at time scale $1/\alpha\sigma^2$, which is driven by the combination of the two terms on the right-hand side in (3).

These observations are made rigorous in the following non-rigorous formulation of our first main result, Theorem 23. We choose two sequences of hyperparameters $\alpha_n \to 0$ and $\sigma_n \to 0$, and for each n we construct a sequence $(w_k^n)_{k\geq 0}$ by (1), starting from (for simplicity) some fixed initial point w_0 sufficiently close to Γ. The theorem then states a convergence result for the time courses $(w_k^n)_{k\geq 0}$ as $n \to \infty$, after accelerating the evolution by a factor $1/\alpha_n\sigma_n^2$, thus capturing the second, slower phase of evolution along Γ.

Formal Theorem A (Non-degenerate case) *Assume $\alpha_n \to 0$ and $\sigma_n \to 0$. Define $W_n : [0,\infty) \to \mathbb{R}^m$ by*

$$W_n(\alpha_n\sigma_n^2 k) := w_k^n\,, \qquad \textit{with piecewise constant interpolation.}$$

Then the sequence W_n converges to a limit $W = (W(t))_{t\geq 0}$, that satisfies

$$\partial_t W(t) = -P_{T_{W(t)}\Gamma}\, \nabla_w \mathrm{Reg}(W(t)), \tag{4a}$$

$$W(t) \in \Gamma \quad \textit{for } t > 0, \tag{4b}$$

$$W(0) = w_0, \tag{4c}$$

$$W(0+) = \Phi(w_0) \in \Gamma \quad \textit{(see Definition 12).} \tag{4d}$$

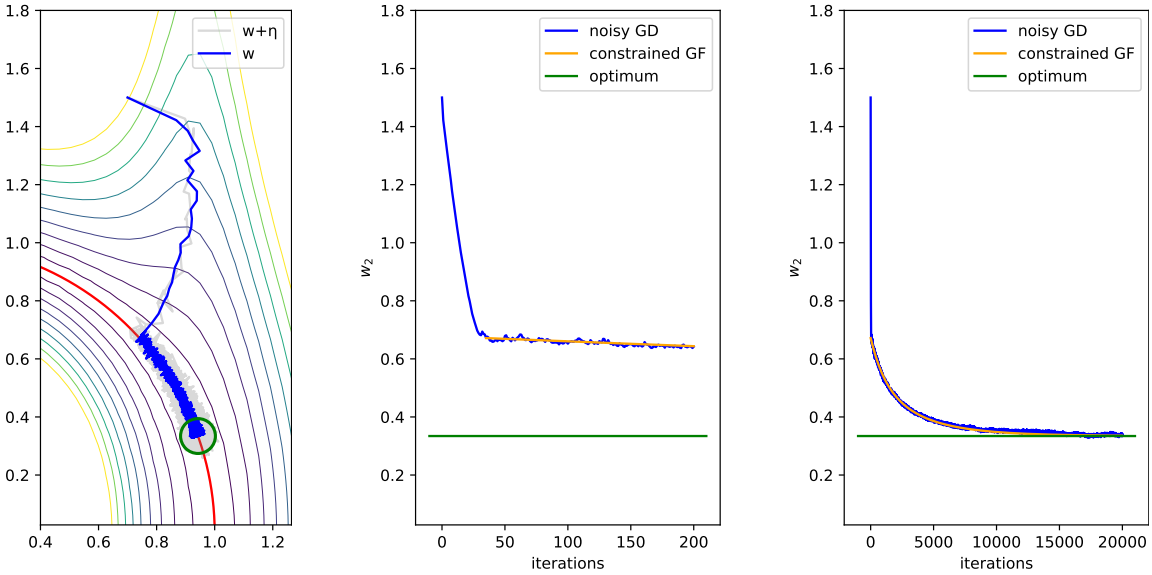
Here $P_{T_{W(t)}\Gamma}$ is the orthogonal projection onto the tangent plane of Γ at $W(t)$, and

$$\mathrm{Reg}(w) := \frac{1}{2}\Delta_\eta \hat{L}(w,0) = \frac{1}{2}\sum_{i=1}^{d} \frac{\partial^2}{\partial \eta_i^2}\hat{L}(w,\eta)\Big|_{\eta=0} \qquad \textit{for } w \in \Gamma.$$

Remark 1 (Biased gradient estimator) *Assume that the limiting system follows the dynamics described in Formal Theorem A with non-trivial $\nabla_w \Delta_\eta \hat{L}(w,0)$. Then, formally, the gradient $\nabla_w \hat{L}(w,\eta)$ satisfies*

$$\mathbb{E}\nabla_w \hat{L}(w,\eta) \approx \nabla_w \hat{L}(w,0) + \frac{1}{2}\sigma^2 \nabla_w \Delta_\eta \hat{L}(w,0) = \nabla_w L(w) + \frac{1}{2}\sigma^2 \nabla_w \Delta_\eta \hat{L}(w,0),$$

so $\nabla_w \hat{L}(w,\eta)$ is a biased gradient estimator. Thus, $\nabla_w \hat{L}(w,\eta)$ being a biased estimator of $\nabla_w L$ is a necessary condition for having a non-trivial dynamics at the time-scale $1/\alpha\sigma^2$. Remarkably, as we discuss in the next subsection, at slower time-scales the drift may arise even if the gradient estimator is unbiased.



(a) Noisy gradient descent as in Fig. 1c

(b) Coordinate w_2 of $w = (w_1, w_2) \in \mathbb{R}^2$ for noisy gradient descent and constrained gradient flow

(c) Same as (b), but over longer time

Figure 2: This figure shows the evolution of Fig. 1c in more detail, and compares it to the solution of the constrained gradient flow. Panel (a) is the same as Fig. 1c; panels (b) and (c) show the vertical coordinate over different periods of time (iterations). Panels (b) and (c) clearly illustrate the difference in time scales between the fast evolution towards Γ and the much slower evolution along Γ. The green circle and lines mark the minimizer of $\Delta_\eta \hat{L}(w, 0)$, which in this case equals $\Delta_w L(w)$. The orange curves are the solution of the constrained gradient flow (4), started from the first point at which w_k came close to Γ. More details are given in Appendix A.

Remark 2 (Constrained gradient flow) *The evolution* (4) *is a* constrained gradient flow *on* Γ*, driven by the function* Reg*. For all positive time* t *we have* $W(t) \in \Gamma$*, and the projection* $P_{T_{W(t)}\Gamma}$ *ensures that the right-hand side of* (4a) *is tangent to* Γ*, thus keeping* $W(t)$ *on* Γ*.*

If the initial point w_0 *is not on* Γ*, then the function* W *is discontinuous at* $t = 0$*; in that case the finite-*n *functions* W_n *show a fast evolution towards* Γ*, which becomes instantaneous in the limit* $n \to \infty$*. This is illustrated by the combination of* (4c) *and* (4d)*, and also by the fast initial transient in Figures 2b and 2c.*

Remark 3 (Projection onto the tangent plane) *The appearance of the projector* $P_{T_{W(t)}\Gamma}$ *in* (4) *can be understood in two complementary ways. To start with, if* W *remains in* Γ*, then the right-hand side in* (4) *has to be a tangent vector; since* $\nabla_w \Delta_\eta \hat{L}(w, 0)$ *has no reason to be tangent at* w*, the projection is necessary.*

The formal calculation (3) *also shows how this projection is generated in the evolution: if the increment* $-(\alpha\sigma^2/2)\, \nabla_w \Delta_\eta \hat{L}(w_k, 0)$ *pushes* w_{k+1} *away from* Γ*, then in the next iteration*

the first term $-\alpha\,\nabla_w L(w_{k+1})$ will not be zero, but point 'back to Γ'. Since the prefactors of the first and second term on the right-hand side of (3) *differ by a factor $\sigma^2 \ll 1$, the first term is asymptotically much stronger than the second, and this generates in the limit the projection.*

Example 1 (The example in Figures 1 and 2) *In the examples of Figures 1-2 we have chosen $\hat{L}(w,\eta) := L(w+\eta)$, and therefore the function* Reg *in Theorem A is*

$$\mathrm{Reg}(w) := \frac{1}{2}\Delta_\eta L(w+\eta)\Big|_{\eta=0} = \frac{1}{2}\Delta_w L(w).$$

The gradient flow (4) *of* Reg *along Γ therefore evolves towards points on Γ where $\Delta_w L$ is smaller, i.e. where the 'valley' around Γ is wider. In Figure 2 one can recognize the widening of the valley in the spreading of the level curves of L, and indeed the evolution converges to the minimizer of $\frac{1}{2}\Delta_w L$ on Γ, indicated by the green circle and line.*

Example 2 (Dropout) Dropout *is a particular form of noise injection that consists of 'dropping out' parameters or neurons randomly at each iteration. In Section 5 we discuss various types of Dropout; here we give one example, following Zhang and Xu (2024), in which the Dropout is only applied to the final layer, and in which Theorem 23 yields a strengthened version of (Zhang and Xu, 2024, Th. 7).*

To set up this form of Dropout, let f_w be a deep feedforward neural network that we write as

$$f_w : \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}, \qquad f_w(x) = \sum_{j=1}^{n} a_j \phi_j(x; w),$$

where $\phi_j(\,\cdot\,; w)$ represents the j^{th} output coordinate of all the layers of the network except the last one (i.e. the output of the j^{th} neuron of the penultimate layer), and the a_j are the weights of the last layer (see Section 5.3.2 for more details).

The 'last-layer Dropout' modification f_w^{drop} of f_w is of the form

$$f_w^{\mathrm{drop}}(x,\eta) = \sum_{j=1}^{n} a_j(1+\eta_j)\phi_j(x; w),$$

where η_j are i.i.d. random variables with law

$$\eta_j = \begin{cases} -1 & \text{w.p. } p\,, \\ \dfrac{p}{1-p} & \text{w.p. } 1-p\,. \end{cases}$$

With this choice, $\eta_j = -1$ corresponds to 'dropping' or 'killing' the output coordinate ϕ_j, and the other value $\eta_j = p/(1-p)$ corresponds to a rescaling such that $\mathbb{E}\,\eta_j = 0$. This choice satisfies (1b) *with $\sigma^2 = p/(1-p)$, and the limit $\sigma \to 0$ is the one in which a vanishingly small number of parameters are dropped.*

We consider the mean-square error loss for f_w^{drop},

$$\hat{L}(w,\eta) = \frac{1}{N}\sum_{i=1}^{N}(f_w^{\mathrm{drop}}(x_i,\eta) - y_i)^2.$$

For this case Theorem A establishes convergence of the noisy gradient descent, after acceleration by $1/\alpha_n\sigma_n^2$, to the constrained gradient flow on $\Gamma = \{w : \hat{L}(w,0) = 0\}$, driven by the function Reg*,*

$$\mathrm{Reg}(w) := \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{N} a_j^2\,\phi_j(x_i;w)^2$$

(see Corollary 52).

1.3 The degenerate case

In at least three important examples of noise injection, namely minibatching, label noise, and stochastic gradient Langevin descent, the procedure above applies, but the limiting constrained gradient flow is trivial: $\partial_t W = 0$. This is because for those examples it happens that $\Delta_\eta \hat{L}(w,\eta)$ is constant in w and η, and for this reason we call these types of noise *degenerate*.

To determine the behaviour for these degenerate forms of noise we return to the formal calculation (2). Instead of taking the expectation we write, assuming $\nabla_\eta^2 \hat{L} \equiv 0$ for simplicity,

$$\begin{aligned} w_{k+1} - w_k &\approx -\alpha \nabla_w \hat{L}(w_k,0) - \alpha \partial_\eta \nabla_w \hat{L}(w_k,0)[\eta_k] - \frac{\alpha}{2}\partial_\eta^2 \nabla_w \hat{L}(w_k,\eta^*)[\eta_k,\eta_k] \\ &= -\alpha\,\nabla_w L(w_k) - \alpha \partial_\eta \nabla_w \hat{L}(w_k,0)[\eta_k]\,. \end{aligned} \tag{5}$$

Take for instance the case that the $\eta_{k,i}$ are i.i.d. centered normal random variables with variance σ^2. Then the second term on the right-hand side is again a centered normal random variable, with variance that scales as $\alpha^2\sigma^2$. As a result, we expect that the noise has a non-trivial contribution when the accumulated quadratic variation is of order one, namely after $k \sim 1/\alpha^2\sigma^2$ steps.

These remarks motivate the following formal version of the second main result, Theorem 29. For this theorem we assume that the function $\hat{L}$ has the more specific form

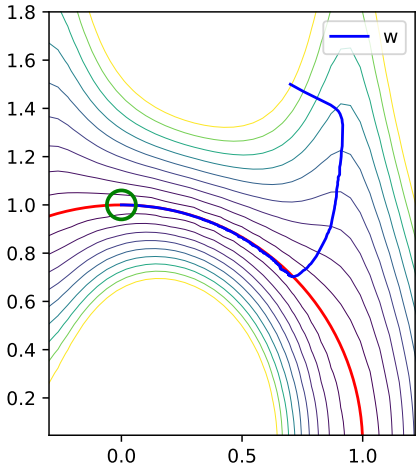
$$\hat{L}(w,\eta) = L(w) + f(w)\cdot\eta + \tfrac{1}{2}H(w) : (\eta\otimes\eta) + g(\eta), \tag{6}$$

for certain smooth maps f, H, and g. We assume that $g(0) = 0$ and that each diagonal element H_{ii} vanishes, so that $\Delta_\eta \hat{L}(w,0) = \sum_{i=1}^d H_{ii}(w) = 0$. It follows that the non-degenerate evolution (4) is trivial, and by the argument above we expect the evolution to take place at the slower time scale $1/\alpha^2\sigma^2$ (see Definition 28 for details).

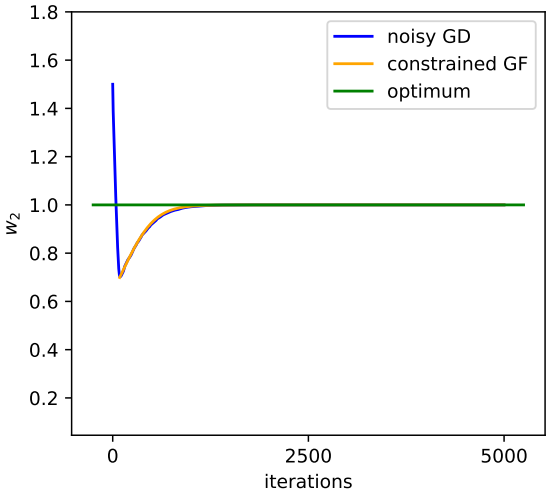
As for Formal Theorem A, we consider two sequences α_n and σ_n and the corresponding paths $(w_k^n)_{k\geq 0}$ given by (1), starting from some w_0 close to Γ.

Formal Theorem B (Degenerate case) *Let $\hat{L}$ be a degenerate loss function as described above. If $\alpha_n \to 0$ and $\sigma_n \to \sigma_0 \geq 0$, then define $W_n : [0,\infty) \to \mathbb{R}^m$ by*

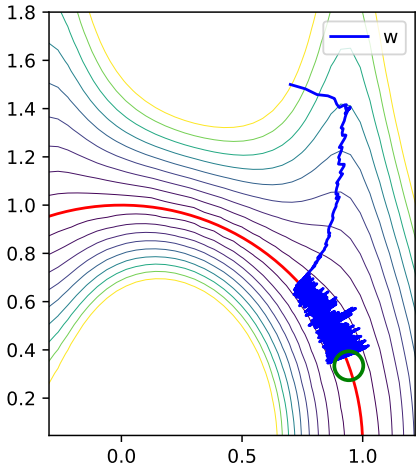
$$W_n(\alpha_n^2\sigma_n^2 k) := w_k\,, \qquad \textit{with piecewise constant interpolation.}$$



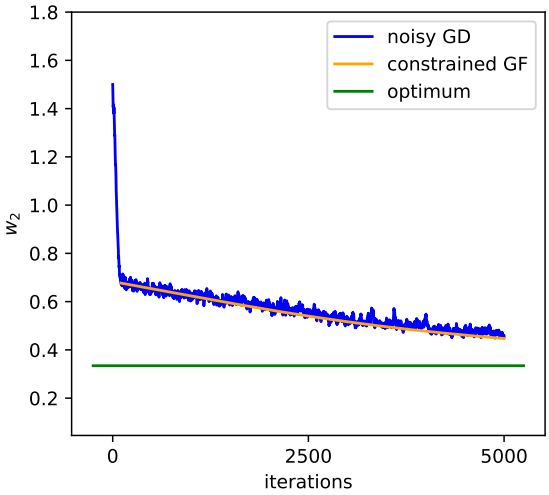
a: Non-degenerate noisy GD, $\hat{L}(w,\eta) = L(w) + \frac{1}{2}a(w)\eta^2$



b: Coordinate w_2 of noisy GD and constrained gradient flow



c: Degenerate noisy GD, $\hat{L}(w,\eta) = L(w) + \frac{1}{2}|w|^2\eta$



d: Coordinate w_2 of noisy GD and constrained gradient flow

Figure 3: Non-degenerate (top row) and degenerate evolution (bottom row); note how the evolution is much faster in the non-degenerate top row than in the degenerate bottom row. In both cases $\alpha = 0.1$ and η is a scalar centered normal random variable with variance $\sigma^2 = 0.01$. In the top row, $\hat{L}(w,\eta) = L(w) + \frac{1}{2}a(w)\eta^2$ with $a(w) = 1 - 0.7\cos 2w_1$. For this case Theorem A applies and gives the regularizer as $\frac{1}{2}\Delta_\eta \hat{L}(w,0) = 2a(w)$. In the bottom row $\hat{L}(w,\eta) = L(w) + \frac{1}{2}|w|^2\eta$, and Theorem B applies, with the limiting evolution (7) reducing to the deterministic constrained gradient flow $\partial_t W = -P_\Gamma \nabla_w \frac{1}{4}\log \Delta_w L(W)$. The green circles and lines mark the minimizers of the respective regularizers. More details are given in Appendix A.

Then the sequence W_n converges to a limit $W = (W(t))_{t>0}$ that satisfies

$$\mathrm{d}W(t) = P_{T_{W(t)}\Gamma}\beta(W(t))\,\mathrm{d}B_t + F(W(t)) : \beta(W(t))\beta(W(t))^\top\,\mathrm{d}t, \tag{7a}$$

$$W(t) \in \Gamma \quad \textit{for } t > 0, \tag{7b}$$

$$W(0) = w_0, \tag{7c}$$

$$W(0+) = \Phi(w_0) \in \Gamma \quad \textit{(see Definition 12)}. \tag{7d}$$

Here β and F are given in terms of f and H, and B is a multidimensional standard Brownian motion.

Remark 4 (Time scale) *The time scale is now $1/\alpha^2\sigma^2$, a factor $1/\alpha$ longer than in the non-degenerate case. Figure 3 illustrates this difference in speed with two functions $\hat{L}$ that are constructed from the same L as in Figures 1 and 2.*

Remark 5 (Limiting equation) *In contrast to the non-degenerate case, the limiting equation* (7) *is a constrained stochastic differential equation (SDE), not a constrained ODE. This difference arises from the fact that the second term on the right-hand side in* (5) *is now a mean-zero random variable of variance $O(\alpha^2\sigma^2)$, and the time scaling is such that we observe a sum of $O(1/\alpha^2\sigma^2)$ of these, leading to a random variable of variance $O(1)$.*

Example 3 (Minibatch noise) *In Section 5.2.1 we apply Theorem B to the case of minibatch noise. The resulting evolution is trivial* even on the time scale $1/\alpha_n^2\sigma_n^2$, *meaning that the evolution on Γ induced by minibatch noise is even slower than this. This also is consistent with the observation by Wojtowytsch (2023, 2024) that for small learning rates minibatch SGD may collapse onto Γ and become completely stationary.*

Example 4 (Label noise) *The result by Li et al. (2022a) also is a case of degenerate noise, and we revisit it in Section 5.2.2. For this case the limiting constrained SDE* (7) *is in fact a deterministic equation (i.e. the first term vanishes), in the form of a constrained gradient flow driven by the function* $\mathrm{Reg}(w) := (2N)^{-1}\Delta_w L(w)$.

See Section 5 for more examples.

1.4 Contributions

The main contributions of the paper are:

1. We introduce a specific class of 'gradient-descent algorithms with noise injection' that unifies a number of existing noise-injection schemes, such as minibatch SGD, Bernoulli and Gaussian Dropout, Bernoulli and Gaussian DropConnect, label noise, stochastic gradient Langevin descent, 'anti-correlated perturbed gradient descent', and others.

2. For this class of noise-injected gradient-descent schemes we give an explicit and rigorous characterization of the regularization induced by the noise in the non-degenerate (Theorem 23) and degenerate cases (Theorem 29), and we identify the corresponding time scales.

3. In particular, we prove convergence of gradient descent with Dropout and DropConnect noise to manifold-constrained gradient flows (Sections 5.1 and 5.3).

4. We demonstrate that a version of minibatch SGD shows no regularization effect on the two fastest time scales (Section 5.2.1) and does not change the form of the regularizer induced by Dropout or label noise (Section 5.2.2 and Remark 54).

2. Related Work

Convergence to stationary points. In the asymptotic analysis of stochastic gradient schemes, such as studied by Tadić (2015) and Fehrman et al. (2020), a large body of literature considers also the question of convergence of the iterates w_k to a single point as $k \to \infty$. For instance Tadić (2015, Section 3) proves such a statement for *stochastic gradient algorithms with Markovian dynamics*, which takes a form similar to the noisy gradient descent (1a). However, the main difference is that Tadić (2015) chooses the learning rate α to be k-dependent and to tend to zero as k tends to infinity, whereas we consider a fixed learning rate.

More recent works (Wojtowytsch, 2023; Dereich and Kassing, 2024) prove convergence of similar systems to a single point under a Łojasiewicz condition on the loss manifold in comparison to the local quadratic behavior, which we assume below (see Assumption 10).

These convergence results are consistent with the characterizations that we give in Theorems A and B, since vanishing learning rates may generate different types of behaviour from the fixed learning rates that we consider.

Noise injection: Minibatch SGD. The most common source of noise in training algorithms results from the use of random minibatches in gradient descent, often simply called 'stochastic gradient descent' (SGD). Experimentally this noise is known to lead to better generalization, and various studies have focused on determining the dependence of this effect on parameters such as the batch size (Keskar et al., 2017; Jastrzębski et al., 2017; Wu et al., 2018; Wojtowytsch, 2023) and learning rate (Jastrzębski et al., 2017; Hoffer et al., 2017; Wu et al., 2018; Wojtowytsch, 2023; Andriushchenko et al., 2023). The particular structure of minibatch noise has been shown to create a 'collapse' effect (Wojtowytsch, 2023, 2024) when the learning rate is small, and the same noise structure makes minibatch SGD incapable of selecting narrow minima (Wu et al., 2022). In this paper we also apply the techniques to a modification of minibatch SGD (Section 5.2.1).

Noise injection: Dropout. The most common form of Dropout is 'Bernoulli Dropout', i.e. randomly 'dropping' neurons, input nodes, or weights with probability p (Hinton et al., 2012; Wager et al., 2013; Wan et al., 2013), but the Gaussian Dropout in Section 5.3 has also been reported to give good results (Srivastava et al., 2014; Molchanov et al., 2017). Random networks generated by Dropout noise have the same universal-approximation properties as deterministic ones (Manita et al., 2022), and the convergence of Dropout-Gradient-Descent algorithms has been rigorously characterized (Senen-Cerda and Sanders, 2025, 2022).

A common explanation of the regularizing effect of Dropout centers on an interpretation of the Dropout-SGD iterates as a Monte-Carlo sampling of a deterministic augmented loss, where the loss term resembles L^2-type weight penalization. Various forms of this additional loss term have been derived, sometimes while assuming the Dropout noise to be

'small' (Baldi and Sadowski, 2013; Wager et al., 2013; Wang and Manning, 2013; Goodfellow et al., 2016; Mianjy et al., 2018; Mianjy and Arora, 2019; Zhang and Xu, 2024). Dropout-noise fluctuations have also been shown to have a significant effect in addition to this regularization-in-expectation (Wei et al., 2020; Zhang and Xu, 2024; Mianjy and Arora, 2020; Clara et al., 2024). Finally, very recently the effectiveness of Dropout regularization has been connected to 'weight expansion' (Jin et al., 2022).

Other types of noise injection. 'Label noise', the situation in which 'mistakes' are present in the data set, is a challenge for the training of classification methods (Frénay and Verleysen, 2013). One method to deal with this is to artifically perturb the labels during training; this can also be seen as a form of noise injection, and it recently has also been applied to the context of regression (Blanc et al., 2020; Li et al., 2022a). We comment on label noise in Section 5.2.2.

The 'anti-correlated' noise injection by Orvieto et al. (2022) is of the same form as we study in this paper, and we comment on this connection in Section 5.1.3. The algorithm studied by Duchi et al. (2012) is of similar form.

Evolution along the zero-loss manifold. For overparameterized networks the training takes place in close proximity of the zero-loss set. The framework by Katzenberger (1991) that we use provides a powerful tool to characterize such behaviour, and has been used extensively in the probability and other literature (see e.g. Funaki and Nagai (1993); Funaki (1995); Calzolari and Marchetti (1997); Parsons (2012); Parsons and Rogers (2015) and also Fatkullin et al. (2010)). Li, Wang, and Arora (2022a) used the same framework to characterize the behaviour of gradient descent with label noise in the limit of small learning rate. This same point of view has also been used to analyze 'local SGD' by Gu et al. (2023) and the impact of normalization by Li et al. (2022b).

In addition to the asymptotically continuous-time approaches above, the discrete-time nature of gradient descent and stochastic gradient descent generates an additional implicit bias, which is related to the form of the loss landscape close to zero loss (Arora et al., 2022; Wu et al., 2022).

Training and flatness of minima. There is growing experimental and theoretical evidence that 'flatter minima generalize better' (see e.g. Hochreiter and Schmidhuber (1997); Keskar et al. (2017); Jastrzębski et al. (2018); Chaudhari et al. (2019); Jiang et al. (2019)), and various properties of SGD and other training algorithms have been interpreted in this light (see e.g. Zhang et al. (2018); Jastrzębski et al. (2018); Wu et al. (2018); Smith et al. (2020); Pesme et al. (2021); Orvieto et al. (2022); Wu et al. (2022)). The results of this paper relate to this 'flatness' hypothesis, since in many cases the regularization term in e.g. (37) can be recognized as some measure of 'flatness'. The examples in the introduction illustrate this: for instance, the choice $\hat{L}(w,\eta) := L(w+\eta)$ in Figure 2 leads to the constrained gradient flow driven by $\Delta_w L(w)$, which is indeed a measure of the 'flatness' of the loss landscape around Γ, and the evolution moves towards the minimizer of this 'flatness', as indicated by the green circle. In the case of label noise, it was already shown by Li et al. (2022a) that the resulting regularizer also is proportional to $\Delta_w L(w)$, with the same effect. As yet another example, the regularizer that we derive for DropConnect in Section 5.1 is closely related to the concept of 'robustness' developed by Petzka et al. (2021), which those authors connect in turn to generalization performance (see Remark 37).

3. Notation and Preliminaries

After setting up the notation in Section 3.1 we introduce the notion of gradient descent algorithms with noise injection in Section 3.2. We discuss the assumptions on the zero-loss set and the behaviour of the system around it in Section 3.3. We give the necessary background and results on the convergence of the suitable type of stochastic processes (sometimes called Katzenberger processes after Katzenberger (1991)) in Section 3.4. Finally we introduce the characterization of the limit map following the derivation by Li et al. (2022a) in Section 3.5.

3.1 Notation

Convergence of random variables. We work with an abstract filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, on which all random variables are defined. For a sequence of $\mathbb{R}^m$-valued random variables $(X_n)_{n\in\mathbb{N}}$ we denote with $X_n \Rightarrow X$ the convergence in distribution to some random variable X, that is $\mathbb{E}f(X_n) \to \mathbb{E}\, f(X)$ for all $f \in C_b(\mathbb{R}^m)$.

We denote with $D_{\mathbb{R}^m}[0,\infty)$ the space of m-dimensional càdlàg processes equipped with the Skorokhod topology (see e.g. Billingsley (1968, Chapter 3)).

For two càdlàg paths $F, M \in D_{\mathbb{R}^m}[0,\infty)$, we denote with $\mathrm{V}_t(F)$ and $[M](t)$ the total and quadratic variation (see (22) for their definitions in the case of pure-jump paths).

Estimates and bounds. We use the floor operation by $\lfloor\cdot\rfloor : \mathbb{R} \to \mathbb{Z}$ to give the largest integer smaller than the argument, that is $\lfloor x\rfloor = \max\{z \in \mathbb{Z} : z \le x\}$. Indices in sums will be always integers and we use shorthand notation $\sum_{k\le x}$ for real positive x to denote the sum $\sum_{k=0}^{\lfloor x\rfloor}$.

The notation $a \lesssim b$ is understood as $a \le Cb$ for some constant C depending on stated assumptions, but never on the hyperparameters α, σ. We also use the Landau notation $f = O(g)$ and $f = o(g)$ to denote that $|f|/|g|$ is bounded or converges to zero, respectively, with the limit under consideration made clear in the context.

Vector and matrix operations. We denote with $|\cdot|$ the usual Euclidean norm on $\mathbb{R}^m$. For some $v, w \in \mathbb{R}^m$, the usual vector product is denoted with $v\cdot w := \sum_{i=1}^m v_i w_i$. In addition, the vector $v \odot w \in \mathbb{R}^m$ is defined by component-wise multiplication $(v\odot w)_i := v_i w_i$ for $i = 1,\dots,m$. Likewise, we denote with $v \otimes w \in \mathbb{R}^{m\times m}$ the standard tensor product of vectors defined by $(v\otimes w)_{i,j} = v_i w_j$ for $i,j = 1,\dots,m$. For $x_0 \in \mathbb{R}^m$ and $r \in \mathbb{R}^+$ we use $B(x_0, r)$ to denote a ball with center x_0 and radius r.

The product of two matrices $A, B \in \mathbb{R}^{m\times n}$ is given by $A : B := \sum_{i=1}^m \sum_{j=1}^n A_{ij}B_{ij}$. For a linear map $A : \mathbb{R}^n \to \mathbb{R}^m$ we write $A^\dagger : \mathbb{R}^m \to \mathbb{R}^n$ for its Moore-Penrose pseudoinverse. For a symmetric positive definite matrix $A \in \mathbb{R}^{m\times m}_{\mathrm{sym}}$, we denote with $|A|_+$ the product of the non-zero eigenvalues.

Derivatives. For a map $\Psi : \mathbb{R}^m \to \mathbb{R}^n$, the first and second variations in $x \in \mathbb{R}^m$ are denoted by $\partial\Psi(x)$ and $\partial^2\Psi(x)$. The action of these linear maps on directions $v \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m\times m}$ is denoted by

$$\partial\Psi(x)[v] := \sum_{i=1}^m \partial_{x_i}\Psi(\xi)v_i \qquad \text{and} \qquad \partial^2\Psi(x)[\Sigma] := \sum_{i=1}^m\sum_{j=1}^m \partial^2_{x_i,x_j}\Psi(x)\Sigma_{ij}\,.$$

By contrast, for a scalar map $\Psi : \mathbb{R}^m \to \mathbb{R}$, the notations $\nabla\Psi$ and $\nabla^2\Psi$ indicate the vector and matrix of first and second partial derivatives. This distinction in notation allows us to write e.g. $\partial_\eta \nabla_w L(w,\eta)[v]$ to indicate that the direction v is to be contracted against the derivatives in η. By slight abuse of notation, if $\Sigma = \sigma \otimes \sigma$ for some $\sigma \in \mathbb{R}^m$, then we also write $\partial^2\Psi(x)[\sigma,\sigma]$ instead of $\partial^2\Psi(x)[\sigma\otimes\sigma]$.

3.2 Problem Setting

In this work we study the following class of noisy gradient descent systems.

Definition 6 (Noisy gradient descent) *Given a* loss function $L \in C^3(\mathbb{R}^m, [0,\infty))$*, any function* $\hat{L} \in C^3(\mathbb{R}^{m+d}, [0,\infty))$ *with* $\hat{L}(w,0) = L(w)$ *is called a* noise-injected loss function*. A* noisy gradient descent *for the noise-injected loss* $\hat{L}$ *is the dynamics given by*

$$w_{k+1} = w_k - \alpha \nabla_w \hat{L}(w_k, \eta_k), \qquad \text{for } k \in \mathbb{N}_0; \tag{8a}$$

$$\eta_{k,i} \sim \rho(\sigma), \ \text{i.i.d.}, \qquad \text{for } i = 1,\dots,d\,. \tag{8b}$$

Here $w_0 \in \mathbb{R}^m$ *is a given initialization and* $\alpha > 0$ *is the learning rate. The family of probability distributions* $\{\rho(\sigma) \in \mathcal{P}(\mathbb{R}^m)\}_{\sigma>0}$ *characterizes the noise injection and* $\eta_{k,i} \sim \rho(\sigma)$ *for* $i = 1,\dots,d$ *are i.i.d. random variables distributed according to* $\rho(\sigma)$ *such that* η_k *is an* m*-dimensional vector. The family* $\{\rho(\sigma)\}_{\sigma>0}$ *is assumed to be centered with variance* σ^2*, that is*

$$\mathbb{E}_{\eta\sim\rho(\sigma)}\,\eta = 0 \qquad \text{and} \qquad \mathrm{Var}_{\eta\sim\rho(\sigma)}\,\eta = \sigma^2. \tag{9}$$

We require the noisy loss $\hat{L}$ together with the distribution ρ to satisfy the following compatibility and growth assumptions.

Assumption 7 (Compatibility between $\hat{L}$ and $\rho(\sigma)$) *There exists* $p \in [1,\infty)$ *such that*

1. *for any compact* $K \subset \mathbb{R}^m$ *there exists* $C_1 = C_1(K)$ *and* $C_2 = C_2(K)$ *with*

$$\left|\nabla_w \nabla_\eta^2 \hat{L}(w,\eta) - \nabla_w \nabla_\eta^2 \hat{L}(\tilde{w},\eta)\right| \le C_1(1+|\eta|^p)|w-\tilde{w}|\,, \ \text{for all } w,\tilde{w} \in K \text{ and all } \eta \in \mathbb{R}^d, \tag{10}$$

$$\left|\nabla_w \nabla_\eta^2 \hat{L}(w,\eta) - \nabla_w \nabla_\eta^2 \hat{L}(w,0)\right| \le C_2|\eta|(1+|\eta|^{p-1})\,, \qquad \text{for all } w \in K \text{ and all } \eta \in \mathbb{R}^d. \tag{11}$$

2. *Setting*

$$\mathrm{M}_k(\sigma) := \mathbb{E}_{\eta\sim\rho(\sigma)}|\eta|^k,$$

we have for all $k \in \{3,\dots,2(p+2)\}$*,*

$$\text{if } C_2 > 0\text{:} \quad \mathrm{M}_k(\sigma) = o(\sigma^2)\,, \tag{12a}$$

$$\text{if } C_2 = 0\text{:} \quad \mathrm{M}_k(\sigma) = O(\sigma^2)\,. \tag{12b}$$

If σ *is clear from the context, we briefly write* $\mathrm{M}_k := \mathrm{M}_k(\sigma)$*.*

Remark 8 (Examples of noise distributions) *The conditions* (9) *and bounds* (12) *for any* $p \ge 1$ *are satisfied for the centered Gaussian distribution* $\mathcal{N}(0,\sigma^2)$ *and the uniform distribution* $\mathrm{Unif}(-\sqrt{3}\sigma, \sqrt{3}\sigma)$*. Note that in both cases we have* $\mathrm{M}_p = O(\sigma^p)$ *and thus for* $q \ge 2(p+2) \ge 6$*, for* $p \ge 1$*, we have* $\mathrm{M}_q = O(\sigma^6)$*.*

Remark 9 (More general noise properties) *In the proofs we actually do not require the noisy variables $\eta^n_{k,i}$ to be sampled from the same distribution. It is sufficient that the $\eta_{k,i}$ are independent and that for every i the conditions* (9) *and* (12) *are satisfied. We stick to the assumption of $\eta_{k,i}$ being i.i.d. for the purpose of readability.*

A noisy gradient system as in Definition 6 is characterized by two hyperparameters, the learning rate α and the noise variance σ. For the main results of this paper we give ourselves two positive sequences $(\alpha_n)_{n\in\mathbb{N}}$ and $(\sigma_n)_{n\in\mathbb{N}}$ and study the limit of interpolations $(\tilde{w}_n)_{n\in\mathbb{N}}$, where $\tilde{w}_n : \mathbb{R}_+ \to \mathbb{R}^m$ is the m-dimensional càdlàg process defined by

$$\tilde{w}_n(t) = w^n_{\lfloor \frac{t}{\alpha_n} \rfloor} \qquad \text{for every } t \in \mathbb{R}_+ . \tag{13a}$$

Here $\lfloor x \rfloor$ is the integer part of x, and $\{w^n_k\}_{k\in\mathbb{N}_0}$ is the noisy gradient descent for hyperparameters $(\alpha_n)_{n\in\mathbb{N}}$ and $(\sigma_n)_{n\in\mathbb{N}}$, given by

$$w^n_{k+1} = w^n_k - \alpha_n \nabla_w \hat{L}(w^n_k, \eta^n_k), \qquad \text{for } k \in \mathbb{N}_0; \tag{13b}$$

$$\eta^n_{k,i} \sim \rho(\sigma_n), \qquad \text{for } i = 1, \dots, d . \tag{13c}$$

Depending on the structure of $\hat{L}$ we consider two scaling regimes, where either both $\alpha_n, \sigma_n \to 0$ or only $\alpha_n \to 0$ for constant noise variance $\sigma_n = \sigma \geq 0$. Consider the space $D_{\mathbb{R}^m}[0,\infty)$ of all m-dimensional càdlàg processes equipped with the Skorokhod topology. We study the limiting behaviour of the interpolations $\tilde{w}_n(t)$ in $D_{\mathbb{R}^m}[0,\infty)$ for the two different cases and characterize the limit processes in terms of the noise-injected loss function $\hat{L}$.

3.3 Zero-loss set

The key assumption in our analysis is the structure of the zero-loss set $\Gamma = \{w \in \mathbb{R}^m : L(w) = 0\}$. We consider systems in which the zero-loss set is locally a C^2 manifold satisfying some non-degeneracy assumptions.

For the gradient flow of L,

$$\dot{x}(t) = -\nabla L(x(t)) \quad \text{with} \quad x(0) = x_0 \in \mathbb{R}^m , \tag{14}$$

we define the flow map $\phi : \mathbb{R}^m \times [0,\infty) \to \mathbb{R}^m$ to be the corresponding solution map, i.e. $\phi(x_0, t)$ is the solution $x(t)$ of (14) at time t, and we have the integral representation

$$\phi(x,t) = x - \int_0^t \nabla L(\phi(x,s))\, \mathrm{d}s . \tag{15}$$

For any initial point $x_0 \in \mathbb{R}^m$ we define the ω-limit set by

$$\omega(x_0) = \bigcap_{s>0} \overline{\{\phi(x_0,t) : t > s\}}.$$

To establish local attractiveness of the loss manifold, we require the loss function $L : \mathbb{R}^m \to [0,\infty)$ to satisfy a number of non-degeneracy assumptions.

Assumption 10 (Non-degeneracy of the loss manifold)
The set $\Gamma := \{x \in \mathbb{R}^m : L(x) = 0\}$ *is a* non-degenerate loss manifold *in the following sense:*

1. manifold: *it forms an M-dimensional C^2 manifold;*
2. constant rank: *the rank of $\nabla^2 L$ is constant and maximal on Γ, that is* $\operatorname{rank}(\nabla^2 L(x)) = m - M$ *for all $x \in \Gamma$;*
3. spectral gap: *there exists $\delta > 0$ such that for all $x \in \Gamma$, all non-zero eigenvalues λ of $\nabla^2 L(x)$ satisfy $\lambda > \delta$.*

By the manifold condition in Assumption 10, there exists a tangent space $T_{x_0}\Gamma$ for each $x_0 \in \Gamma$. Moreover, for any $x_0 \in \Gamma$, there exists an $\epsilon > 0$ such that the projection operator $\operatorname{Proj}_\Gamma : B_\epsilon(x_0) \to \Gamma$ is well-defined. Then, the rank and spectral gap condition in Assumption 10 ensure that locally for every $x_0 \in \Gamma$ and every non-tangent direction $v \notin T_{x_0}\Gamma$ the loss function locally grows quadratically, that is there exists some $c > 0$ such that

$$L(x_0 + \epsilon v) \geq c\epsilon^2 \left|(I - P_{T_{x_0}\Gamma})v\right|^2 .$$

The assumption is satisfied by many existing overparameterized systems. In particular, overparameterized linear models (Li et al., 2022a) and feedforward neural networks (Cooper, 2018) have been shown to satisfy conditions similar to Assumption 10.

Remark 11 (Localizing the smoothness requirement) *For many losses, for instance those based on* ReLU *neural networks, the function L and the zero-loss set Γ are not differentiable, and Assumption 10 is not satisfied. For those systems, the results of this paper can be adapted by localizing. Since the main theorems characterize training behaviour up to the time of leaving a compact set K, the results can be applied within a set K such that $\Gamma \cap K$ and $L|_K$ do satisfy Assumption 10.*

We study the behaviour of the noisy gradient descent (13) in the proximity of Γ and thus require initial conditions that guarantee the convergence of the flow map ϕ in (15) to Γ. In terms of the ω-limit, we define a *locally attractive* neighbourhood U of the zero-loss manifold Γ such that the deterministic gradient flow trajectory (14) with initial conditions within U converges to a point on Γ.

Definition 12 *An open set $U \in \mathbb{R}^m$ is a* locally attractive *neighbourhood of Γ if $\Gamma \subset U$ and there exists a map $\Phi \in C^2(U;\Gamma)$ which satisfies $\omega(x) = \{\Phi(x)\}$ for all $x \in U$. In this case, Φ is called the* limit map *and satisfies*

$$\Phi(x) = \lim_{t\to\infty} \phi(x,t) \quad \text{for all } x \in U. \tag{16}$$

The existence of a locally attractive neighbourhood is guaranteed by Katzenberger (1991, Proposition 3.5) whenever Γ satisfies Assumption 10. In addition, we use the regularity of the limit map for which a direct application from Falconer (1983, Theorem 5.1) guarantees that if L is three times differentiable with locally Lipschitz third derivative, the limit map satisfies $\Phi \in C^2(U)$ thanks to Katzenberger (1991, Corollary 5.1).

Remark 13 *Note that for initial conditions $x_0 \in U$ the (unperturbed) gradient flow* (14) *converges to $\Phi(x_0)$. As remarked in the introduction, the noise injection drastically changes this behaviour, since instead of converging to a stationary point the system then approximately follows a manifold-constrained deterministic or stochastic flow.*

Proposition 14 (Exponential convergence of the flow) *Let U be a locally attractive neighbourhood (Definition 12) of the loss manifold Γ satisfying Assumption 10, then there exists $\beta > 0$ such that for any $W(0) \in U$ there exists $C > 0$ with*

$$|\phi(W(0), t) - \Phi(W(0))| \leq Ce^{-\beta t}.$$

In particular, the constants $C > 0$ can be chosen uniformly among $W(0) \in K \cap U$ for any compact set K.

Proof The proof is not literally written out by Katzenberger (1991), however it is implied by Katzenberger (1991, Lemma 3.2) together with a standard localization procedure as used in the proof of Katzenberger (1991, Lemma 3.3). ■

3.4 Katzenberger's Theorem

In this section we present a simplified version of the main tool of our analysis, which is Katzenberger (1991, Theorem 6.3). We fix a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ and consider formal 'stochastic differential equations' of the form

$$\mathrm{d}X_n = -\nabla L(X_n)\,\mathrm{d}A_n + \mathrm{d}Z_n, \tag{17}$$

or in integrated form

$$X_n(t) = X_n(0) - \int_0^t \nabla L(X_n(s))\,\mathrm{d}A_n(s) + Z_n(t),$$

where $(A_n)_{n\in\mathbb{N}}$ is a sequence of deterministic non-decreasing càdlàg processes, also called an *integrator sequence*, and $(Z_n)_{n\in\mathbb{N}}$ is a sequence of $\mathbb{R}^m$-valued semimartingales with respect to $\mathbb{P}$. We assume that $Z_n(0) = 0$.

For the length of this section U_Γ will indicate an arbitrary locally attractive neighbourhood of the zero-loss set Γ (as in Definition 12).

Remark 15 (Generality of (17)) *The setup by Katzenberger (1991) is more general than the equation (17); in particular, it allows for a splitting of the process Z_n into a 'noise' and a 'mobility' part, with different assumptions on these two parts. For our purposes the simpler form (17) suffices.*

We impose the following conditions on the integrator sequence A_n.

Assumption 16 (Assumptions on $(A_n)_{n\in\mathbb{N}}$) *The family $(A_n)_{n\in\mathbb{N}}$ of non-decreasing càdlàg deterministic processes in (17) satisfies*

1. *$A_n(0) = 0$;*
2. *A_n asymptotically puts infinite mass onto every interval, that is, for every $\delta > 0$,*

$$\inf_{0\leq t\leq T}\big(A_n(t+\delta) - A_n(t)\big) \longrightarrow \infty, \qquad \text{as } n \to \infty\,; \tag{18}$$

3. A_n is asymptotically continuous, that is,

$$\sup_{t\geq 0}(A_n(t)-A_n(t-))\longrightarrow 0, \qquad \text{as } n\to\infty, \tag{19}$$

with $A_n(t-) := \lim_{s\nearrow t} A_n(s)$.

For any compact K we define the stopping time

$$\lambda_n(K) = \inf\{t\geq 0 \mid X_n(t)\notin K^\circ\} \tag{20}$$

and for a process $Y:[0,\infty)\to\mathbb{R}^d$ and a stopping time λ the notation Y^λ denotes the stopped process

$$Y^\lambda(t) := Y(t\wedge\lambda) \qquad \text{for all } t\in[0,\infty). \tag{21}$$

The family $(Z_n)_{n\in\mathbb{N}}$ of $\mathbb{R}^m$-valued semimartingales in (17) needs to satisfy the following two assumptions.

Assumption 17 (Vanishing increments of $(Z_n)_{n\in\mathbb{N}}$) *For all $T\in\mathbb{R}_+$ and all compact $K\subset U_\Gamma$ we have*

$$\sup_{0<s\leq T\wedge\lambda_n(K)}|\Delta Z_n(s)|\Rightarrow 0, \qquad \text{as } n\to\infty,$$

where $\Delta Z_n(s) := Z_n(s)-Z_n(s-)$ denotes the increment.

In the following we use the notation $\mathrm{V}_t(F)$ and $[M](t)$ to denote the total and quadratic variation of càdlàg paths $F, M\in D_{\mathbb{R}^d}[0,\infty)$. For pure-jump paths, which is the only case we will be using, these are defined by

$$\mathrm{V}_t(F) := \sum_{0<s<t}|\Delta F(s)| \qquad \text{and} \qquad [M](t) := \sum_{0<s<t}|\Delta M(s)|^2. \tag{22}$$

Assumption 18 *The family $(Z_n)_{n\in\mathbb{N}}$ is a sequence of semimartingales with sample paths in $D_{\mathbb{R}^d}[0,\infty)$. For every $n\geq 1$ there exist stopping times $\left\{\tau_n^k \mid k\geq 1\right\}$ and a decomposition of Z_n into a local martingale M_n plus a finite variation process F_n such that*

$$\mathbb{P}[\tau_n^k\leq k]\leq 1/k \qquad \text{and} \qquad \left([M_n](t\wedge\tau_n^k)+\mathrm{V}_{t\wedge\tau_n^k}(F_n)\right)_{n\in\mathbb{N}}$$

is uniformly integrable for every $t\geq 0, k\geq 1$ and

$$\lim_{\gamma\to 0}\limsup_{n\to\infty}\mathbb{P}\Big[\sup_{0\leq t\leq T}\left(\mathrm{V}_{t+\gamma}(F_n)-\mathrm{V}_t(F_n)\right)>\varepsilon\Big]=0, \tag{23}$$

for every $\varepsilon>0$ and $T>0$.

Remark 19 *Assumption 18 requires both the quadratic variation of the martingale part of the noise and the total variation of the drift to be bounded. One can interpret this as a requirement to accumulate an order one perturbation for any fixed time $t>0$ in the limit $n\to\infty$. In our analysis we encounter cases when one of the components dominates, namely the drift in the general case and the martingale part in the degenerate case. Nevertheless, in the general case Katzenberger's theorem allows for a balanced contribution of both terms.*

Before introducing Katzenberger's result we need to do a shift of the variable X_n. If we consider a solution X_n of the system (17) with $X_n(0) = x_0 \in U_\Gamma$ but $x_0 \notin \Gamma$, then for any $t > 0$ we have $A_n(t) \to \infty$ and thus by definition of U_Γ we have

$$\phi(x_0, A_n(t)) \longrightarrow \Phi(x_0) \in \Gamma \qquad \text{as } n \to \infty\,.$$

As we can take arbitrarily small t, by a simple Gronwall inequality argument, we obtain that the limiting process X for $n \to \infty$ must be discontinuous at $t = 0$. To avoid this discontinuity we consider the shifted process

$$Y_n(t) = X_n(t) - \phi\left(X_n(0), A_n(t)\right) + \Phi\left(X_n(0)\right). \tag{24}$$

Note that at $t = 0$ we have $\phi\left(X_n(0), A_n(0)\right) = X_n(0)$ and thus $Y_n(0) = \Phi\left(X_n(0)\right) \in \Gamma$. The main theorem by Katzenberger (1991) states that the limiting dynamics of $Y_n(t)$ lie on Γ and can be expressed in terms of the limit map Φ and the limit Z of the process Z_n.

Theorem 20 (Katzenberger (1991, Theorem 6.3)) *Assume that the loss manifold Γ satisfies Assumption 10. Assume that $L \in C^3(U_\Gamma)$, for a locally attractive neighbourhood U_Γ of Γ. Assume that the processes $(X_n)_{n\in\mathbb{N}}$ (17) satisfy Assumptions 16, 17, and 18 with $X_n(0) \Rightarrow X(0) \in U_\Gamma$. For a compact $K \subset U_\Gamma$, let*

$$\mu_n(K) := \inf\{t \geq 0 \mid Y_n(t) \notin K^\circ\}\,. \tag{25}$$

Then, for every compact $K \subset U_\Gamma$, the sequence $\left(Y_n^{\mu_n(K)}, Z_n^{\mu_n(K)}, \mu_n(K)\right)_{n\in\mathbb{N}}$ of stopped processes and stopping times is relatively compact in $D_{\mathbb{R}^d\times\mathbb{R}^m}[0,\infty) \times [0,\infty]$. If (Y, Z, μ) is a limit point of this sequence, then (Y, Z) is a continuous semimartingale, $Y(t) \in \Gamma$ for a.e. $t \in [0,\infty)$, $\mu \geq \inf\{t \geq 0 \mid Y(t) \notin K^\circ\}$ a.s., and

$$Y(t) = Y(0) + \int_0^{t\wedge\mu} \partial\Phi(Y)\,\mathrm{d}Z + \frac{1}{2}\sum_{ij}\int_0^{t\wedge\mu} \partial^2_{ij}\Phi(Y)\,\mathrm{d}[Z^i, Z^j]. \tag{26}$$

Remark 21 *For a given process G_t on the locally attractive neighbourhood U_Γ (Definition 12), the process $\Phi(G_t)$ satisfies $\Phi(G_t) \in \Gamma$ by definition. In addition, we note that Φ satisfies $\Phi(x) = x$ for all $x \in \Gamma$ so if $G_t \in \Gamma$ almost surely, we also have $G_t = \Phi(G_t)$ almost surely. Thus, if a sequence of stochastic processes converges to a process on the zero-loss manifold, the limiting process must coincide with its image under map Φ. Thus, Theorem 20 can be interpreted as an application of Itô's lemma to the limit map Φ.*

3.5 Characterization of the limit map

The characterization of the limit behaviour in Theorem 20 makes use of the first and second derivatives of the limit map Φ that was defined in Definition 12. In this section we recall the relevant characterizations of these derivatives that were proved by Li et al. (2022a).

For a linear map A we write $A^\dagger$ for its Moore-Penrose pseudoinverse. For $H \in \mathbb{R}^{k\times k}$, we define the *Lyapunov* operator $\mathcal{L}_H : \mathbb{R}^{k\times k} \to \mathbb{R}^{k\times k}$ by

$$\mathcal{L}_H(X) := H^\top X + XH.$$

For a non-negative symmetric matrix A, we define $|A|_+$ to be its 'pseudo-determinant', the product of its non-zero eigenvalues. With this preliminary considerations, we can refer to the first and second derivatives of Φ contained in Li et al. (2022a, Lem. 4.5 and Cor. 5.1 & 5.2).

Lemma 22 (First and second derivatives of Φ) *Let $L \in C^3(\mathbb{R}^m, [0,\infty))$ and assume that Γ is a lower-dimensional manifold in $\mathbb{R}^m$ of class C^1 satisfying Assumption 10.*

1. *For any $\xi_0 \in \Gamma$ the limit below exists, and the identities hold:*

$$\nabla\Phi(\xi_0) = \lim_{t\to\infty} e^{-t\nabla^2 L(\xi_0)} = P_{T_{\xi_0}\Gamma}. \tag{27}$$

Here $P_{T_{\xi_0}\Gamma}$ is the orthogonal projection onto $T_{\xi_0}\Gamma$. We write $P = P_{T_{\xi_0}\Gamma}$ for short, and $Q := Q_{T_{\xi_0}\Gamma} := I - P_{T_{\xi_0}\Gamma}$ for the corresponding orthogonal projection onto $T_{\xi_0}\Gamma^\perp$.

2. *The second derivative $\partial^2\Phi$ is characterized by*

$$\begin{aligned}\partial^2\Phi(\xi_0)[\Sigma] = (\nabla^2 L)^\dagger \partial^2(\nabla L)[P\Sigma P] - P\partial^2(\nabla L)[\mathcal{L}^\dagger_{\nabla^2 L}(Q\Sigma Q)] \\ + 2P\partial^2(\nabla L)[(\nabla^2 L)^\dagger Q\Sigma P],\end{aligned} \tag{28}$$

   for any symmetric $\Sigma \in \mathbb{R}^{m\times m}$.

3. *For the special case of the identity matrix, $\Sigma = I_m$, we have*

$$\partial^2\Phi(\xi_0)[I_m] = (\nabla^2 L)^\dagger \partial^2(\nabla L)[P] - P\nabla \log|\nabla^2 L|_+ . \tag{29}$$

4. *For the special case $\Sigma = \nabla^2 L(\xi_0)$ we have*

$$\partial^2\Phi(\xi_0)[\nabla^2 L] = -\frac{1}{2} P\nabla\Delta L(\xi_0). \tag{30}$$

4. Main Results

The main theorems of this paper are Theorems 23 and 29 below, which were already mentioned in the introduction as Theorems A and B. Both characterize convergence of time-rescaled noisy gradient systems to a limiting evolution that is a constrained ODE or SDE on Γ.

The starting point for both theorems is the dynamics of the parameters w^n from (1) or equivalently (13b), which we rewrite as

$$w^n_{k+1} = w^n_k - \alpha_n \nabla_w L(w^n_k) + \alpha_n\big(\nabla_w \hat{L}(w^n_k, 0) - \nabla_w \hat{L}(w^n_k, \eta^n_k)\big)\,, \tag{31}$$

where we used that $\hat{L}(w,0) = L(w)$ by Definition 6. As described in the Introduction, in order to follow the evolution on Γ we need to speed up the process, by a factor $1/\alpha_n\sigma_n^2$ in the non-degenerate case or $1/\alpha_n^2\sigma_n^2$ in the degenerate case.

4.1 Non-degenerate case

In the non-degenerate case the rescaling by $1/\alpha_n\sigma_n^2$ leads us to consider the sequence of processes

$$W_n(t) = \tilde{w}_n\left(\frac{t}{\sigma_n^2}\right) = w^n_{\lfloor \frac{t}{\alpha_n\sigma_n^2}\rfloor}, \tag{32}$$

where $\tilde{w}$ and w^n are the solution to (13). Here $\alpha_n, \sigma_n \to 0$ are chosen in some way to be specified. Out of the sequences $(\alpha_n)_{n\in\mathbb{N}}$, $(\sigma_n)_{n\in\mathbb{N}}$ we define the sequence of integrators

$$A_n(t) := \alpha_n \lfloor \frac{t}{\alpha_n\sigma_n^2} \rfloor. \tag{33}$$

For any n the dynamics of W_n can then be written in the form (17) as

$$\mathrm{d}W_n(t) = -\nabla L(W_n(t))\,\mathrm{d}A_n(t) + \mathrm{d}Z_n(t), \tag{34a}$$

$$Z_n(t) = \sum_{s\le t} \Delta Z_n(s) = \alpha_n \sum_{k\le \frac{t}{\alpha_n\sigma_n^2}} (\nabla_w \hat{L}(W_n(\alpha_n\sigma_n^2 k), 0) - \nabla_w \hat{L}(W_n(\alpha_n\sigma_n^2 k), \eta_k^n)), \tag{34b}$$

$$\eta^n_{k,i} \sim \rho(\sigma_n). \tag{34c}$$

Theorem 23 (Main convergence theorem in the non-degenerate case) *Consider a loss function L and noise-injected loss $\hat{L}$ in the sense of Definition 6 with loss manifold Γ satisfying Assumption 10. Let $W_n(0) \Rightarrow W_0 \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ in the sense of Definition 12. Let ρ satisfy the assumption (9) and let $\alpha_n, \sigma_n \to 0$ as $n \to \infty$. Let $\hat{L}$ and ρ satisfy the compatibility Assumption 7 for some $p \ge 1$ and assume*

$$\sup_{k\le 1/\alpha_n\sigma_n^2} \alpha_n |\eta_k^n|^{p+2} \Rightarrow 0\,, \qquad \text{as } n \to \infty. \tag{35}$$

Let W_n be a solution to (34) and the shifted process Y_n be defined as in (24) by

$$Y_n(t) := W_n(t) - \phi(W_n(0), A_n(t)) + \Phi\left(W_n(0)\right). \tag{36}$$

For compact $K \subset U_\Gamma$, define the exit time of K by

$$\mu_n(K) := \inf\{t \ge 0 \mid Y_n(t-) \notin K^\circ \text{ or } Y_n(t) \notin K^\circ\}\,.$$

Then for any compact set $K \subset U_\Gamma$, the sequence $\left(Y_n^{\mu_n(K)}, \mu_n\right)_{n\in\mathbb{N}}$ is relatively compact in the Skorokhod topology. Moreover, for any limit point (Y, μ) of $\left(Y_n^{\mu_n(K)}, \mu_n\right)$, Y is a continuous function of time, it satisfies $Y(t) \in \Gamma$ a.s. for any t, and

$$Y(t) = \Phi(W_0) - \int_0^{t\wedge\mu} P_{T_{Y(s)}\Gamma} \nabla_w \mathrm{Reg}(Y(s))\,\mathrm{d}s, \qquad \mathrm{Reg}(w) := \frac{1}{2}\Delta_\eta \hat{L}(w, 0)\,, \tag{37}$$

where $P_{T_\xi\Gamma}$ is the orthogonal projection onto the tangent space of Γ at the point $\xi \in \Gamma$. In addition,

$$\mu \ge \inf\{t \ge 0 \mid Y(t) \notin K^\circ\}. \tag{38}$$

The limiting equation (37) is a deterministic evolution equation for Y, and as long as $t < \mu$ it can be written in differential form as

$$\frac{d}{dt}Y(t) = -P_{T_{Y(t)}\Gamma}\nabla_w \mathrm{Reg}(Y(t)), \qquad \text{with } Y(t) \in \Gamma \text{ for } t \geq 0, \quad \text{and } Y(0) = \Phi(W_0). \tag{39}$$

It is a constrained gradient flow of the functional Reg, under the constraint that $Y \in \Gamma$; the role of the projection $P_{T_Y\Gamma}$ is to project the vector field $-\nabla\mathrm{Reg}$ onto the tangent space of Γ at Y, which is necessary to maintain $Y \in \Gamma$.

Remark 24 (Convergence of the whole sequence) *If the functional* Reg *is Lipschitz continuous (e.g. if $\hat{L} \in C^4$) then solutions of the constrained gradient flow* (37) *or* (39) *are unique up to time μ. This implies that defining*

$$\tau := \inf\{t \geq 0 \mid Y(t) \notin K^\circ\},$$

we have the Skorokhod convergence of the full sequence up to time τ,

$$Y_n^{\mu_n(K)\wedge\tau} \Rightarrow Y^\tau.$$

In addition, the exponential estimate (50) *below then implies that for any $\delta > 0$, we get the convergence $W_n^{\mu_n(K)\wedge\tau}|_{[\delta,\infty)} \Rightarrow Y^\tau|_{[\delta,\infty)}$ in Skorokhod topology.*

Note that we can not exclude the existence of different limit points (Y, μ). Imagine, for instance, that ∂K contains part of a solution curve of the gradient flow (39)*; then one can easily understand how for some n the hitting time $\mu_n(K)$ may be triggered earlier than for others. The uniqueness argument above, however, shows that as long as $Y(t) \in K^\circ$, i.e. $t < \tau$, all limit points Y coincide.*

Remark 25 (Skorokhod and locally-uniform convergence) *Convergence to a continuous process in Skorokhod topology implies uniform convergence on compact time intervals. Moreover, convergence in distribution to a deterministic object implies convergence in probability, so for any $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{0\leq s\leq t} \left|Y_n^{t\wedge\mu_n(K)}(s) - Y^{t\wedge\mu(K)}(s)\right| > \varepsilon\right] \to 0.$$

Remark 26 (Convergence of Z_n) *In the proof we also show that the sequence of noise processes Z_n converges to a deterministic process. This behaviour corresponds to the case in which the total variation term in Assumption 18 dominates the quadratic variation.*

Proof [Proof of Theorem 23] The proof consists of two parts: the first part is an application of Theorem 20, which leads to the characterization (26). In the second step we convert that equation to a more explicit form, by giving an explicit characterization of the limit process Z.

For both parts, we collect a number of properties of the process Z_n and set

$$\Delta\mathsf{Z}_n(w,\eta) := \alpha_n\big(\nabla_w\hat{L}(w,0) - \nabla_w\hat{L}(w,\eta)\big)\,; \tag{40a}$$

$$\Delta\mathsf{F}_n(w) := \mathbb{E}_{\eta\sim\rho(\sigma_n)}[\Delta\mathsf{Z}_n(w,\eta)]\,; \tag{40b}$$

$$\mathrm{Reg}(w) := \frac{1}{2}\Delta_\eta\hat{L}(w,0)\,. \tag{40c}$$

We note that from (40a), we get $\Delta Z_n(\alpha_n\sigma_n^2 k) = \Delta \mathsf{Z}_n(W_n(\alpha_n\sigma_n^2 k), \eta_k^n)$. Similarly to Z_n we assemble the jumps $\Delta \mathsf{F}_n$ into a process with jumps at times separated by $\alpha_n\sigma_n^2$,

$$F_n(t) := \sum_{s\le t} \Delta F_n(s) := \sum_{k\le \frac{t}{\alpha_n\sigma_n^2}} \Delta \mathsf{F}_n(W_n(\alpha_n\sigma_n^2 k)).$$

Lemma 27 (Properties of Z_n and F_n) *Let the compact set $K \subset \mathbb{R}^m$ and the sequences α_n, σ_n be as in Theorem 23. Let $T > 0$. Let $\hat{K} = K + \overline{B(0,C)}$, where C is the corresponding constant in Proposition 14. Then $W_n(t) \in \hat{K}$ for all $t \le T \wedge \mu_n(K)$ and*

$$\sup_{t\le T\wedge\mu_n(K)} |\Delta Z_n(t)| \Rightarrow 0 \qquad \text{as } n\to\infty\,; \tag{41}$$

$$\sup_{w\in\hat{K}} \Big|\Delta \mathsf{F}_n(w) + \alpha_n\sigma_n^2 \nabla \mathrm{Reg}(w)\Big| = o(\alpha_n\sigma_n^2) \qquad \text{as } n\to\infty\,; \tag{42}$$

$$\mathbb{E}_{\eta\sim\rho(\sigma_n)}\Big[\sup_{w\in\hat{K}} \big(\Delta \mathsf{Z}_n(w,\eta)\big)^2\Big] + \sup_{w\in\hat{K}} \big(\Delta \mathsf{F}_n(w)\big)^2 = o(\alpha_n\sigma_n^2) \qquad \text{as } n\to\infty\,. \tag{43}$$

We prove this lemma below, and continue in the meantime with the proof of Theorem 23. To apply Theorem 20 we verify Assumptions 16, 17, and 18. Note that we fix the set K for once and for all, and we show that the *stopped* processes $Z_n^{\mu_n(K)}$ satisfy Assumptions 17 and 18.

Part 1: Verification of the assumptions of Theorem 20.

Verification of Assumption 16. The condition (18) on the integrator sequence $A_n(t) = \alpha_n \lfloor \frac{t}{\alpha_n\sigma_n^2} \rfloor$ is verified for $\sigma_n \to 0$ and for any $\delta > 0$ by the estimate

$$\inf_{0<t\le T} \big(A_n(t+\delta) - A_n(t)\big) = \inf_{0<t\le T} \left(\alpha_n \left\lfloor \frac{t+\delta}{\alpha_n\sigma_n^2}\right\rfloor - \alpha_n \left\lfloor \frac{t}{\alpha_n\sigma_n^2}\right\rfloor\right) \ge \alpha_n \left\lfloor \frac{\delta}{\alpha_n\sigma_n^2}\right\rfloor \to \infty\,.$$

Likewise, we verify the assumption (19) by using at the same time $\alpha_n \to 0$ to deduce

$$\sup_{t\ge 0} \big(A_n(t) - A_n(t-)\big) = \alpha_n \to 0\,.$$

Verification of Assumption 17. Assumption 17 for the stopped process $Z_n^{\mu_n(K)}$ follows immediately from (41).

Verification of Assumption 18. We use the martingale decomposition of $Z_n = M_n + F_n$, with martingale part

$$\begin{aligned} M_n(t) := Z_n(t) - F_n(t) &= \sum_{s\le t} \big(\Delta Z_n(s) - \Delta F_n(s)\big) \\ &= \sum_{k\le \frac{t}{\alpha_n\sigma_n^2}} \Big[\Delta \mathsf{Z}_n(W_n(\alpha_n\sigma_n^2 k, \eta_k^n)) - \Delta \mathsf{F}_n(W_n(\alpha_n\sigma_n^2 k))\Big]. \end{aligned}$$

The expected quadratic variation of $M_n^{\mu_n(K)}$ is estimated by

$$\begin{aligned}\mathbb{E}[M_n](t\wedge\mu_n(K)) &= \mathbb{E}[Z_n - F_n](t\wedge\mu_n(K)) \le 2\,\mathbb{E}\big[[Z_n]+[F_n]\big](t\wedge\mu_n(K)) \\ &= 2\ \mathbb{E}\sum_{k\le\frac{t\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\Big[\Delta \mathsf{Z}_n(W_n(\alpha_n\sigma_n^2 k,\eta_k^n))^2 + \Delta\mathsf{F}_n(W_n(\alpha_n\sigma_n^2 k))^2\Big] \\ &\xrightarrow{(43)} 0 \qquad \text{as } n\to\infty\,.\end{aligned} \tag{44}$$

This proves the uniform integrability of $[M_n](t\wedge\mu_n(K))$ in Assumption 18 (where we can take $\tau_n^k = +\infty$ for all k and n).

We next show condition (23). From (42) and the regularity of Reg we have

$$\sup_{w\in\hat K}\big|\Delta\mathsf{F}_n(w)\big| \le \sup_{w\in\hat K}\big|\Delta\mathsf{F}_n(w)+\alpha_n\sigma_n^2\nabla\mathrm{Reg}(w)\big| + \alpha_n\sigma_n^2\sup_{w\in\hat K}\big|\nabla\mathrm{Reg}(w)\big| \lesssim \alpha_n\sigma_n^2.$$

Therefore, since $W_n(t)\in\hat K$ as defined in Lemma 27,

$$V_{t+\gamma}(F_n^{\mu_n(K)}) - V_t(F_n^{\mu_n(K)}) = \sum_{\frac{t\wedge\mu_n(K)}{\alpha_n\sigma_n^2}<k\le\frac{(t+\gamma)\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\big|\Delta\mathsf{F}_n(W_n(\alpha_n\sigma_n^2 k))\big| \ \lesssim\ \gamma\,, \tag{45}$$

and the property (23) follows. Finally, the estimate (45) for $t=0$ implies that $V_t(F_n^{\mu_n(K)})$ is almost surely bounded uniformly in n, thus establishing the unform integrability condition on $V_t(F_n^{\mu_n(K)})$ in Assumption 18 (again with $\tau_n^k=+\infty$).

Having verified the three assumptions, we apply Theorem 20 and conclude that the triplet $(Y_n^{\mu_n(K)}, Z_n^{\mu_n(K)}, \mu_n(K))$ converges along a subsequence to a limit that satisfies the equation (26). We now show that that equation reduces to (37).

Part 2: Characterization of Z and identification of the limiting dynamics.

To recover the limiting dynamics we study the limit of the sequence of processes Z_n. Note that Theorem 20 establishes joint convergence of triplets $(Y_n^{\mu_n}, Z_n^{\mu_n}, \mu_n)$ but does not state any explicit result on the process W_n in (34a) and its limit. Even though Z_n is defined through W_n but not Y_n in (36), the initial jump does not play a role for the limiting behaviour of Z_n. Hence, we introduce an intermediate process defined in terms of Y_n by

$$Z_n^Y(t) := \sum_{k\le\frac{t}{\alpha_n\sigma_n^2}}\alpha_n\big(\nabla_w\hat L(Y_n(\alpha_n\sigma_n^2 k),0) - \nabla_w\hat L(Y_n(\alpha_n\sigma_n^2 k),\eta_k^n)\big)\,. \tag{46}$$

We then estimate

$$\begin{aligned}&\sup_{0<t\le T\wedge\mu_n(K)}\Big|Z_n(t)+\int_0^t\nabla\mathrm{Reg}(Y_n(s))\,\mathrm{d}s\Big| \\ &\quad\le \sup_{0<t\le T\wedge\mu_n(K)}\Big|Z_n(t)-Z_n^Y(t)\Big| + \sup_{0<t\le T\wedge\mu_n(K)}\Big|Z_n^Y(t)+\int_0^t\nabla\mathrm{Reg}(Y_n(s))\,\mathrm{d}s\Big|.\end{aligned} \tag{47}$$

We first show that the second term vanishes in probability. Note that Y_n is again a pure-jump process, so that

$$\int_0^t\nabla\mathrm{Reg}(Y_n(s))\,\mathrm{d}s = \alpha_n\sigma_n^2\sum_{k\le\frac{t}{\alpha_n\sigma_n^2}}\nabla\mathrm{Reg}(Y_n(\alpha_n\sigma_n^2 k))\,.$$

Writing the corresponding martingale

$$M_n^Y(t) = Z_n^Y(t) - F_n^Y(t) = \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} \left(\Delta \mathsf{Z}_n(Y_n(\alpha_n\sigma_n^2 k, \eta_k^n)) - \Delta \mathsf{F}_n(Y_n(\alpha_n\sigma_n^2 k)) \right),$$

we have by the same argument as in (44) that

$$\mathbb{E}[M_n^{Y,\mu_n(K)}](t) \longrightarrow 0 \qquad \text{as } n \to \infty,$$

and by Doob's inequality that for all t

$$\mathbb{E}\Big[\sup_{0\le s\le t} |M_n^{Y,\mu_n(K)}|^2(s) \Big] \le 4\, \mathbb{E}\big[[M_n^{Y,\mu_n(K)}](t)\big] \longrightarrow 0.$$

We then estimate

$$\begin{aligned}
&\mathbb{E} \sup_{0<t\le T\wedge\mu_n(K)} \left| Z_n^Y(t) + \int_0^t \nabla \mathrm{Reg}(Y_n(s))\,\mathrm{d}s \right| \\
&\quad \le \mathbb{E} \sup_{0<t\le T\wedge\mu_n(K)} |M_n(t)| + \mathbb{E} \sup_{0<t\le T\wedge\mu_n(K)} \left| F_n^Y(t) + \int_0^t \nabla \mathrm{Reg}(Y_n(s))\,\mathrm{d}s \right| \\
&\quad \le o(1) + \mathbb{E} \sum_{k \le \frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}} \left| \Delta \mathsf{F}_n(Y_n(\alpha_n\sigma_n^2 k)) + \alpha_n\sigma_n^2 \nabla \mathrm{Reg}(Y_n(\alpha_n\sigma_n^2 k)) \right|,
\end{aligned}$$

and the right-hand side vanishes by (42).

For the first term in the splitting (47), we get

$$\begin{aligned}
\left| Z_n(t) - Z_n^Y(t) \right| &= \alpha_n \Bigg| \sum_{k \le \frac{t}{\alpha_n\sigma_n^2}} \big(\nabla_w \hat{L}(W_n(\alpha_n\sigma_n^2 k), 0) - \nabla_w \hat{L}(W_n(\alpha_n\sigma_n^2 k), \eta_k^n) \big) \\
&\qquad\qquad - \big(\nabla_w \hat{L}(Y_n(\alpha_n\sigma_n^2 k), 0) - \nabla_w \hat{L}(Y_n(\alpha_n\sigma_n^2 k), \eta_k^n) \big) \Bigg| \\
&\le \alpha_n \Bigg| \sum_{k \le \frac{t}{\alpha_n\sigma_n^2}} \Big(\nabla_w \nabla_\eta \hat{L}(W_n(\alpha_n\sigma_n^2 k), 0) - \nabla_w \nabla_\eta \hat{L}(Y_n(\alpha_n\sigma_n^2 k), 0) \Big) \eta_k^n \Bigg| \qquad (48) \\
&\quad + \alpha_n \Bigg| \sum_{k \le \frac{t}{\alpha_n\sigma_n^2}} \sum_{i,j=1}^d g_{i,j}(W_k^n, Y_k^n, \eta_k^n) \eta_{k,i}^n \eta_{k,j}^n \Bigg|, \qquad (49)
\end{aligned}$$

where we use Taylor's theorem to write the remainder term as

$$g_{ij}(W_k^n, Y_k^n, \eta_k^n) := \int_0^1 (1-r) \partial_{\eta_i\eta_j} \big(\nabla_w \hat{L}(W_n(\alpha_n\sigma_n^2 k), r\eta_k^n) - \nabla_w \hat{L}(Y_n(\alpha_n\sigma_n^2 k), r\eta_k^n) \big)\,\mathrm{d}r\,.$$

Recall that the definition of Y_n in (36) implies

$$W_n(t) - Y_n(t) = \phi\left(W_n(0), A_n(t)\right) - \Phi\big(W_n(0)\big)\,.$$

Then for the first term (48) we have

$$\alpha_n \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} \Big(\nabla_w \nabla_\eta \hat{L}(W_n(\alpha_n \sigma_n^2 k), 0) - \nabla_w \nabla_\eta \hat{L}(Y_n(\alpha_n \sigma_n^2 k), 0) \Big) \eta_k^n$$
$$= \sqrt{\alpha_n} \int_0^t \Big(\nabla_w \nabla_\eta \hat{L}(W_n(s), 0) - \nabla_w \nabla_\eta \hat{L}(Y_n(s), 0) \Big) \, \mathrm{d}h_n.$$

Here h_n is defined as

$$h_n(t) := \sqrt{\alpha_n} \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} \eta_k^n$$

and h_n converges to a standard Brownian motion by the Functional Central Limit Theorem (Billingsley, 1968, Th. 8.2). Since $\nabla_w \nabla_\eta \hat{L}(w, 0)$ is bounded on $\hat{K}$, the characterization by Katzenberger (1991, Prop. 4.4) then implies that the term (48) converges to zero in probability.

For the second term (49) we use Assumption 7 and obtain

$$\begin{aligned} |g_{ij}(W_k^n, Y_k^n, \eta_k^n)| &= \left| \int_0^1 (1-r) \partial_{\eta_i \eta_j} \big(\nabla_w \hat{L}(W_n(\alpha_n \sigma_n^2 k), r\eta_k^n) - \nabla_w \hat{L}(Y_n(\alpha_n \sigma_n^2 k), r\eta_k^n) \big) \, \mathrm{d}r \right| \\ &\le \sup_{0 \le r \le 1} \left| \nabla_\eta^2 \big(\nabla_w \hat{L}(W_n(\alpha_n \sigma_n^2 k), r\eta_k^n) - \nabla_w \hat{L}(Y_n(\alpha_n \sigma_n^2 k), r\eta_k^n) \big) \right| \\ &\lesssim (1 + |\eta_k^n|^p) |W_n(\alpha_n \sigma_n^2 k) - Y_n(\alpha_n \sigma_n^2 k)|. \end{aligned}$$

By applying this bound to the residual in estimate (49), we obtain

$$\left| \alpha_n \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} \sum_{i,j=1}^d g_{ij}(W_k^n, Y_k^n, \eta_k^n) \eta_{k,i}^n \eta_{k,j}^n \right| \lesssim \alpha_n \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} (|\eta_k^n|^2 + |\eta_k^n|^{p+2}) |W_n(\alpha_n \sigma_n^2 k) - Y_n(\alpha_n \sigma_n^2 k)|.$$

By an application of Proposition 14, we have the exponential convergence

$$|W_n(t) - Y_n(t)| = \big| \phi\big(W_n(0), A_n(t)\big) - \Phi\big(W_n(0)\big) \big| \lesssim e^{-\beta A_n(t)} \,. \tag{50}$$

Hence, we conclude

$$\begin{aligned} \mathbb{E}\Big[\alpha_n \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} (|\eta_k^n|^2 + |\eta_k^n|^{p+2}) e^{-\beta A_n(t)} \Big] &= \alpha_n \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} e^{-\beta A_n(t)} \mathbb{E}\big[|\eta_k^n|^2 + |\eta_k^n|^{p+2} \big] \\ &\le (\sigma_n^2 + \mathrm{M}_{p+2}(\sigma_n)) \int_0^t e^{-\beta A_n(t)} \, \mathrm{d}A_n(t) \longrightarrow 0 \qquad \text{as } n \to \infty \,, \end{aligned}$$

because $M_{p+2}(\sigma_n) = O(\sigma_n^2)$ by Assumption 7 and $\sigma_n \to 0$. So we conclude

$$(49) \;\lesssim\; \alpha_n \sum_{k \le \frac{t}{\alpha_n \sigma_n^2}} (|\eta_k^n|^2 + |\eta_k^n|^{p+2}) |W_n(\alpha_n \sigma_n^2 k) - Y_n(\alpha_n \sigma_n^2 k)| \Rightarrow 0.$$

Hence, we have shown that

$$Z_n \Rightarrow Z, \qquad \text{where} \qquad Z(t) := -\int_0^t \nabla \mathrm{Reg}(Y(s)) \, \mathrm{d}s.$$

Note that the limit Z is a continuous process of finite variation and therefore $[Z,Z]=0$. Combining this with (27) we find that (26) reduces to (37). Since the right-hand side of (37) is continuous in t for any Y, the process Y is a continuous function of t. By the definition (36) we have $Y_n(0)=\Phi(W_n(0))\Rightarrow\Phi(W_0)$, implying that $Y(0)=\Phi(W_0)$.

This concludes the proof of Theorem 23. ∎

We still owe the reader the proof of Lemma 27.

Proof [Proof of Lemma 27] To begin with, note that the stopping time $\mu_n(K)$ restricts $Y_n(t\wedge\mu_n(K))$ to the compact set K, which combined with Proposition 14 guarantees that $W_n(t\wedge\mu_n(K))$ is restricted to the larger but still compact set $\hat{K}$. We first prove (41). We use Taylor expansion in η to obtain

$$\nabla_w\hat{L}(w,0)-\nabla_w\hat{L}(w,\eta)=-\nabla_w\partial_\eta\hat{L}(w,0)[\eta]-\frac{1}{2}\nabla_w\partial_\eta^2\hat{L}(w,0)[\eta,\eta]-R_n(w,\eta)\,, \tag{51}$$

where the error term R_n is given by

$$R_n(w,\eta):=\int_0^1(1-\xi)\Big(\partial_\eta^2\nabla_w\hat{L}(w,\xi\eta)[\eta,\eta]-\partial_\eta^2\nabla_w\hat{L}(w,0)[\eta,\eta]\Big)\,\mathrm{d}\xi\,. \tag{52}$$

Using the compactness of $\hat{K}$ and Assumption 7, we then bound

$$\begin{aligned}\sup_{0<t\le T\wedge\mu_n(K)}|\Delta Z_n(t)|\le&\sup_{k\le\frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\alpha_n|\nabla_w\partial_\eta\hat{L}(W_n(\alpha_n\sigma_n^2k),0)[\eta_k^n]|\\&+\sup_{k\le\frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\alpha_n\Big|\frac{1}{2}\nabla_w\partial_\eta^2\hat{L}(W_n(\alpha_n\sigma_n^2k),0)[\eta_k^n,\eta_k^n]\Big|\\&+\sup_{k\le\frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\alpha_n\big|R_n(W_n(\alpha_n\sigma_n^2k),\eta_k^n)\big|\\\lesssim&\sup_{k\le\frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\alpha_n|\eta_k^n|+\sup_{k\le\frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\alpha_n|\eta_k^n|^2+\sup_{k\le\frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}\alpha_n\big(|\eta_k^n|^3+|\eta_k^n|^{p+2}\big)\,.\end{aligned} \tag{53}$$

Thus, by the assumption (35), we conclude with the estimate

$$\sup_{0<t\le T\wedge\mu_n(K)}|\Delta Z_n(t)|\lesssim\alpha_n\Big\{1+\sup_{k\le\frac{T\wedge\mu_n(K)}{\alpha_n\sigma_n^2}}|\eta_k^n|^{p+2}\Big\}\Rightarrow 0.$$

We next prove (42). First note that for any $w\in\hat{K}$

$$\mathbb{E}_\eta|R_n(w,\eta)|\overset{(11)}{\lesssim}\mathbb{E}_\eta\,C_2\int_0^1(1-\xi)(|\eta|^3+|\eta|^{p+2})\,\mathrm{d}\xi=\frac{1}{2}C_2(\mathrm{M}_3+\mathrm{M}_{p+2})\overset{(12)}{=}o(\sigma_n^2)\,. \tag{54}$$

We then write, using the independence of the coordinates of η,

$$\begin{aligned}|\Delta\mathsf{F}_n(w)+\alpha_n\sigma_n^2\nabla\mathrm{Reg}(w)|&=\alpha_n\Big|\mathbb{E}_\eta[\nabla_w\hat{L}(w,0)-\nabla_w\hat{L}(w,\eta)]+\frac{1}{2}\sigma_n^2\nabla_w\Delta_\eta\hat{L}(w,0)\Big|\\&=\alpha_n\Big|\underbrace{-\frac{1}{2}\,\mathbb{E}_\eta\Big[\nabla_w\partial_\eta^2\hat{L}(w,0)[\eta,\eta]\Big]+\frac{1}{2}\sigma_n^2\nabla_w\Delta_\eta\hat{L}(w,0)}_{=0}-\mathbb{E}_\eta\,R_n(w,\eta)\Big|\\&=o(\alpha_n\sigma_n^2)\qquad\text{as } n\to\infty.\end{aligned}$$

Since the estimates are independent of $w \in \hat{K}$, this proves (42).

We finally show (43). Since $\nabla \mathrm{Reg}$ is continuous on $\hat{K}$, it is bounded, and therefore

$$\big|\Delta \mathsf{F}_n(w)\big|^2 \le 2\big|\Delta \mathsf{F}_n(w) + \alpha_n\sigma_n^2 \nabla \mathrm{Reg}(w)\big|^2 + 2\alpha_n^2\sigma_n^4 \big|\nabla \mathrm{Reg}(w)\big|^2 \overset{(42)}{=} o(\alpha_n\sigma_n^2)\,.$$

For $\Delta \mathsf{Z}_n$ we use a similar calculation as in (53) above and estimate

$$\mathbb{E} \sup_{w\in K} \big|\Delta Z_n(w,\eta)\big|^2 \lesssim \alpha_n^2\, \mathbb{E}\Big[|\eta|^2 + |\eta|^4 + |\eta|^{2p+4}\Big] = \alpha_n^2(\mathrm{M}_2 + \mathrm{M}_4 + \mathrm{M}_{2p+4}) = o(\alpha_n\sigma_n^2)\,.$$

This concludes the proof of Lemma 27. ∎

4.2 Degenerate case

As discussed in Section 1, for some forms of noisy gradient descent the limiting dynamics of W_n characterized by Theorem 23 is trivial. Examples of this are label noise, minibatching, and stochastic gradient Langevin descent; we discuss these in more detail in Section 5. In this section we present the second main convergence result, in the more strongly accelerated regime, which applies to functions $\hat{L}$ for which the evolution (39) is trivial.

Label noise does not only have a trivial limiting dynamics according to Theorem 23, but it happens to have an even more specific structure which allows to prove convergence under milder assumptions. Namely, label noise has a loss function that is polynomial in η, and this both significantly simplifies the calculations and does not require the variance of the noise to vanish. It is enough to consider the limit of infinitesimally small learning rate $\alpha_n \to 0$ with σ_n either constant or converging to some $\sigma_0 \ge 0$. The same structure is shared by a number of other degenerate functions $\hat{L}$ that we study in Section 5.

To characterize this structure we make a more specific assumption on the noise-injected loss function $\hat{L}(w,\eta)$ that augments Definition 6.

Definition 28 (Degenerate quadratic noise-injected loss) *For a given* loss function $L \in C^3(\mathbb{R}^m, [0,\infty))$*, a noise-injected loss function* $\hat{L} \in C^3(\mathbb{R}^{m+d}, [0,\infty))$ *in the sense of Definition 6 is called* degenerate quadratic *provided that*

$$\hat{L}(w,\eta) = L(w) + f(w)\cdot\eta + \tfrac{1}{2}H(w) : (\eta\otimes\eta) + g(\eta) \tag{55}$$

for some $f \in C^3(\mathbb{R}^m, \mathbb{R}^d)$ *and a field of quadratic forms* $H \in C^3(\mathbb{R}^m, \mathbb{R}^{d\times d}_{sym})$ *with zero diagonal entries* $H_{ii}(w) = 0$ *for all* $w \in \mathbb{R}^m$*, where* $H : (\eta\otimes\eta) = \sum_{i,j} H_{ij}\eta_i\eta_j$*.*

Moreover, $g \in C^3(\mathbb{R}^d, \mathbb{R})$ *with* $g(0) = 0$*.*

For any $\hat{L}$ of the form given in Definition 28 we obtain $\nabla_w \Delta_\eta \hat{L}(w,\eta) = 0$, indeed leading to trivial dynamics on the time scale $\alpha_n\sigma_n^2$. We define the corresponding sequence of integrators $\hat{A}_n(t) = \alpha_n \lfloor \frac{t}{\alpha_n^2\sigma_n^2} \rfloor$ (note the difference with (33) in the power of α_n in the denominator). The corresponding accelerated process of noisy gradient descent (13) are then of the form

$$\hat{W}_n(t) = w^n_{\lfloor \frac{t}{\alpha_n^2\sigma_n^2} \rfloor}\,. \tag{56a}$$

Due to the specific structure of the noise-injected loss from Definition 28, we can bring the process into the form (17) (compare also with (34) in the first case), and obtain a simplified expression for the noise process $\tilde{Z}_n$ given by

$$d\hat{W}_n = -\nabla L(\hat{W}_n)\,\mathrm{d}\hat{A}_n + \mathrm{d}\hat{Z}_n, \tag{56b}$$

$$\hat{Z}_n(t) = \sum_{s\le t} \Delta\hat{Z}_n(s) = \alpha_n \sum_{k\le \frac{t}{\alpha_n^2\sigma_n^2}} \left(\nabla_w f(\hat{W}_n(\alpha_n^2\sigma_n^2 k))\cdot \eta_k^n + \tfrac{1}{2}\nabla_w H(\hat{W}_n(\alpha_n^2\sigma_n^2 k))[\eta_k^n,\eta_k^n]\right), \tag{56c}$$

$$\eta_{k,i}^n \sim \rho(\sigma_n). \tag{56d}$$

We formulate the analogue of Theorem 23 for the rescaled process defined by (56).

Theorem 29 (Main convergence theorem in the degenerate case) *Assume that L, $\hat{L}$ are $\mathcal{C}^3$, $\hat{L}$ is of the form (55), and L satisfies Assumption 10. Let $\hat{W}_n(0) \Rightarrow \hat{W}(0) \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Fix a sequence $(\sigma_n)_{n\in\mathbb{N}} \subset (0,\infty)$ such that $\sigma_n \to \sigma_0 \ge 0$ as $n\to\infty$. Let $(\rho_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$ satisfy $\mathbb{E}_{\eta\sim\rho_n}[\eta] = 0$ and $\mathbb{E}_{\eta\sim\rho_n}|\eta|^2 = \sigma_n^2$, and let $\alpha_n \to 0$ such that*

$$\alpha_n \sup_{k\le 1/\alpha_n^2\sigma_n^2} |\eta_k^n|^2 \Rightarrow 0 \qquad \text{for i.i.d. } \eta_{k,i}^n \sim \rho_n \text{ as } n\to\infty\,. \tag{57}$$

Let $\hat{W}_n$ be a solution of (56) and define

$$\hat{Y}_n(t) = \hat{W}_n(t) - \phi(\hat{W}_n(0), \hat{A}_n(t)) + \Phi(\hat{W}_n(0))\,.$$

For compact $K \subset U_\Gamma$, define the exit time

$$\hat{\mu}_n(K) = \inf\left\{t\ge 0 \mid \hat{Y}_n(t-) \notin K^\circ \text{ or } \hat{Y}_n(t) \notin K^\circ\right\}.$$

Then for any compact set $K \subset U_\Gamma$ the sequence $(\hat{Y}_n^{\hat{\mu}_n(K)}, \hat{\mu}_n)$ is relatively compact in the Skorokhod topology. Moreover, any limit point $(\hat{Y},\hat{\mu})$ of $(\hat{Y}_n,\hat{\mu}_n)$ satisfies $\hat{Y}(t)\in\Gamma$ a.s. for any t and solves

$$\begin{aligned}\hat{Y}(t) = \hat{Y}(0) &+ \int_0^{t\wedge\hat{\mu}} P_{T_{\hat{Y}(s)}\Gamma}\Big(\nabla_w f(\hat{Y}(s))\cdot \mathrm{d}b_s + \sigma_0\nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s\Big)\\ &+ \frac{1}{2}\int_0^{t\wedge\hat{\mu}} \partial^2\Phi(\hat{Y})[\Sigma(\hat{Y}(s))]\,\mathrm{d}s\,,\end{aligned} \tag{58}$$

where $(b_s)_{s\ge0}$ is a d-dimensional Brownian motion and $B_{s,k,\ell}$ for $1\le k<\ell\le d$ are Brownian motions with $B_{s,\ell,k} := B_{s,k,\ell}$, and all Brownian motions are independent; the mobility Σ is defined as

$$\Sigma(w)_{ij} := \partial_{w_i} f(w)\cdot\partial_{w_j} f(w) + \sigma_0^2 \partial_{w_i} H(w) : \partial_{w_j} H(w)\,, \qquad \text{for } 1\le i,j\le m\,.$$

$P_{T_y\Gamma}$ is the orthogonal projection onto the tangent space $T_y\Gamma$ of Γ at y.

In addition, we have

$$\hat{\mu} \ge \inf\{t\ge 0 \mid \hat{Y}(t)\notin K^\circ\} \qquad \text{almost surely.} \tag{59}$$

Remark 30 (No Assumption 7) *Theorem 23 requires Assumption 7, which specifies relations between the modified loss $\hat{L}$ and the structure of the noise, as given by $\rho(\sigma)$. For Theorem 29 we do not need this assumption, because $\hat{L}$ given by Definition 28 automatically has similar properties.*

Remark 31 (Convergence of the whole sequence) *Similarly to the remark after Theorem 23, whenever the weak solution of* (58) *is unique, the usual argument implies that the whole sequence $\hat{Y}_n^{\mu_n(K)}$ converges up to the time τ,*

$$\tau := \inf\{t \geq 0 \mid \hat{Y}(t) \notin K^\circ\}.$$

A sufficient condition for such weak uniqueness is local Lipschitz continuity of the drift and mobility functions in (58) *(see e.g. Hsu (2002, Th. 1.1.10)).*

Remark 32 (Implicit Itô correction terms) *Since* (58) *is a stochastic differential equation in Itô form, the noise term should be accompanied by a corresponding drift term in order to preserve the condition $\hat{Y}(t) \in \Gamma$. Li, Wang, and Arora (2022a) discuss this aspect in some detail, and show how the final integral contains, as part of this integral, the necessary 'Itô correction terms'.*

Remark 33 (Comparison with Li et al. (2022a)) *At a high level of abstraction, Theorem 29 above is similar to Li et al. (2022a, Th. 4.6). However, in Theorem 29 the conditions on $\hat{L}$ and on the noise are much more general, opening the door to e.g. the applications to minibatching and stochastic gradient Langevin descent in Section 5.*

Proof As in the proof for the general case in Theorem 23, we begin by checking the assumptions of Theorem 20.

Verification of Assumption 16. By definition of $\hat{A}_n$ we get for $\alpha_n \to 0$ and bounded σ_n for any $\delta > 0$

$$\inf_{0<t\leq T}(\hat{A}_n(t+\delta) - \hat{A}_n(t)) = \inf_{0<t\leq T}\left(\alpha_n\left\lfloor\frac{t+\delta}{\alpha_n^2\sigma_n^2}\right\rfloor - \alpha_n\left\lfloor\frac{t}{\alpha_n^2\sigma_n^2}\right\rfloor\right) \geq \alpha_n\left\lfloor\frac{\delta}{\alpha_n^2\sigma_n^2}\right\rfloor \to \infty.$$

At the same time

$$\sup_{t\geq 0} \hat{A}_n(t) - \hat{A}_n(t-) = \alpha_n \to 0 \qquad \text{for } \alpha_n \to 0\,.$$

Verification of Assumption 17. Since $\hat{L}$ is quadratic in η, we get from assumption (57) the estimate

$$\sup_{0<t\leq T\wedge\hat{\mu}_n(K)} |\Delta\hat{Z}_n| = \alpha_n\left(1 + \sup_{0<t\leq T\wedge\hat{\mu}_n(K)} |\eta_k^n|^2\right) \Rightarrow 0.$$

Verification of Assumption 18. Note that due to the structure of $\hat{L}$ the noise process is a martingale because $\nabla_w H_{i,i} = 0$ and has a simplified form:

$$\Delta\hat{Z}_n = \alpha_n \nabla_w f(\hat{W}_n^{\hat{\mu}_n(K)}(\alpha_n^2\sigma_n^2 k))\eta_k^n + \tfrac{1}{2}\alpha_n \nabla_w H(\hat{W}_n^{\hat{\mu}_n(K)}(\alpha_n^2\sigma_n^2 k))[\eta_k^n, \eta_k^n]$$

We verify that $[\hat{Z}_n^{\hat{\mu}_n(K)}](t)$ is uniformly integrable for any $t \in \mathbb{R}^+$ by estimating

$$[\hat{Z}_n^{\hat{\mu}_n(K)}](t) \leq 2\alpha_n^2 \sum_{k \leq \frac{t}{\alpha_n^2\sigma_n^2}} \sup_{w\in K} |\nabla_w f(w)|^2 \sum_{i=1}^d (\eta_{k,i}^n)^2 + \alpha_n^2 \sum_{k \leq \frac{t}{\alpha_n^2\sigma_n^2}} \sup_{w\in K} |\nabla_w H(w)|^2 \sum_{\substack{i,j=1 \\ i\neq j}}^d (\eta_{k,i}^n \eta_{k,j}^n)^2.$$

By assumption $\operatorname{Var}(\eta_{k,i}^n) = \sigma_n^2$ and as $\mathbb{E}\,\eta_{k,i}^n = 0$ also $\operatorname{Var}(\eta_{k,i}^n\eta_{k,j}^n) = \operatorname{Var}(\eta_{k,i}^n)\operatorname{Var}(\eta_{k,j}^n) = \sigma_n^4$, so by the Functional Central Limit Theorem (see e.g. Billingsley (1968, Th. 8.2)) we have for $i = 1, \dots, d$ and $1 \leq \ell \neq m \leq d$ the convergence

$$\sum_{k \leq \frac{t}{\alpha_n^2\sigma_n^2}} \alpha_n \eta_{k,i}^n =: b_{t,i}^n \Rightarrow b_{t,i}, \quad \text{and} \quad \sum_{k \leq \frac{t}{\alpha_n^2\sigma_n^2}} \alpha_n \eta_{k,\ell}^n \eta_{k,m}^n =: B_{t,\ell,m}^n \xRightarrow{n\to\infty} \sigma_0 B_{t,\ell,m}\,, \tag{60}$$

where $b_{t,i}$ for $i = 1, \dots, d$ and $B_{t,\ell,m}$ for $1 \leq \ell \neq m \leq d$ are standard Brownian motions. We have the symmetry $B_{t,\ell,m} = B_{t,m,\ell}$. We note that $b_{t,i}^n$ for $1 \leq i \leq d$ and $B_{t,\ell,m}^n$ for $1 \leq \ell < m \leq d$ are uncorrelated for any n. Hence, so are the limits and as normal random variables, we find that $b_{t,i}$ for $i = 1, \dots, d$ and $B_{t,\ell,m}$ for $1 \leq \ell < m \leq d$ are independent standard Brownian motions. We then combine Katzenberger (1991, Proposition 4.3) and Katzenberger (1991, Proposition 4.4) and conclude that $\hat{Z}_n^{\hat{\mu}_n(K)}$ satisfies Assumption 18.

We then apply Theorem 20 and obtain the convergence of $\hat{Y}_n$ to a limiting process $\hat{Y}$ that follows the limiting evolution (26).

Identification of the limiting dynamics. For studying the limiting dynamic of $\hat{Z}_n$, we introduce the intermediate process

$$\hat{Z}_n^Y(t) = \alpha_n \sum_{k \leq \frac{t}{\alpha_n^2\sigma_n^2}} \left(\nabla_w f(\hat{Y}_n(s)) \cdot \eta_k^n + \tfrac{1}{2}\nabla_w H(\hat{Y}_n(s))[\eta_k^n, \eta_k^n]\right),$$

where we abbreviate $s := s(k) := \alpha_n^2\sigma_n^2 k$. For the intermediate process we obtain the estimate

$$\begin{aligned} &\sup_{0<t\leq T}\left|\hat{Z}_n(t) - \int_0^t \nabla_w f(\hat{Y}(s)) \cdot \mathrm{d}b_s - \sigma_0 \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s\right| \qquad (61)\\ &\leq \sup_{0<t\leq T}\left|\hat{Z}_n(t) - \hat{Z}_n^Y(t)\right| + \sup_{0<t\leq T}\left|\hat{Z}_n^Y(t) - \int_0^t \nabla_w f(\hat{Y}(s)) \cdot \mathrm{d}b_s - \sigma_0 \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s\right|, \end{aligned}$$

where $(b_{s,i})_{s\geq 0}$, $(B_{s,k,l})$ for $i = 1, \dots, d$ and $1 \leq k < l \leq d$ are independent Brownian motions and we set $B_{s,l,k} := B_{s,k,l}$. For the first term analogously to the main theorem we get

$$\left|\hat{Z}_n(t) - \hat{Z}_n^Y(t)\right| \leq \alpha_n \left|\sum_{k \leq \frac{t}{\alpha_n^2\sigma_n^2}} \left(\nabla_w f(\hat{W}_n(s)) - \nabla_w f(\hat{Y}_n(s))\right) \cdot \eta_k^n\right| \tag{62}$$

$$+ \frac{1}{2}\alpha_n \left|\sum_{k \leq \frac{t}{\alpha_n^2\sigma_n^2}} \left(\nabla_w H(\hat{W}_n(s)) - \nabla_w H(\hat{Y}_n(s))\right)[\eta_k^n, \eta_k^n]\right|. \tag{63}$$

Recall from Proposition 14 that

$$\hat{W}_n(t) - \hat{Y}_n(t) = \phi\big(\hat{W}_n(0), A_n(t)\big) - \Phi\big(\hat{W}_n(0)\big) = u(t)e^{-\beta A_n(t)}$$

for some $u \in C_b([0,T])$. Note that both terms are martingales, so $\mathbb{E}\big((X_T^n)^2\big) \to 0$ implies $\sup_{t\le T} |X_t^n| \Rightarrow 0$. Then by compactness of K and regularity of $\hat{L}$ similarly to the proof of Theorem 23, we conclude that

$$\begin{aligned}
&\mathbb{E}\Big[\alpha_n \Big| \sum_{k \le \frac{t}{\alpha_n^2\sigma_n^2}} \Big(\nabla_w f(\hat{W}_n(s)) - \nabla_w f(\hat{Y}_n(s))\Big) \cdot \eta_k^n \Big|\Big]^2 \\
&\quad \lesssim \alpha_n^2 \sum_{k \le \frac{t}{\sigma_n^2}} e^{-2\beta A_n(\alpha_n^2\sigma_n^2 k)}\, \mathbb{E}|\eta_k^n|^2 = \alpha_n \sigma_n^2 \int_0^t e^{-2\beta A_n(s)}\, \mathrm{d}A_n(s) \\
&\quad = \frac{\alpha_n\sigma_n^2}{\beta}(1 - e^{-\beta A_n(t)}) \to 0 \qquad \text{as } n \to \infty\,.
\end{aligned}$$

Hence, we obtain for term (62) the convergence

$$\sup_{0\le t\le T} \alpha_n \Big| \sum_{k \le \frac{t}{\alpha_n^2\sigma_n^2}} \Big(\nabla_w f(\hat{W}_n(s)) - \nabla_w f(\hat{Y}_n(s))\Big)\eta_k^n \Big| \Rightarrow 0\,.$$

Analogous bound holds for the second term (63).

We split the second component in (61) into the two contributions

$$\begin{aligned}
&\sup_{0<t\le T} \Big| \hat{Z}_n^Y(t) - \int_0^t \nabla_w f(\hat{Y}(s)) \cdot \mathrm{d}b_s - \sigma_0 \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s \Big| \qquad (64)\\
&\quad \le \sup_{0<t\le T} \Big| \alpha_n \sum_{k \le \frac{t}{\alpha_n^2\sigma_n^2}} \nabla_w f(\hat{Y}_n(s)) \cdot \eta_k^n - \int_0^t \nabla_w f(\hat{Y}(s)) \cdot \mathrm{d}b_s \Big| \\
&\qquad + \sup_{0<t\le T} \Big| \alpha_n \sum_{k \le \frac{t}{\alpha_n^2\sigma_n^2}} \nabla_w H(\hat{Y}_n(s))[\eta_k^n, \eta_k^n] - \sigma_0 \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s \Big|.
\end{aligned}$$

By using the processes $(b_t^n)_{t\ge 0}$ and $(B_t^n)_{t\ge 0}$ from (60), we decompose the error terms further into

$$\begin{aligned}
&\sup_{0<t\le T} \Big| \hat{Z}_n^Y(t) - \int_0^t \nabla_w f(\hat{Y}(s)) \cdot \mathrm{d}b_s - \sigma_0 \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s \Big| \qquad (65)\\
&\quad \le \sup_{0<t\le T} \Big| \int_0^t \nabla_w f(\hat{Y}(s))\, \mathrm{d}b_s^n - \int_0^t \nabla_w f(\hat{Y}(s)) \cdot \mathrm{d}b_s \Big| \qquad (66)\\
&\qquad + \sup_{0<t\le T} \Big| \int_0^t \nabla_w f(\hat{Y}_n(s)) \cdot \mathrm{d}b_s^n - \int_0^t \nabla_w f(\hat{Y}(s)) \cdot \mathrm{d}b_s^n \Big| \qquad (67)\\
&\quad + \sup_{0<t\le T} \Big| \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s^n - \sigma_0 \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s \Big| \qquad (68)\\
&\qquad + \sup_{0<t\le T} \Big| \int_0^t \nabla_w H(\hat{Y}_n(s)) : \mathrm{d}B_s^n - \int_0^t \nabla_w H(\hat{Y}(s)) : \mathrm{d}B_s^n \Big|. \qquad (69)
\end{aligned}$$

We recall the convergence from (60) and observe that all the integrals are relatively compact by Katzenberger (1991, Proposition 4.4). By noting that $\hat{Y}(s), \hat{Y}_n(s) \in K$ with K compact, by using assumption on the regularity of $\hat{L}$, by Kurtz and Protter (1991, Theorem 2.2) we conclude that the terms (66) and (68) converge to zero. The terms (67) and (69) convergence by the continuous mapping theorem,

$$\nabla_w f(\hat{Y}_n(s)) \Rightarrow \nabla_w f(\hat{Y}(s)) \qquad \text{and} \qquad \nabla_w H(\hat{Y}_n(s)) \Rightarrow \nabla_w H(\hat{Y}(s))\,.$$

Hence, by an application of Kurtz and Protter (1991, Theorem 2.2), we get the convergence

$$\begin{aligned}
&\sup_{0<t\leq T}\left|\int_0^t \nabla_w f(\hat{Y}_n(s))\cdot \mathrm{d}b_s^n - \int_0^t \nabla_w f(\hat{Y}(s))\cdot \mathrm{d}b_s^n\right| \\
&\qquad \leq \sup_{0<t\leq T}\left|\int_0^t \nabla_w f(\hat{Y}_n(s))\cdot \mathrm{d}b_s^n - \int_0^t \nabla_w f(\hat{Y}(s))\cdot \mathrm{d}b_s\right| \\
&\qquad\quad + \sup_{0<t\leq T}\left|\int_0^t \nabla_w f(\hat{Y}(s))\cdot \mathrm{d}b_s^n - \int_0^t \nabla_w f(\hat{Y}(s))\cdot \mathrm{d}b_s\right| \Rightarrow 0.
\end{aligned}$$

Similarly (69) $\Rightarrow 0$. Finally note that

$$\mathrm{d}[\hat{Z}^i, \hat{Z}^j]_s = \partial_{w_i} f(\hat{Y}(s))\cdot \partial_{w_j} f(\hat{Y}(s)) + \sigma_0 \nabla_{w_i} H(\hat{Y}(s)) : \sigma_0 \nabla_{w_j} H(\hat{Y}(s)) = \Sigma(Y(s))_{ij}\,\mathrm{d}s\,.$$

Hence, the stated result in Theorem 29 follows by applying Theorem 20. ∎

5. Examples

In this section we apply Theorems 23 and 29 to a number of examples. Table 1 gives an overview.

In each of the examples below we apply either Theorem 23 or Theorem 29 to obtain a convergence result. Because the statements of those theorems involve a specific type of convergence, which deals with the initial jump and requires a proper speedup and restriction to a compact set, we introduce the notion of *Katzenberger convergence* to simplify the formulation of the results below.

As in Section 4, consider a sequence of stochastic processes $(W_n)_{n\in\mathbb{N}}$ and a sequence of integrators $(A_n)_{n\in\mathbb{N}}$, and let Y_n be the process after the jump correction

$$Y_n(t) := W_n(t) - \phi(W_n(0), A_n(t)) + \Phi\left(W_n(0)\right). \tag{70}$$

For a compact set K we again set

$$\mu_n(K) := \inf\{t \geq 0 \mid Y_n(t-) \notin K^\circ \text{ or } Y_n(t) \notin K^\circ\}.$$

Let Y be a deterministic or stochastic process with continuous sample paths and set

$$\tau := \inf\{t \geq 0 \mid Y(t) \notin K^\circ\}.$$

Theorems 23 and 29 provide convergence in distribution; by the Skorokhod representation theorem, we can assume without loss of generality that the processes W_n, Y_n, and Y are defined on the same probability space.

Type	$\hat{L}(w,\eta)$	$\mathrm{Reg}(w)$	**Rate**	**Section**
Independent of the structure of L, with η having the same dimension as w:				
Gaussian D-Connect	$L(w\odot(1+\eta))$	$\frac{1}{2}\partial^2 L(w)[w,w]$	$\alpha_n\sigma_n^2$	5.1.1
Bernoulli D-Connect	$L(w\odot(1+\eta))$	(79)	$\alpha_n\sigma_n^2$	5.1.1
SGLD	$L(w)+w^\top\eta$	$\frac{1}{2}\log\lvert\nabla^2 L\rvert_+$	$\alpha_n^2\sigma_n^2$	5.1.2
Anti-PGD	$L(w+\eta)$	$\frac{1}{2}\Delta_w L(w)$	$\alpha_n\sigma_n^2$	5.1.3
Variables η indexed by i, the index of the samples:				
Mini-batching	$\frac{1}{N}\sum_{i=1}^N(1+\eta_i)\ell(w,x_i,y_i)$	0	$\ll\alpha_n^2\sigma_n^2$	5.2.1
Label noise	$\frac{1}{N}\sum_{i=1}^N(f_w(x_i)-y_i-\eta_i)^2$	$\frac{1}{2N}\Delta L(w)$	$\alpha_n^2\sigma_n^2$	5.2.2
Classical Dropout, with mean-square loss:				
OLM-DO	$f_w^{\mathrm{drop}}(x,\eta)=\langle u^{\odot 2}-v^{\odot 2},x\odot(1+\eta)\rangle$	$\frac{1}{2N}\sum_{i,j}(u_j^2-v_j^2)^2x_{ij}^2$	$\alpha_n\sigma_n^2$	5.3.1
ShNN-DO	$f_w^{\mathrm{drop}}(x,\eta)=\sum_j a_j(1+\eta_j)s(b_j^\top x)$	$\frac{1}{N}\sum_{i,j}a_j^2 s(b_j^\top x_i)^2$	$\alpha_n\sigma_n^2$	5.3.2
DeepNN-DO	(94)	$\frac{1}{2}\Delta_\eta\hat{L}(w,0)$	$\alpha_n\sigma_n^2$	5.3.2

Table 1: List of examples discussed in Section 5. The column 'Rate' indicates the rate at which iterations should be mapped to continuous time in order to follow the evolution; for instance, '$\alpha_n\sigma_n^2$' means that step k is mapped to time $t_k := \alpha_n\sigma_n^2 k$. Therefore smaller numbers mean slower evolution, requiring stronger speedup to see non-trivial evolution. See the referenced sections for details.

Definition 34 (Katzenberger convergence) *We say that W_n converges* in the sense of Katzenberger *in U to Y if for every compact subset $K\subset U$ we have*

$$Y_n^{\mu_n(K)\wedge\tau}\Rightarrow Y^\tau \qquad \text{and for all } \delta>0,\quad W_n^{\mu_n(K)\wedge\tau}\big|_{[\delta,\infty)}\Rightarrow Y^\tau\big|_{[\delta,\infty)}\,.$$

Here the convergences are in Skorokhod topology.

In Section 4 we discussed that Theorems 23 and 29 provide such convergence when the limiting process Y has a uniquely defined deterministic or stochastic evolution up to time τ.

5.1 Methods independent of L: DropConnect and SGLD

5.1.1 DropConnect

Dropout is a regularization technique originally introduced for neural networks by Wager et al. (2013) and Srivastava et al. (2014). The idea is to balance the importance of all the neurons by temporarily disabling some of them at every iteration of the optimization. In this section we study the specific case of *DropConnect* (Wan et al., 2013), which is more general than other types of Dropout, in the sense that it can be applied to 'any' loss function L, not just those generated by neural networks.

In DropConnect, given a loss function L, we construct the noise-injected loss $\hat{L}$ by

$$\hat{L}(w,\eta) := L(w\odot(1+\eta))\,. \tag{71}$$

In the most common form the filters η are chosen to be i.i.d. Bernoulli random variables,

$$\eta_i = \begin{cases} -1 & \text{w.p. } p\,, \\ \dfrac{p}{1-p} & \text{w.p. } 1-p\,. \end{cases} \tag{72}$$

Note that $\mathbb{E}\eta_i = 0$ and $\sigma^2 := \mathrm{Var}(\eta_i) = p/(1-p) \to 0$ as $p \to 0$. When $\eta_i = -1$, the corresponding parameter is zeroed, or dropped out, which explains the name. Note that the limit of small σ^2 is the limit in which $p \downarrow 0$, i.e. in which vanishingly few parameters are dropped.

As an alternative for Bernoulli DropConnect we also consider Gaussian DropConnect, in which the η_i are i.i.d. normal random variables

$$\eta_i \sim \mathcal{N}(0, \sigma^2). \tag{73}$$

Gaussian DropConnect. The case of Gaussian noise variables η fits directly into the structure of Theorem 23, and we now state the corresponding result.

Corollary 35 (Convergence for Gaussian DropConnect) *Let $L \in C^4(\mathbb{R}^m)$ satisfy Assumption 10, and assume that L has at most polynomial growth. Define $\hat{L}$ by (71), and let W_n be the process characterized in (34), where the $\eta_{k,i}$ are i.i.d. Gaussian variables as in (73). Assume that $W_n(0) \Rightarrow W_0 \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Finally, let α_n and σ_n be arbitrary sequences converging to zero.*

Then for any compact set $K \subset U_\Gamma$, the process W_n converges to Y in the Katzenberger sense (see Definition 34), where Y solves the constrained gradient-flow equation

$$\frac{dY}{dt} = -P_{T_Y\Gamma}\nabla_w \mathrm{Reg}(Y), \qquad Y(t) \in \Gamma, \qquad \textit{and} \qquad Y(0) = \Phi(W_0),$$

with

$$\mathrm{Reg}(w) := \frac{1}{2}\Delta_\eta \hat{L}(w,0) = \frac{1}{2}\partial^2 L(w)[w,w] = \frac{1}{2}\sum_{j=1}^m w_j^2 \partial^2_{w_j w_j} L(w)\,. \tag{74}$$

Proof [Proof of Corollary 35] The result follows fairly directly from Theorem 23. If L grows at most polynomially, then $\hat{L}$ and the Gaussian random variables satisfy Assumption 7 for some p, and by Lemma 55 the condition (35) also is satisfied. The assertion of Corollary 35 is then a translation of the assertion of Theorem 23. ∎

Remark 36 (Empirical loss) *If L is an empirical loss of the form*

$$L(w) := \frac{1}{N}\sum_{i=1}^N \big(f(w, x_i) - y_i\big)^2\,, \tag{75}$$

then on Γ the function Reg *can be written as*

$$\mathrm{Reg}(w) = \frac{1}{N}\sum_{i=1}^N \big|w \odot \nabla_w f(w, x_i)\big|^2\,. \tag{76}$$

Note that although $\mathrm{Reg}(w)$ *is itself a second derivative of* $\hat{L}$*, only first derivatives of* f *enter in the expression* (76)*, because the second derivatives of* f *are multiplied by factors* $(f(w,x_i)-y_i)$ *that vanish on* Γ*.*

Remark 37 (Connection with generalization) *The form of the function* Reg *in* (74) *is very similar to the 'relative flatness' introduced by Petzka et al. (2021), and for which they prove rigorous generalization benefits.*

Petzka and co-authors consider functions f *of the form* $f(w,x)=g(w\phi(x))$*, where* w *is organized as a matrix, with* $w\in\mathbb{R}^{m_1\times m_2}$ *and* $x\in\mathbb{R}^{m_2}$*. Taking* $\phi(x)=x$ *to connect with this paper, the 'relative flatness' (Petzka et al., 2021, Def. 3) of the empirical loss of such an* f *can be written as*

$$\frac{2}{N}\sum_{i=1}^{N}\sum_{s,s'=1}^{m_1}\sum_{j,j'=1}^{m_2} w_{sj}w_{s'j}\partial_{w_{sj'}}f(w,x_i)\partial_{w_{s'j'}}f(w,x_i) \tag{77}$$

To compare, we can write the expression (76) *in a very similar form,*

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{s,s'=1}^{m_1}\sum_{j,j'=1}^{m_2} w_{sj}w_{s'j}\partial_{w_{sj'}}f(w,x_i)\partial_{w_{s'j'}}f(w,x_i)\ \delta_{ss'}\delta_{jj'}\,. \tag{78}$$

Note how (78) *is a 'diagonal' form of* (77)*.*

The two expressions (78) *and* (77) *share a number of properties that are consistent with good generalization behaviour. To start with, as also remarked by Petzka et al. (2021), both are invariant under coordinate-wise rescaling of the parameters* w*, because each derivative is accompanied by a multiplication with the corresponding coordinate of* w*. They also scale quadratically in* f*, just as the loss function* (75)*, and if one considers* f *to be a neural network of depth* D *(see below) then the scaling in terms of* w *is of the form* w^{2D}*, both for the loss and for the expressions* (78) *and* (77)*. These scaling properties are necessary for a functional to be admissible as a measure of generalization ability.*

Bernoulli DropConnect. The case of Bernoulli-distributed η as in (72) is not covered by Theorem 23, because all higher moments M_k for $k\geq 3$ scale as σ^2, and therefore violate the condition (12a). In general, we have $C_2>0$ in Assumption 7, and therefore they also do not satisfy condition (12b), except for the special case of shallow neural networks—see Section 5.3.2 below. As it turns out, however, the proof of Theorem 23 does apply to this situation, provided we prove the three statements of Lemma 27 for this particular case. This leads to the following proposition.

Proposition 38 (Convergence for Bernoulli DropConnect) *Let* $L\in C^4(\mathbb{R}^m)$ *satisfy Assumption 10. Define* $\hat{L}$ *by* (71)*, and let* W_n *be the process characterized in* (34)*, where the* $\eta_{k,i}$ *are i.i.d. Bernoulli random variables as in* (72)*, with dropout probability* p_n*. Assume that* $W_n(0)\Rightarrow W_0\in U_\Gamma$ *for some locally attractive neighbourhood* U_Γ *of* Γ*. Finally, let* $\alpha_n\to 0$ *and* $\sigma_n^2:=p_n/(1-p_n)\to 0$*.*

Then for any compact set $K\subset U_\Gamma$*, the process* W_n *converges to* Y *in the Katzenberger sense (see Definition 34), where* Y *solves the constrained gradient-flow equation*

$$\frac{dY}{dt}=-P_{T_Y\Gamma}\nabla_w\mathrm{Reg}(Y),\qquad Y(t)\in\Gamma,\qquad \textit{and}\qquad Y(0)=\Phi(W_0),$$

with

$$\mathrm{Reg}(w) := \partial L(w)[w] + \sum_{j=1}^{m} \big(L(w \odot \mathfrak{d}_j) - L(w)\big). \tag{79}$$

Here $\mathfrak{d}_j$ *is the inverted unit vector in* $\mathbb{R}^m$*,*

$$\mathfrak{d}_j := (1, \dots, 1, \underbrace{0}_{j^{\text{th}} \text{ position}}, 1, \dots, 1) = 1 - \mathrm{e}_j.$$

Remark 39 (Structure of the Bernoulli DropConnect regularizer)
The functional form of Reg *in* (79) *can be interpreted as a version of the Gaussian regularizer* (74) *with the local derivative replaced by a non-local one. Indeed, writing* $\mathfrak{d}_j = 1 - \mathrm{e}_j$ *we Taylor-develop* L *as*

$$\begin{aligned} L(w \odot \mathfrak{d}_j) &= L(w \odot (1 - \mathrm{e}_j)) \\ &= L(w) - \partial L(w)[\mathrm{e}_j \odot w] + \frac{1}{2}\partial^2 L(w)[\mathrm{e}_j \odot w, \mathrm{e}_j \odot w] + O(|\mathrm{e}_j \odot w|^3). \end{aligned}$$

If we pretend for the moment that e_j *is 'small', and discard the final term, then*

$$\begin{aligned} \sum_{j=1}^{m} \big(L(w \odot \mathfrak{d}_j) - L(w)\big) &\approx \sum_{j=1}^{m} \Big[-\partial L(w)[\mathrm{e}_j \odot w] + \frac{1}{2}\partial^2 L(w)[\mathrm{e}_j \odot w, \mathrm{e}_j \odot w]\Big] \\ &= -\partial L(w)[w] + \frac{1}{2}\sum_{j=1}^{m} w_j^2 \partial^2_{w_j w_j} L(w). \end{aligned}$$

In this approximation, therefore, the Bernoulli regularizer (79) *reduces to the Gaussian regularizer* (74)*.*

Proof As described above, we re-prove the assertions of Lemma 27, i.e. properties (41–43), for this case. Given those assertions the proof of Theorem 23 applies to this situation.

We first estimate $\Delta \mathsf{Z}_n(w, \eta) = \alpha_n \nabla L(w) - \alpha_n \nabla_w (L(w \odot (1 + \eta)))$ for this setup. Let $\tilde{K} \supset \hat{K}$ be a compact set that is large enough to contain $w(1 + \eta)$ for all η and all $w \in \hat{K}$. Using the boundedness of derivatives of $\hat{L}$ on $\tilde{K}$, we then have for all η and all $w \in \hat{K}$

$$|\Delta \mathsf{Z}_n(w, \eta)| \le \alpha_n \Big(1 + \frac{1}{1 - p_n}\Big) \|\nabla L\|_{L^\infty(\tilde{K})} = O(\alpha_n),$$

which proves (41). To prove the estimate on $|\Delta \mathsf{Z}_n|^2$ in (43) we note that on the event E that all coordinates of η are non-zero, which has probability $(1 - p_n)^m = 1 - O(\sigma_n^2)$, we have the additional estimate

$$\sup_{w \in K} |\Delta \mathsf{Z}_n(w, \eta)| \le \alpha_n \|\nabla^2 L\|_{L^\infty(\tilde{K})} \underbrace{\frac{p_n}{1 - p_n}}_{\sigma_n^2} \sup_{w \in K} |w| = O(\alpha_n \sigma_n^2).$$

It follows that for all $w \in \hat{K}$,

$$\mathbb{E}_\eta \sup_{w \in \hat{K}} |\Delta \mathsf{Z}_n(w, \eta)|^2 \le O(\alpha_n^2 \sigma_n^4)\, \mathbb{P}(E) + O(\alpha_n^2)\, \mathbb{P}(E^c) = o(\alpha_n \sigma_n^2),$$

implying the first estimate in (43).

We next estimate $\Delta \mathsf{F}_n$ as defined in (40b). The expectation over the set of independent Bernoulli random variables can be expressed in terms of $\{0,1\}^m$-valued outcomes b in the form

$$\begin{aligned}\Delta \mathsf{F}_n(w) &= \mathbb{E}_\eta\, \Delta \mathsf{Z}_n(w,\eta) = \alpha_n \nabla L(w) - \alpha_n\, \mathbb{E}_\eta\, \nabla_w (L(w \odot (1+\eta))) \\ &= \alpha_n \nabla L(w) - \alpha_n \sum_{b \in \{0,1\}^m} p_n^{m-|b|} (1-p_n)^{|b|-1}\, b \odot \nabla L\Big(\frac{1}{1-p_n} w \odot b\Big).\end{aligned}$$

We estimate the term inside the sum according to the number of zero coordinates in b:

$$b \odot (\nabla L)\Big(\frac{1}{1-p_n} w \odot b\Big) = \begin{cases} \nabla L(w) + \frac{p_n}{1-p_n} \partial \nabla L(w)[w] + O(p_n^2) & \text{if } |b| = m, \\ \mathfrak{d}_j \odot (\nabla L)(w \odot \mathfrak{d}_j) + O(p_n) & \text{if } |b| = m-1,\ b = \mathfrak{d}_j, \\ O(1) & \text{if } |b| \le m-2, \end{cases}$$

with the O symbols being uniform in $w \in \hat{K}$.

Using the expression for the gradient of the regularizer (79),

$$\nabla \mathrm{Reg}(w) = \partial \nabla L(w)[w] - (m-1)\nabla L(w) + \sum_{j=1}^m \mathfrak{d}_j \odot (\nabla L)(w \odot \mathfrak{d}_j).$$

we then split the sum over b into three parts and estimate accordingly

$$\Delta \mathsf{F}_n(w) + \alpha_n \sigma_n^2 \nabla \mathrm{Reg}(w) = \mathrm{I} + \mathrm{II} + \mathrm{III},$$

with

$$\begin{aligned}\frac{1}{\alpha_n}\mathrm{I} &= \nabla L(w) - (1-p_n)^{m-1} \nabla L(w) - p_n (1-p_n)^{m-2} \partial \nabla L(w)[w] + O(p_n^2) \\ &\qquad + \sigma_n^2 \partial \nabla L(w)[w] - \sigma_n^2 (m-1) \nabla L(w), \\ \frac{1}{\alpha_n}\mathrm{II} &= \sum_{j=1}^m \Big[-p_n (1-p_n)^{m-2}\, \mathfrak{d}_j \odot (\nabla L)(w \odot \mathfrak{d}_j) + \sigma_n^2\, \mathfrak{d}_j \odot (\nabla L)(w \odot \mathfrak{d}_j) \Big] + O(p_n^2), \\ \frac{1}{\alpha_n}\mathrm{III} &= \sum_{|b| \le m-2} p_n^{m-|b|} (1-p_n)^{|b|-1} b \odot \nabla L\Big(\frac{1}{1-p_n} w \odot b\Big).\end{aligned}$$

We estimate the terms one by one. For I we find

$$\begin{aligned}\frac{1}{\alpha_n}\mathrm{I} &= \Big[1 - (1 - (m-1)p_n + O(p_n^2)) - \sigma_n^2 (m-1)\Big] \nabla L(w) \\ &\qquad + (-p_n (1-p_n)^{m-2} + \sigma_n^2) \partial \nabla L(w)[w] \\ &= O(p_n^2) = O(\sigma_n^4), \qquad \text{uniformly in } w \in \hat{K}.\end{aligned}$$

For II we write

$$\frac{1}{\alpha_n}\mathrm{II} = \sum_{j=1}^m \Big[-p_n (1-p_n)^{m-2} + \sigma_n^2 \Big]\, \mathfrak{d}_j \odot (\nabla L)(w \odot \mathfrak{d}_j) + O(p_n^2) = O(p_n^2) = O(\sigma_n^4),$$

and for the third term we immediately find $\alpha_n^{-1}\mathrm{III} = O(p_n^2) = O(\sigma_n^4)$. Combining all these estimates we conclude that

$$\sup_{w\in K}\left|\Delta \mathsf{F}_n(w) + \alpha_n\sigma_n^2\nabla\mathrm{Reg}(w)\right| = O(\alpha_n\sigma_n^4),$$

thereby establishing (42).

Finally, to prove the second estimate in $(43)_2$ we write

$$|\Delta \mathsf{F}_n(w)| \le |\Delta \mathsf{F}_n(w) + \alpha_n\sigma_n^2\nabla\mathrm{Reg}(w)| + |\alpha_n\sigma_n^2\nabla\mathrm{Reg}(w)|,$$

and the estimate $(43)_2$ follows from (42) and the regularity of Reg. ■

5.1.2 Stochastic Gradient Langevin Descent

Stochastic Gradient Langevin Descent (Gelfand and Mitter, 1991; Raginsky et al., 2017; Mei et al., 2018) is a form of gradient descent in which at each iteration a centered Gaussian perturbation is added to the gradient. In our notation this corresponds to

$$\hat{L}(w,\eta) := L(w) + w^\top\eta, \qquad \text{with } \eta \sim \mathcal{N}(0,\sigma^2 I_m) \text{ i.i.d..} \tag{80}$$

This structure is of the degenerate form (55) and therefore is covered by Section 4.2. This results in the following Corollary.

Corollary 40 (Convergence for Stochastic Gradient Langevin Descent) *Let $L \in C^4(\mathbb{R}^m)$ satisfy Assumption 10. Define $\hat{L}$ by (80), and let W_n be the process characterized in (34), where the $\eta_{k,i}$ are i.i.d. centered normal random variables with variance σ_n^2. Assume that $W_n(0) \Rightarrow W_0 \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Finally, let $\alpha_n \to 0$ and $\sigma_n \to \sigma_0 \ge 0$.*

Then for any compact set $K \subset U_\Gamma$, the process W_n converges to Y in the Katzenberger sense (see Definition 34), where Y solves the SDE

$$\mathrm{d}Y(t) = P_{T_Y\Gamma}\,\mathrm{d}b(t) + \frac{1}{2}(\nabla^2 L(Y))^\dagger\partial^2\nabla L(Y)[P_{T_Y\Gamma}]\,dt - \frac{1}{2}P_{T_Y\Gamma}\nabla\log|\nabla^2 L(Y)|_+\,\mathrm{d}t, \tag{81}$$

constrained to $Y(t)\in\Gamma$, with $Y(0) = \Phi(W_0)$. (Recall that $|A|_+$ is the product of the non-zero eigenvalues of A, and $\dagger$ indicates the Moore-Penrose pseudoinverse).

Remark 41 *As observed by Li et al. (2022a), the first two terms in (81) form a geometrically intrinsic Brownian motion on the manifold Γ, while the final term acts as a drift parallel to Γ.*

Proof The assumptions of Theorem 29 are satisfied, with $f(w) = w$ and $H(w) = 0$ in (55) and using Lemma 55 to show (57). We conclude that the evolution converges to the constrained SDE (58), which reduces in this case to

$$\mathrm{d}Y(t) = P_{T_Y\Gamma}\,\mathrm{d}b(t) + \frac{1}{2}\partial^2\Phi(Y(t))[I_m]\,\mathrm{d}t, \qquad \text{and} \qquad Y(t)\in\Gamma,$$

where b is an m-dimensional Brownian motion and I_m is the identity matrix. The formulation (81) follows from applying the characterization (29). ■

Remark 42 (SGLD and generalization) *Raginsky, Rakhlin, and Telgarsky (2017) show quantitative generalization bounds on SGLD, using optimal transport and Logarithmic Sobolev inequalities. Their result is meaningful in a limit of very many data points, and therefore focuses on an underparameterized setting, rather than the overparameterized setting as in this paper.*

5.1.3 'Anti-correlated perturbed gradient descent'

Orvieto et al. (2022) gave the name 'anti-correlated gradient descent' to the simple scheme used in the introduction,

$$\hat{L}(w,\eta) := L(w+\eta).$$

As discussed there, this is an example of non-degenerate noisy gradient descent, with convergence to the constrained gradient flow driven by

$$\mathrm{Reg}(w) := \frac{1}{2}\Delta_w L(w).$$

Orvieto et al. (2022) and Orvieto et al. (2023) also directly investigate the generalization properties of this type of noise injection.

5.2 Examples where η is indexed by the sample

We now consider the supervised-learning context, in which the loss L is an empirical average of a local loss function $\ell \geq 0$ over a set $\{(x_i, y_i)_{1\leq i\leq N}\}$ of 'training data points',

$$L(w) = \frac{1}{N}\sum_{i=1}^{N} \ell(w, x_i, y_i). \tag{82}$$

In this section we consider noise variables η indexed by the same indices as the data points (x_i, y_i), i.e. η is a random element of $\mathbb{R}^N$.

5.2.1 Minibatching

One of the most widely used noisy gradient descent algorithms is the one with noise induced by the random sampling of the data points, usually referred simply as *stochastic gradient descent* (SGD). In the standard setting the whole dataset is randomly split into disjoint 'minibatches' $\{B_k\}$ of the same size m, and at every iteration the gradient is calculated only for samples in one minibatch B_k:

$$w_{k+1} = w_k - \alpha\nabla_w\Big(\frac{1}{m}\sum_{(x_i,y_i)\in B_k} \ell(w, x_i, y_i)\Big).$$

We slightly modify this formulation by introducing i.i.d. random variables $\eta_{k,i}$ at every iteration, with distribution

$$\eta_{k,i} = \begin{cases} -1 & \text{w.p. } 1-\dfrac{m}{N}, \\ \dfrac{N-m}{m} & \text{w.p. } \dfrac{m}{N}, \end{cases}$$

and the corresponding noisy loss function $\hat{L}$:

$$\hat{L}(w,\eta) = \frac{1}{N}\sum_{i=1}^{N}(1+\eta_i)\ell(w,x_i,y_i).$$

One can see that indeed $\hat{L}(w,0) = L(w)$. Moreover, if we write the corresponding 'minibatch' as

$$B_k = \left\{(x_i,y_i) : \eta_{k,i} = \frac{N-m}{m}\right\},$$

then the noisy gradient descent (1) becomes

$$w_{k+1} = w_k - \alpha\nabla_w\hat{L}(w,\eta_k) = w_k - \alpha\nabla_w\Big(\frac{1}{m}\sum_{(x_i,y_i)\in B_k}\ell(w,x_i,y_i)\Big).$$

This version of SGD can be interpreted as deciding at every iteration which data point to include, independently for each data point and independently for each iteration. In contrast to the standard SGD algorithm, in such a setting the minibatch size is not fixed, and during every epoch the same data point might not occur or can occur multiple times. The parameter m also is not a deterministic minibatch size but the expectation of the minibatch size.

We show that for this modification of minibatch SGD the limiting dynamics are trivial on the time scale $1/\alpha_n^2$ for any fixed parameter m, by applying Theorem 29. Similarly, one can show that the joint limit $m_n \to N$, $\alpha_n \to 0$ also results in a trivial process by applying Theorem 23. Thus, we argue that minibatch noise affects the training dynamics around the zero-loss manifold only in a weak way, and additional forms of noise injection might be required to ensure good generalization properties on this time scale.

Corollary 43 *Let $\hat{L}, \eta_{k,i}$ be as defined above. Assume that ℓ is C^3 in w and that L satisfies Assumption 10. Let $\hat{W}_n(0) \Rightarrow W_0 \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Let $\alpha_n \to 0$ and $\sigma_n \to \sigma_0 \geq 0$. Then $\hat{W}_n$ given by* (56a) *converges in the sense of Katzenberger (see Definition 34) to the trivial process*

$$\hat{Y}(t) = \Phi(W_0) \qquad \text{for all } t \geq 0.$$

Proof

The function $\hat{L}(w,\eta)$ is of the form (55) with $f(w)_i = N^{-1}\ell(w,x_i,y_i)$, $i = 1,\dots,N$, and applying Theorem 29 we find the limiting dynamics

$$\hat{Y}(t) = \Phi(\hat{W}(0)) + \frac{1}{N}\int_0^{t\wedge\mu} P_{T_Y\Gamma}\nabla_w f(w)db_s + \frac{1}{2}\int_0^{t\wedge\mu}\partial^2\Phi(\hat{Y}(s))[\Sigma(s)]ds, \tag{83}$$

where

$$\Sigma(s) = \frac{1}{N^2}\sum_{i=1}^{N}\nabla_w\ell(\hat{Y}(s),x_i,y_i)\otimes\nabla_w\ell(\hat{Y}(s),x_i,y_i)\,.$$

As $\hat{Y}(s) \in \Gamma$ a.s., we have

$$L(\hat{Y}(s)) = \frac{1}{N}\sum_{i=1}^{N}\ell(\hat{Y}(s),x_i,y_i) = 0,$$

implying that

$$\nabla_w \ell(\hat{Y}(s), x_i, y_i) = 0. \tag{84}$$

In turn this implies that f and Σ vanish on Γ, resulting in the trivial dynamics $\hat{Y}(t) = \Phi(\hat{W}(0))$ for all t. ■

5.2.2 Label Noise

Label noise is the specific case of mean squared error L with noisy perturbation of the labels y_i, which in our setting takes the form

$$\hat{L}(w, \eta) = \frac{1}{N} \sum_{i=1}^{N} (f_w(x_i) - y_i - \eta_i)^2. \tag{85}$$

The consequences of label noise for the training iterates were studied by Blanc et al. (2020) and the already-mentioned Li, Wang, and Arora (2022a).

The function $\hat{L}$ is of the degenerate form (55), with $f(w)_i = -2(f_w(x_i) - y_i)/N$, $H(w) = 0$, and $g(\eta) = N^{-1} \sum_{i=1}^{N} \eta_i^2$. As a result, Theorem 29 provides the limiting dynamics at rate $1/\alpha_n^2 \sigma_n^2$, both for $\sigma_n \to 0$ and $\sigma_n \to \sigma_0 > 0$.

Corollary 44 (Convergence for Label Noise) *Let $\hat{L}$ be as defined in* (85) *for some family of functions f_w such that L is of class C^3 and satisfies Assumption 10. Let $\tilde{W}_n(0) \Rightarrow W_0 \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Let $\alpha_n \to 0$ and $\sigma_n \to \sigma_0 \geq 0$. Let $\eta_{k,i} \sim \rho(\sigma_n)$ i.i.d., where ρ satisfies* (9) *and* (57).

Then $\hat{W}_n$ given by (56a) *converges in U_Γ in the sense of Katzenberger to the constrained gradient flow Y given by*

$$\frac{dY}{dt} = -P_{T_Y \Gamma} \nabla_w \mathrm{Reg}(Y), \qquad Y(t) \in \Gamma, \qquad \textit{and} \qquad Y(0) = \Phi(W_0),$$

with

$$\mathrm{Reg}(w) = \frac{1}{2N} \Delta L(w). \tag{86}$$

Corollary 44 is effectively the same result as by (Li et al., 2022a, Cor. 5.2); we give the proof both for completeness and to allow us to re-use the arguments in Corollary 45.

Proof As remarked above, the function $\hat{L}$ in (85) is of the form (55) with $f(w)_i = -2(f_w(x_i) - y_i)/N$, $H(w) = 0$, and $g(\eta) = N^{-1} \sum_{i=1}^{N} \eta_i^2$. Theorem 29 then provides Katzenberger convergence to the limiting SDE (58).

To simplify this SDE, note that

$$\nabla_w f(w)_i = -\frac{2}{N} \nabla_w (f_w(x_i) - y_i - \eta_i) = -\frac{2}{N} \nabla_w f_w(x_i),$$

and for $w \in \Gamma$

$$\nabla^2 L(w) = \frac{2}{N} \nabla_w \sum_{i=1}^{N} (f_w(x_i) - y_i) \nabla_w f_w(x_i) = \frac{2}{N} \sum_{i=1}^{N} \nabla_w f_w(x_i) \nabla_w f_w(x_i)^\top.$$

It follows that $\nabla_w f(w) \in \mathrm{Range}(\nabla^2 L(w))$, and therefore $P_{T_w\Gamma}\nabla_w f(w) = 0$, implying that the noise term in (58) vanishes.

We also find that

$$\Sigma_{k\ell} := \sum_{i=1}^{N} \partial_{w_k} f(w)_i \partial_{w_\ell} f(w)_i = \frac{4}{N^2} \sum_{i=1}^{N} \partial_{w_k} f_w(x_i) \partial_{w_\ell} f_w(x_i),$$

implying that $\Sigma = 2\nabla^2 L(w)/N$. Applying (30) to the final term in (58) yields the expression (86). ∎

A minor modification of the discussion above shows that combination of minibatching and label noise results in the same limiting dynamics. Consider the following loss that combines minibatching noise variables $\tilde{\eta}$ and label noise variables η,

$$\hat{L}(w, \tilde{\eta}, \eta) = \frac{1}{N} \sum_{i=1}^{N} (1 + \tilde{\eta}_i)(f_w(x_i) - y_i - \eta_i)^2, \tag{87}$$

and for simplicity let both distributions have the same variance $\mathrm{Var}(\eta_{k,i}) = \mathrm{Var}(\tilde{\eta}_{k,j}) = \sigma^2$. The loss (87) has the degenerate form (55),

$$\hat{L}(w, \eta, \tilde{\eta}) = L(w) + f(w) \cdot \eta + \tilde{f}(w) \cdot \tilde{\eta} + H(w) : (\eta \otimes \tilde{\eta}) + g(\eta, \tilde{\eta}),$$

with

$$f(w)_i = -\frac{2}{N}(f_w(x_i) - y_i), \qquad \tilde{f}(w)_i = \frac{1}{N}(f_w(x_i) - y_i)^2,$$
$$\text{and} \quad H(w)_{ij} = -\frac{2}{N}(f_w(x_i) - y_i)\delta_{ij}.$$

From Theorem 29 we obtain the following characterization.

Corollary 45 (Convergence for combined Label Noise and Minibatching) *Let $\hat{L}$ be as defined in* (87) *for some family of functions f_w such that L is of class C^3 and satisfies Assumption 10. Let $\tilde{W}_n(0) \Rightarrow W_0 \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Let $\alpha_n \to 0$ and $\sigma_n \to \sigma_0 \geq 0$. Let $\eta_{k,i}, \tilde{\eta}_{k,i} \sim \rho(\sigma_n)$ i.i.d., where ρ satisfies* (9) *and* (57).

Then $\hat{W}_n$ given by (56a) *converges in the sense of Katzenberger to the constrained gradient flow Y given by*

$$\frac{dY}{dt} = -P_{T_Y\Gamma}\nabla_w \mathrm{Reg}(Y), \qquad Y(t) \in \Gamma, \qquad \text{and} \qquad Y(0) = \Phi(W_0),$$

with

$$\mathrm{Reg}(w) = \frac{1 + \sigma_0^2}{2N} \Delta L(w). \tag{88}$$

Proof Similarly to the cases of Corollaries 43 and 44 we have on Γ that

$$\nabla_w f(w)_i = -\frac{2}{N} \nabla_w f_w(x_i),$$
$$\nabla_w \tilde{f}(w)_i = -\frac{2}{N} \nabla_w (f_w(x_i) - y_i)^2 = 0,$$

and

$$\nabla_w H(w)_{ij} = -\frac{2}{N}\nabla_w f_w(x_i)\delta_{ij},$$

and thus by the same orthogonality argument as in Corollary 44 we conclude that the noise is orthogonal to $T\Gamma$. We apply Theorem 29 and find the resulting evolution (81), where the first integral vanishes by orthogonality and we have the expression for Σ,

$$\begin{aligned}\Sigma_{k\ell} &= \frac{4}{N^2}\sum_{i=1}^{N}\partial_{w_k} f_w(x_i)\partial_{w_\ell} f_w(x_i) + \sigma_0^2\frac{4}{N^2}\sum_{i=1}^{N}\partial_{w_k} f_w(x_i)\partial_{w_\ell} f_w(x_i)\\ &= \frac{4}{N^2}(1+\sigma_0^2)\sum_{i=1}^{N}\partial_{w_k} f_w(x_i)\partial_{w_\ell} f_w(x_i).\end{aligned}$$

Similarly to Corollary 44 it follows that $\Sigma = 2(1+\sigma_0^2)\Delta L(w)/N$, resulting in the expression (88). ■

5.3 Classical Dropout

We finally discuss the case of 'classical' Dropout as applied in neural networks and other systems. We consider the mean-square error loss for a function $f_w^{\mathrm{drop}}(x,\eta)$ that depends both on the parameter w and the noise variable η:

$$\hat{L}(w,\eta) = \frac{1}{N}\sum_{i=1}^{N}(f_w^{\mathrm{drop}}(x_i,\eta) - y_i)^2. \tag{89}$$

5.3.1 Overparameterized linear models

The first example of this type is the class of *overparameterized linear models* (Baldi and Hornik, 1989; Woodworth et al., 2020), in which $w = (u,v)$, $u,v \in \mathbb{R}^{d_{\mathrm{in}}}$, and f_w has the following form:

$$f_w(x) = \left\langle u^{\odot 2} - v^{\odot 2}, x\right\rangle, \qquad x \in \mathbb{R}^{d_{\mathrm{in}}}. \tag{90}$$

Note that the model is linear in x but non-linear in parameters w. We introduce Dropout filters $\eta \in \mathbb{R}^{d_{\mathrm{in}}}$ and a corresponding function f_w^{drop} as

$$f_w^{\mathrm{drop}}(x,\eta) = \left\langle u^{\odot 2} - v^{\odot 2}, x\odot(1+\eta)\right\rangle. \tag{91}$$

In comparison to DropConnect, the filter η_i alters the magnitude of the i^{th} feature of the input vector, rather than the corresponding feature of w. The following corollary is a direct consequence of Theorem 23.

Corollary 46 (Convergence for Dropout in overparameterized linear models) *Let f_w^{drop} be as defined in (91) and $\hat{L}$ as in (89). Assume that $L(w) := \hat{L}(w,0)$ satisfies Assumption 10. Let $\tilde{W}_n(0) \Rightarrow W_0 \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Let $\eta_{k,i}$ be Bernoulli or Gaussian dropout noisy variables. Let $\alpha_n, \sigma_n \to 0$.*

Then $\tilde{W}_n$ as defined in (34) *converges to Y in the sense of Katzenberger (see Def. 34), where Y is the constrained gradient flow defined as*

$$\frac{dY}{dt} = -P_{T_Y\Gamma}\nabla_w \mathrm{Reg}(Y), \qquad Y(t) \in \Gamma, \qquad \text{and} \qquad Y(0) = \Phi(W_0),$$

and

$$\mathrm{Reg}(w) := \frac{1}{2}\Delta_\eta \hat{L}(w,0) = \frac{1}{2N}\sum_{j=1}^{d_{\mathrm{in}}}(u_j^2 - v_j^2)^2 \sum_{i=1}^{N} x_{ij}^2, \tag{92}$$

and where x_{ij} is the j^{th} feature of the i^{th} data point x_i.

Remark 47 *The expression* (92) *coincides with a weighted L^2-regularization term in linear models—treating $u^{\odot 2} - v^{\odot 2}$ as a single linear parameter—in which the weights are chosen equal to the average amplitude of the corresponding feature $\frac{1}{N}\sum_{i=1}^N x_{ij}^2$.*

Remark 48 *Li et al. (2022a, Lemma 6.2, Lemma 6.3) provide sufficient conditions for overparameterized linear models to satisfy the manifold Assumption 10.*

Remark 49 (Two Laplacians as regularizer) *Both for Dropout and for label noise the regularizer is a Laplacian, but one is with respect to η and the other with respect to w; thus these two forms of regularization may lead to different types of behaviour. We illustrate this on the example of overparameterized linear models, where $f_w(x) = \langle u^{\odot 2} - v^{\odot 2}, x\rangle$ as in* (90).

With Dropout noise we obtain the regularizer (92). *With label noise the regularizer is given by* (86) *instead, and we now make this label-noise regularizer more explicit. On the zero-loss manifold we have $\langle u^{\odot 2} - v^{\odot 2}, x_i\rangle = y_i$ for all i and thus*

$$\begin{aligned}\frac{1}{2N}\Delta_w L &= \frac{1}{2N^2}\Delta_u \sum_{i=1}^{N}\Big(\Big\langle u^{\odot 2} - v^{\odot 2}, x_i\Big\rangle - y_i\Big)^2 + \frac{1}{2N^2}\Delta_v \sum_{i=1}^{N}\Big(\Big\langle u^{\odot 2} - v^{\odot 2}, x_i\Big\rangle - y_i\Big)^2 \\ &= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{d_{\mathrm{in}}} 8u_j^2 x_{ij}^2 + \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{d_{\mathrm{in}}} 8v_j^2 x_{ij}^2 = \frac{4}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{d_{\mathrm{in}}}(u_j^2 + v_j^2)x_{ij}^2. \end{aligned} \tag{93}$$

Note how the Dropout regularizer (92) *regularizes the difference $(u_j^2 - v_j^2)^2$, and the label noise regularizer above penalizes both u and v separately. Also note how the two regularizers differ in their scaling in N, with label noise leading to an additional slow-down factor $1/N$.*

5.3.2 Feedforward neural networks

We say that $f_w : \mathbb{R}^{d_{\mathrm{in}}} \to \mathbb{R}$ is a p-layered feedforward neural network if it is a composition of p blocks in which each block consists of a linear layer and a pointwise nonlinearity. Each block is a mapping $\phi^k : y^k \to y^{k+1}$ ($\mathbb{R}^{d_k} \to \mathbb{R}^{d_{k+1}}$) of the form:

$$\begin{aligned} z^{k+1} &= W^{k+1}y^k + b^{k+1}, \\ y^{k+1} &= s(z^{k+1}), \end{aligned}$$

where $W^{k+1} \in \mathbb{R}^{d_{k+1}\times d_k}$, $b^{k+1} \in \mathbb{R}^{d_{k+1}}$, the input dimension is $d_0 = d_{\mathrm{in}}$ and the output dimension $d_p = 1$. The map $f_w(x)$ then takes the form:

$$f_w(x) = \phi^p(\phi^{p-1}(\cdots \phi^1(x))).$$

We introduce dropout filters η^k that perturb the input features at the k-th layer. The resulting maps $\phi_d^k : (\eta^k, y^k) \to y^{k+1}$ ($\mathbb{R}^{2d_k} \to \mathbb{R}^{d_{k+1}}$) then are

$$\begin{aligned} z^{k+1} &= W^{k+1}(y^k \odot (1+\eta^k)) + b^{k+1}, \\ y^{k+1} &= s(z^{k+1}), \end{aligned}$$

and the corresponding f_w^{drop} is given by

$$f_w^{\text{drop}}(x,\eta) = \phi_d^p(\eta^p, \phi_d^{p-1}(\eta^{p-1} \cdots \phi_d^1(\eta^1, x))), \tag{94}$$

Here we distinguish the same two choices as in Section 5.1, namely Bernoulli and Gaussian filters.

In the Bernoulli case $\eta_{k,i}$ are i.i.d. Bernoulli random variables for some $0 < p < 1$:

$$\eta_{k,i} = \begin{cases} -1, & \text{w.p. } p \\ \frac{p}{1-p} & \text{w.p. } 1-p. \end{cases}$$

In the Gaussian case $\eta_{k,i}$ are i.i.d. normal random variables,

$$\eta_{k,i} \sim \mathcal{N}(0, \sigma^2).$$

With Gaussian filters, no input is ever ignored, as $\mathbb{P}(\eta_{k,i} = -1) = 0$. Note that Gaussian noise also allows to change sign of some inputs.

First consider feedforward neural networks with a single hidden layer:

$$f_w^{\text{drop}}(x,\eta) = \sum_{j=1}^{n} a_j(1+\eta_j)s(b_j^\top x), \tag{95}$$

where dropout is only applied to the output layer. As an example we consider the *smooth rectified linear unit* as activation function,

$$s(x) = \begin{cases} 0, & x \le 0 \\ xe^{-1/x}, & x > 0. \end{cases} \quad . \tag{96}$$

Corollary 50 *Let f_w^{drop} be as defined in* (95) *with rectified smooth activation function* (96)*. Assume that $L(w) := \hat{L}(w,0)$ satisfies Assumption 10. Let $\tilde{W}_n(0) \Rightarrow \tilde{W}(0) \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Let $\eta_{k,i}$ be Bernoulli or Gaussian Dropout noisy variables as defined above. Let $\alpha_n, \sigma_n \to 0$.*

Then $\tilde{W}_n$ as defined in (34) *converges to Y in the sense of Katzenberger, where Y is the constrained gradient flow defined as*

$$\frac{dY}{dt} = -P_{T_Y\Gamma}\nabla_w \text{Reg}(Y), \qquad Y(t) \in \Gamma, \qquad \textit{and} \qquad Y(0) = \Phi(W_0), \tag{97}$$

with

$$\text{Reg}(w) = \frac{1}{2}\Delta_\eta \hat{L}(w,0) = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{N} a_j^2\, s(b_j^\top x_i)^2. \tag{98}$$

Remark 51 (Known properties of the zero-loss manifold) *Cooper (2018) proves for overparameterized shallow neural networks with the rectified smooth activation function* (96) *that the zero-loss set of $L(w) := \hat{L}(w, 0)$ is a smooth manifold and L satisfies Assumption 10. The result also holds for deep feedforward neural networks if the size of the last layer is greater than the size of the training dataset, i.e. $d_k > N$.*

Similar results are known for the ReLU *activation function* $\mathrm{ReLU}(y) = \max(0, y)$ *(Petzka et al., 2020; Dereich and Kassing, 2022). Even though the* ReLU *is not differentiable in $y = 0$, the manifold is locally smooth away from the hyperplanes $Q_i = \{x : w_i^\top x = 0\}$, so that the analysis holds for every set K satisfying $K \cap Q_i = \emptyset$ for every Q_i. Boursier et al. (2022) study the structure of the neighbourhood U and propose an initialization scheme guaranteeing convergence of the gradient flow for shallow* ReLU *networks.*

Proof First note that the smoothness of s guarantees the regularity of $\hat{L}$, and since s has sublinear growth, $\hat{L}$ has at most growth of order 6 in (w, η).

Next, note that we can write $\hat{L}$ in the form

$$\hat{L}(w, \eta) = L(w) + f(w) \cdot \eta + \frac{1}{2} H(w) : (\eta \otimes \eta),$$

with

$$f(w)_j = \frac{2}{N} a_j \sum_{i=1}^{N} (f_w(x_i) - y_i)\, s(b_j^\top x_i) \quad \text{and} \quad H_{jj'} = \frac{2}{N} \sum_{i=1}^{N} a_j s(b_j^\top x_i) a_{j'} s(b_{j'}^\top x_i).$$

This structure is similar to (55), which we require for the degenerate case, but note that for the degenerate case we require $\operatorname{tr} H = 0$, while here $\operatorname{tr} H$ does not vanish. Consequently the evolution takes place at the time scale $1/\alpha_n \sigma_n^2$. This function $\hat{L}$ satisfies conditions (10) and (11) with $C_2 = 0$.

We now turn to condition (12). Gaussian noise satisfies this condition as mentioned in Remark 8, and for the q-th absolute moment of Bernoulli random variables, where $q \geq 1$, we have

$$\mathrm{M}_q = 1 \cdot p + \frac{p^q}{(1-p)^q} \cdot (1-p) = \frac{p^q + p \cdot (1-p)^{q-1}}{(1-p)^{q-1}} = O(p) = O(\sigma^2) \qquad \text{as } p \to 0.$$

Since $C_2 = 0$, condition (12) also is satisfied. The condition (35) in Theorem 23 is satisfied for Bernoulli filters by construction and for Gaussian filters by Lemma 55.

Theorem 23 then yields Katzenberger convergence to the constrained gradient flow (97), with $\mathrm{Reg}(w) = \frac{1}{2} \Delta_\eta \hat{L}(w, 0) = \frac{1}{2} \operatorname{tr} H(w)$ as given in (98). ∎

A very similar result holds for the case of deep networks in which Dropout is only applied to the last layer, as described in Example 2:

Corollary 52 *Let f_w^{drop} be as defined in* (2) *with smooth earlier layers ϕ_j with at most polynomial growth. Assume that $L(w) := \hat{L}(w, 0)$ satisfies Assumption 10. Let $\tilde{W}_n(0) \Rightarrow \tilde{W}(0) \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Let $\eta_{k,i}$ be Bernoulli or Gaussian Dropout noisy variables as defined above. Let $\alpha_n, \sigma_n \to 0$.*

Then $\tilde{W}_n$ as defined in (34) *converges to Y in the sense of Katzenberger, where Y is the constrained gradient flow defined as*

$$\frac{dY}{dt} = -P_{T_Y\Gamma}\nabla_w \mathrm{Reg}(Y), \qquad Y(t) \in \Gamma, \qquad \text{and} \qquad Y(0) = \Phi(W_0), \tag{99}$$

with

$$\mathrm{Reg}(w) = \frac{1}{2}\Delta_\eta \hat{L}(w,0) = \frac{1}{N}\sum_{j=1}^{n}\sum_{i=1}^{N} a_j^2\,\phi_j(x_i;w)^2. \tag{100}$$

The proof is practically identical to that of Corollary 50, since very few of the properties of the earlier layers are used, and we omit it.

For deep neural networks with Dropout applied to more layers than just the final one, Theorem 23 implies a convergence result for the case of Gaussian filters. Due to the complexity of Dropout in deeper layers we do not provide an explicit form of the regularizer in this case.

Corollary 53 (Convergence for Dropout noise in deep neural networks) *Let f_w^{drop} be as defined in* (94) *for some $p \geq 2$ with rectified smooth activation function s* (96)*. Assume that $L(w) := \hat{L}(w,0)$ satisfies Assumption 10. Let $\tilde{W}_n(0) \Rightarrow \tilde{W}(0) \in U_\Gamma$ for some locally attractive neighbourhood U_Γ of Γ. Let $\eta_{k,i}$ be Gaussian filter variables as defined above. Let $\alpha_n, \sigma_n \to 0$.*

Then W_n as defined in (34) *converges to Y in the sense of Katzenberger, where Y is the constrained gradient flow defined as*

$$\frac{dY}{dt} = -\frac{1}{2}P_{T_Y\Gamma}\nabla_w \Delta_\eta \hat{L}(Y,0), \qquad Y(t) \in \Gamma, \qquad \text{and} \qquad Y(0) = \Phi(W_0).$$

Proof The proof is analogous to the proof of Corollary 50 with the only difference caused by the composition of several layers. Using the global boundedness of the derivatives of activation function $|s'(x)|, |s''(x)| < C$ for all $x \in \mathbb{R}$ one can see that for any fixed $w \in K$ the expression $|\nabla_w \nabla_\eta^2 \hat{L}(w,\eta_1) - \nabla_w \nabla_\eta^2 \hat{L}(w,\eta_2)|$ scales at most polynomially in $\eta_{1,2}$ and thus the assumptions of Theorem 23 are satisfied. ■

Remark 54 (Combining Dropout with Minibatching) *It is easy to see that the combination of Dropout noise and minibatching results in the same dynamics as Dropout gradient descent without minibatching. Consider the loss*

$$\hat{L}(w,\tilde{\eta},\eta) = \frac{1}{N}\sum_{i=1}^{N}(1+\tilde{\eta}_i)\hat{\ell}(w,\eta,x_i,y_i).$$

As $\hat{L}$ is linear in the minibatching noise variables $\tilde{\eta}$, we have $\Delta_{\tilde{\eta}}\hat{L}(w,\tilde{\eta},\eta) = 0$ and thus the regularizer takes the same form as the regularizer of the corresponding dropout gradient descent $\frac{1}{2}\Delta_{\tilde{\eta},\eta}\hat{L}(w,\tilde{\eta},\eta) = \frac{1}{2}\Delta_\eta \hat{L}(w,\tilde{\eta},\eta)$. The same holds for the DropConnect case.

6. Discussion and outlook

In this paper we define a class of noisy gradient-descent systems and prove their convergence in the limit of small learning rate and in some cases also small noise. This class of systems unifies a broad collection of existing training algorithms in a common structure, and the convergence theorems thus give a more global understanding of the effect of noise in various overparameterized training situations.

In this section we add some more remarks and discuss various generalizations.

6.1 Constant learning rates

It is common to change the learning rate α during training, for instance to generate phases of larger and smaller noise in minibatch SGD. For some such non-constant learning rate algorithms the results of this paper should continue to hold. The essential properties that need to be verified are Assumptions 16, 17, and 18, which are all formulated in terms of the integrator sequences A_n and $\hat{A}_n$ and therefore are also meaningful for non-constant learning rates.

6.2 Correlated noise

Similarly, we have chosen to make the coordinates $\eta_{k,i}$ of each iterate η_k independent from each other, but this again only for simplicity of formulation.

To give an example of a generalization, consider again the example of Figures 1 and 2 in the introduction. Figure 4 compares the effect of uncorrelated (left) and correlated noise. Here we choose as uncorrelated noise

$$\eta_{k,i} \sim \mathcal{N}(0, \sigma^2 I_2)$$

and as correlated noise

$$\eta_k \sim \mathcal{N}(0, C), \qquad C = \frac{\sigma^2}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

The corresponding regularizers are

$$\begin{aligned} \text{uncorrelated: } & \frac{1}{2}\Delta_w L(w) \\ \text{correlated: } & \frac{1}{2}\nabla_w^2 L(w) : C = \frac{1}{4}(\partial_{11}L + 2\partial_{12}L + \partial_{22}L), \end{aligned}$$

and the minimizers of these two functions are indicated by green circles.

6.3 Constrained gradient flow and regularization

In this paper we have used the term 'regularizer' and the notation 'Reg' for the function that drives the limiting constrained gradient flow (4). This terminology is inspired by its relationship with Tychonov regularization of inverse problems.

To explain this, note that solutions of the constrained gradient flow tend to converge to local minimizers of Reg, or if one is 'lucky', even to a global minimizer, i.e. a solution of the constrained problem

$$\min_w \{\mathrm{Reg}(w) : w \in \Gamma\}.$$

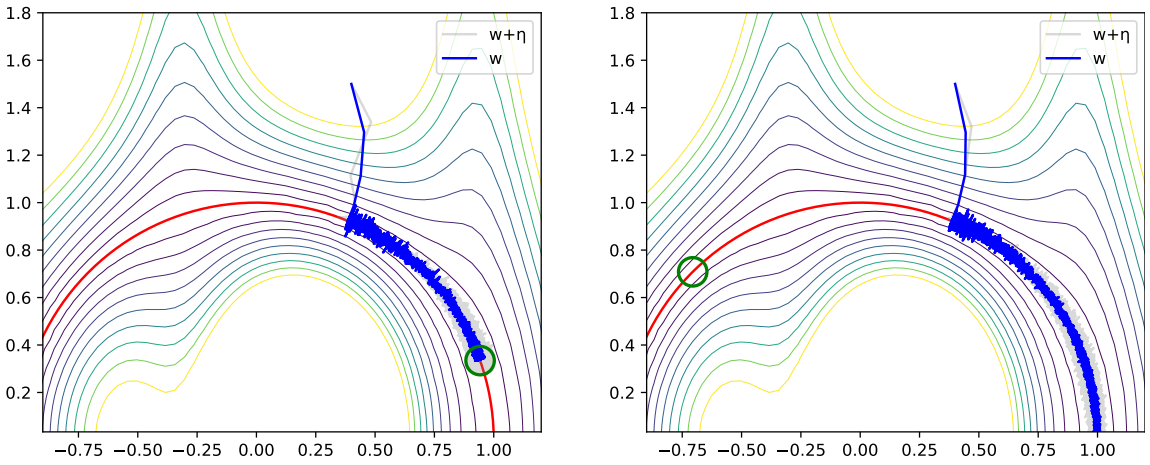


Figure 4: When the coordinates of the vector η_k are correlated, the resulting regularizer Reg is modified, and the evolution follows a different path along Γ. In both diagrams each η_k is a centered normal two-dimensional random variable with covariance C, independent for each k; in the left-hand picture $C = \sigma^2 I_2$, implying independence of $\eta_{k,1}$ from $\eta_{k,2}$, while in the right-hand picture $C = \frac{1}{2}\sigma^2 \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, implying that $\eta_{k,1}$ and $\eta_{k,2}$ are fully correlated.

This situation is reminiscent of regularized inverse problems of the form

$$\min_u \|Tu - f\|^2 + \lambda R(u),$$

for which in the 'weak-regularization' limit $\lambda \to 0$ the minimizers u_λ converge to the solution of the constrained minimization problem

$$\min_u \{R(u) : Tu = f\}.$$

This is why we call Reg, the driving functional in the constrained gradient flow, the (implicit) regulariser of the noisy gradient descent.

Theorem A opens the door to a form of reverse engineering. Given a loss L, the choice of $\hat{L}$ is only limited by the consistency $\hat{L}(w, 0) = L(w)$, and one therefore has a wide freedom to tailor $\hat{L}$ to have particular properties. Assuming that one has an understanding of what 'good' and 'bad' points $w \in \Gamma$ look like, Theorem A suggests to look for functions $\hat{L}$ such that $\Delta_\eta \hat{L}$ gives high value to 'bad' points and low value to 'good' ones.

Some pointers to how 'good' and 'bad' points can be recognized or characterized are given by the 'robustness' criterion by Petzka et al. (2020) (see Remark 37) or the discussion by Orvieto et al. (2022) of the connection with PAC-Bayes bounds. We leave this aspect to future work.

6.4 Convergence results in Skorokhod spaces and Katzenberger's theorem

The first (to our knowledge) application of Katzenberger's theorem to machine learning models was by Li et al. (2022a). In that paper the dynamics the noise is assumed to be of minibatch-type, namely

$$dW_n = -\nabla L(W_n) d\hat{A}_n + \sqrt{\Sigma}\sigma_{\eta_k}(W_n),$$

where η_k is sampled uniformly from the set $B = \{1, 2, \dots, N\}$, and for every $i \in B$, $\sigma_i(W_n)$ is a deterministic function and $\mathbb{E}\sigma_{\eta_k}(W) = 0$. Note that by definition the noise process in such a setting is a martingale. Our definition of noisy gradient-descent systems generalizes this, and allows us to study for instance the effect of dropout noise.

The result by Li et al. (2022a) has also been generalized to SGD with momentum. Cowsik et al. (2022) study the interplay between the momentum parameter and the noise distribution. The structure of the noise is the same as in the work by Li et al. (2022a), and one direction of future work is the study of SGD with momentum with general noise.

Another type of convergence results in the sense of processes are so-called mean-field limit results. In contrast to this work, mean-field convergence describes the behaviour of the models with variable number of parameters. For example, for shallow neural networks it studies the limiting training dynamic of models

$$f_n(W_n, x) = \frac{1}{n}\sum_{i=1}^{n} a_i \sigma(b_i^\top x + c_i),$$

where $w_i = (a_i, b_i, c_i)$ and $W = (w_1, \dots, w_n)$. It turns out that under suitable assumptions for μ_n defined as

$$\mu_n(s) = \frac{1}{n}\sum_{i=1}^{n} \delta_{w_i(s)},$$

it holds that $\mu_n \Rightarrow \mu$ in Skorokhod topology on $[0, T]$, where μ is a solution of a measure-valued evolution equations characterized by the loss function (Sirignano and Spiliopoulos, 2020; Rotskoff and Vanden-Eijnden, 2022). Similar results have been derived for deep neural networks (Sirignano and Spiliopoulos, 2022; Nguyen, 2019). Note that the mean-field setting does not involve rescaling of time, implying that only the main (the fast) time scale is considered. Another direction of future work is to study the slow time scale dynamics in the measure-valued setting. Methods such as those developed by Duong et al. (2017, 2018) could be useful for this.

Acknowledgments

Mark Peletier and André Schlichting gratefully acknowledge several interesting discussions with Anton Arnold, Vlado Menkowski, Gijs Peletier, and Frank Redig. The work was done while Anna Shalova was at Eindhoven University of Technology. Mark Peletier and Anna Shalova are supported by the Dutch Research Council (NWO), in the framework of the program 'Unraveling Neural Networks with Structure-Preserving Computing' (file number OCENW.GROOT.2019.044). André Schlichting is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2044 –390685587, Mathematics Münster: Dynamics–Geometry–Structure and his research is partially based upon work from COST Action 24122 mSPACE, supported by COST (European Cooperation in Science and Technology), www.cost.eu.

Appendix A. Details of numerical simulations

The function $L : \mathbb{R}^2 \to \mathbb{R}$ depicted in Figures 1, 2, and 3 is

$$L(w) = \frac{(|w|^2-1)^2}{(|w|^2+1)^2}(1 + a\sin(bw_1)), \qquad a = 0.7,\ b = 5.$$

The learning rate in those figures is $\alpha = 0.3$, and η has a centered normal distribution with covariance $\sigma^2 I_2$ with $\sigma = 0.03$.

The constrained gradient flow in Figures 2 and 3 is implemented by parameterizing by Γ by polar angle θ and writing the regularizer as a function of θ.

Appendix B. Auxiliary results

The following lemma shows that for i.i.d. Gaussian noise variables η the two conditions (35) and (57) follow from the assumption $\alpha_n \to 0$.

Lemma 55 (Gaussian filters satisfy the noise-decay condition)
Let α_n and σ_n be positive sequences such that $\alpha_n \to 0$ and $\sigma_n \to \sigma_0 \geq 0$. Let for each n, Y_k^n, $k \in \mathbb{N}$, be i.i.d. centered Gaussian random variables with variance σ_n^2. Then for any $T \in \mathbb{R}_+$, for any $p \geq 1$ the following convergence holds in probability and in distribution:

$$\sup_{k \leq \frac{T}{\alpha_n^2 \sigma_n^2}} \alpha_n |Y_k^n|^p \longrightarrow 0 \qquad \text{as } n \to \infty$$

Proof

$$\begin{aligned}
\mathbb{P}\left(\sup_{k \leq \frac{T}{\alpha_n^2\sigma_n^2}} \alpha_n |Y_k^n|^p > \varepsilon \right) &\leq \sum_{k \leq \frac{T}{\alpha_n^2\sigma_n^2}} \mathbb{P}\left(|Y_k^n|^p > \frac{\varepsilon}{\alpha_n} \right) \leq \frac{T}{\alpha_n^2\sigma_n^2} \mathbb{P}\left(|Y_1^n| > \frac{\varepsilon^{1/p}}{\alpha_n^{1/p}} \right) \\
&= \frac{T}{\alpha_n^2\sigma_n^2} \operatorname{erfc}\left(\frac{\varepsilon^{1/p}}{\sigma_n \alpha_n^{1/p}\sqrt{2}} \right) \\
&\leq \frac{\alpha_n^{-2+1/p}\, T\, \sqrt{2}}{\varepsilon^{1/p}\sigma_n\sqrt{\pi}} \exp\left(-\frac{\varepsilon^{2/p}}{2\sigma_n^2 \alpha_n^{2/p}} \right) \\
&= \frac{\beta_n^{-p+1/2}\, T\, \sqrt{2}}{\varepsilon^{1/p}\sigma_n\sqrt{\pi}} \exp\left(-\frac{\varepsilon^{2/p}}{2\sigma_n^2 \beta_n} \right), \qquad \text{with } \beta_n = \alpha_n^{2/p},
\end{aligned}$$

and for every fixed $\varepsilon > 0$ this vanishes as $\beta_n \to 0$. ∎

References

M. Andriushchenko, A. Varre, L. Pillaud-Vivien, and N. Flammarion. SGD with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925, 2023.

S. Arora, Z. Li, and A. Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

P. Baldi and P. J. Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26, 2013.

P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., 1968.

G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In *Conference on learning theory*, pages 483–513. PMLR, 2020.

E. Boursier, L. Pillaud-Vivien, and N. Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pages 20105–20118, 2022.

A. Calzolari and F. Marchetti. Limit motion of an Ornstein–Uhlenbeck particle on the equilibrium manifold of a force field. *Journal of Applied Probability*, 34(4):924–938, 1997.

P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

G. Clara, S. Langer, and J. Schmidt-Hieber. Dropout regularization versus ℓ_2-penalization in the linear model. *Journal of Machine Learning Research*, 25(204):1–48, 2024.

Y. Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

A. Cowsik, T. Can, and P. Glorioso. Flatter, faster: scaling momentum for optimal speedup of sgd. *arXiv preprint arXiv:2210.16400*, 2022.

A. Damian, T. Ma, and J. D. Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, volume 34, pages 27449–27461, 2021.

S. Dereich and S. Kassing. On minimal representations of shallow ReLU networks. *Neural Networks*, 148:121–128, 2022.

S. Dereich and S. Kassing. Convergence of stochastic gradient descent schemes for Lojasiewicz-landscapes. *Journal of Machine Learning*, 3(3):245–281, 2024.

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

M. H. Duong, A. Lamacz, M. A. Peletier, and U. Sharma. Variational approach to coarse-graining of generalized gradient flows. *Calc. Var. Partial Differ. Equ.*, 56(4):65, 2017.

M. H. Duong, A. Lamacz, M. A. Peletier, A. Schlichting, and U. Sharma. Quantification of coarse-graining error in Langevin and overdamped Langevin dynamics. *Nonlinearity*, 31 (10):4517–4566, 2018.

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

I. Fatkullin, G. Kovacic, and E. Vanden-Eijnden. Reduced dynamics of stochastically perturbed gradient flows. *Comm. Math. Sci.*, 2010.

B. Fehrman, B. Gess, and A. Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21 (136):1–48, 2020.

B. Frénay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

T. Funaki. The scaling limit for a stochastic PDE and the separation of phases. *Probability Theory and Related Fields*, 102(2):221–288, 1995.

T. Funaki and H. Nagai. Degenerative convergence of diffusion process toward a submanifold by strong drift. *Stochastics: An International Journal of Probability and Stochastic Processes*, 44(1-2):1–25, 1993.

S. B. Gelfand and S. K. Mitter. Simulated annealing type algorithms for multivariate optimization. *Algorithmica*, 6(1):419–436, June 1991.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

X. Gu, K. Lyu, L. Huang, and S. Arora. Why (and when) does Local SGD generalize better than SGD? In *International Conference on Learning Representations*, 2023.

G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

E. P. Hsu. *Stochastic Analysis on Manifolds*. Number 38. American Mathematical Soc., 2002.

S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Finding flatter minima with SGD. In *International Conference on Learning Representations - Workshop track*, 2018.

Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2019.

G. Jin, X. Yi, P. Yang, L. Zhang, S. Schewe, and X. Huang. Weight expansion: A new perspective on dropout and generalization. *arXiv preprint arXiv:2201.09209*, 2022.

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

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations.*, 2017.

T. G. Kurtz and P. Protter. Weak limit theorems for stochastic integrals and stochastic differential equations. *The Annals of Probability*, pages 1035–1070, 1991.

Z. Li, T. Wang, and S. Arora. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022a.

Z. Li, T. Wang, and D. Yu. Fast mixing of Stochastic Gradient Descent with normalization and weight decay. *Advances in Neural Information Processing Systems*, 35:9233–9248, 2022b.

O. A. Manita, M. A. Peletier, J. W. Portegies, J. Sanders, and A. Senen-Cerda. Universal approximation in dropout neural networks. *Journal of Machine Learning Research*, 23 (19):1–46, 2022.

S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

P. Mianjy and R. Arora. On dropout and nuclear norm regularization. In *International conference on machine learning*, pages 4575–4584. PMLR, 2019.

P. Mianjy and R. Arora. On convergence and generalization of dropout training. *Advances in Neural Information Processing Systems*, 33:21151–21161, 2020.

P. Mianjy, R. Arora, and R. Vidal. On the implicit bias of dropout. In *International conference on machine learning*, pages 3540–3548. PMLR, 2018.

D. Molchanov, A. Ashukha, and D. Vetrov. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*, pages 2498–2507. PMLR, 2017.

P.-M. Nguyen. Mean field limit of the learning dynamics of multilayer neural networks. *arXiv preprint arXiv:1902.02880*, 2019.

A. Orvieto, H. Kersting, F. Proske, F. Bach, and A. Lucchi. Anticorrelated noise injection for improved generalization. In *International Conference on Machine Learning*, pages 17094–17116, 2022.

A. Orvieto, A. Raj, H. Kersting, and F. Bach. Explicit regularization in overparametrized models via noise injection. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 7265–7287. PMLR, 25–27 Apr 2023.

T. Parsons. *Asymptotic analysis of some stochastic models from population dynamics and population genetics*. PhD thesis, University of Toronto, 2012.

T. L. Parsons and T. Rogers. Dimension reduction via timescale separation in stochastic dynamical systems. *Arxiv preprint arXiv:01510.07031*, 2015.

S. Pesme, L. Pillaud-Vivien, and N. Flammarion. Implicit bias of SGD for diagonal linear networks: A provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

H. Petzka, M. Trimmel, and C. Sminchisescu. Notes on the symmetries of 2-layer relu-networks. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 1, pages 6–6, 2020.

H. Petzka, M. Kamp, L. Adilova, C. Sminchisescu, and M. Boley. Relative flatness and generalization. *Advances in neural information processing systems*, 34:18420–18432, 2021.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

G. M. Rotskoff and E. Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

A. Senen-Cerda and J. Sanders. Asymptotic convergence rate of dropout on shallow linear neural networks. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(2):1–53, 2022.

A. Senen-Cerda and J. Sanders. Almost sure convergence of dropout algorithms for neural networks. *Journal of Machine Learning Research*, 26(283):1–53, 2025.

J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

J. Sirignano and K. Spiliopoulos. Mean field analysis of deep neural networks. *Mathematics of Operations Research*, 47(1):120–152, 2022.

S. Smith, E. Elsen, and S. De. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, pages 9058–9067. PMLR, 2020.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

V. B. Tadić. Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. *Stochastic Processes and their Applications*, 125(5): 1715–1755, 2015.

S. Wager, S. Wang, and P. S. Liang. Dropout training as adaptive regularization. *Advances in neural information processing systems*, 26, 2013.

L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

S. Wang and C. Manning. Fast dropout training. In *international conference on machine learning*, pages 118–126. PMLR, 2013.

C. Wei, S. Kakade, and T. Ma. The implicit and explicit regularization effects of dropout. In *International conference on machine learning*, pages 10181–10192. PMLR, 2020.

S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type part I: Discrete time analysis. *Journal of Nonlinear Science*, 33(3):45, 2023.

S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type part II: Continuous time analysis. *Journal of Nonlinear Science*, 34(1):16, 2024.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.

L. Wu, C. Ma, and W. E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

L. Wu, M. Wang, and W. Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35: 4680–4693, 2022.

C. Zhang, Q. Liao, A. Rakhlin, B. Miranda, N. Golowich, and T. Poggio. Theory of deep learning IIb: Optimization properties of SGD. *arXiv preprint arXiv:1801.02254*, 2018.

Z. Zhang and Z.-Q. J. Xu. Implicit regularization of dropout. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.