STOCHASTIC MODIFIED EQUATIONS AND DYNAMICS OF DROPOUT ALGORITHM

Zhongwang Zhang[1], **Yuqing Li**[1,2] *, **Tao Luo**[1,2,3,4,5] †, **Zhi-Qin John Xu**[1,3,4,6] ‡
[1] School of Mathematical Sciences, Shanghai Jiao Tong University
[2] CMA-Shanghai, Shanghai Jiao Tong University
[3] Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University
[4] Qing Yuan Research Institute, Shanghai Jiao Tong University
[5] Shanghai Artificial Intelligence Laboratory
[6] Shanghai Seres Information Technology Company, Ltd

ABSTRACT

Dropout is a widely utilized regularization technique in the training of neural networks, nevertheless, its underlying mechanism and impact on achieving good generalization abilities remain to be further understood. In this work, we start by undertaking a rigorous theoretical derivation of the stochastic modified equations, with the primary aim of providing an effective approximation for the discrete iterative process of dropout. Meanwhile, we experimentally verify SDE's ability to approximate dropout under a wider range of settings. Subsequently, we empirically delve into the intricate mechanisms by which dropout facilitates the identification of flatter minima. This exploration is conducted through intuitive approximations, exploiting the structural analogies inherent in the Hessian of loss landscape and the covariance of dropout. Our empirical findings substantiate the ubiquitous presence of the Hessian-variance alignment relation throughout the training process of dropout.

1 INTRODUCTION

Dropout is a technique integrated into gradient-based algorithms for training neural networks (NNs) (Hinton et al., 2012; Srivastava et al., 2014). It constitutes a pivotal component contributing to the attainment of state-of-the-art test performance in deep learning (Tan and Le, 2019; Helmbold and Long, 2015). The key idea behind dropout is to randomly deactivate a subset of neurons during the training process. Specifically, the output of each neuron is multiplied by a random variable that takes the value $1/p$ with probability p and zero otherwise. This random variable is independently sampled at each feedforward operation. Despite its widespread adoption and empirical success, the mechanism by which dropout enhances generalization in deep learning remains an ongoing area of research.

The noise structure introduced by stochastic algorithms plays a crucial role in understanding their training behaviors. A series of recent works reveal that the noise structure inherent in stochastic gradient descent (SGD) is vital for exploring flatter solutions (Keskar et al., 2016; Feng and Tu, 2021; Zhu et al., 2018). Analogously, the dropout algorithm introduces a specific form of noise, acting as an implicit regularizer that facilitates improved generalization abilities (Hinton et al., 2012; Srivastava et al., 2014; Wei et al., 2020; Zhang and Xu, 2022; Zhu et al., 2018).

In this paper, we first employ the stochastic modified equations (SMEs) (Li et al., 2017) framework to analyze the dynamics of the dropout algorithm applied to two-layer NNs. By application of SMEs, we embark on an exhaustive quantification of the leading order dynamics governing dropout, and we fortify this analytical approach through some empirical validations. In addition, we calculate the covariance matrix associated with the noise introduced by dropout. Hence our analytical exploration

*Corresponding author: liyuqing_551@sjtu.edu.cn
†Corresponding author: luotao41@sjtu.edu.cn
‡Corresponding author: xuzhiqin@sjtu.edu.cn

is further enriched by an investigation of the alignment relation between this covariance matrix and the Hessian matrix, a relationship conceptually framed as the Hessian-variance alignment relation (Zhu et al., 2018; Wu et al., 2022). We emphasize that this alignment property occupies a central role in sculpting the flatness attributes inherent in the solutions favored by NN models, and it has been firmly established that flatter solutions tend to exhibit enhanced generalization capabilities (Keskar et al., 2016; Neyshabur et al., 2017).

2 Related works

A flurry of recent works aims to shed light on the regularization effect conferred by dropout. Wager et al. (2013) show that dropout performs a form of adaptive regularization in the context of linear regression and logistic problems. McAllester (2013) propose a PAC-Bayesian bound, whereas Wan et al. (2013); Mou et al. (2018) derive some Rademacher-complexity-type error bounds specifically tailored for dropout. Cavazza et al. (2018); Mianjy and Arora (2020); Wei et al. (2020); Arora et al. (2021) demonstrate that dropout regularizes the inductive bias under different settings. Jin et al. (2022) try to explain the generalization ability of dropout from the new perspective of weight expansion. Finally, Zhang and Xu (2022) establish that dropout facilitates condensation (Luo et al., 2021; Zhou et al., 2021; 2022) through an additional regularization term endowed by dropout.

Continuous formulations have been extensively utilized to study the dynamical behavior of stochastic algorithms. Li et al. (2017; 2019) present an entirely rigorous and self-contained mathematical formulation of the SME framework that applies to a wide class of stochastic algorithms. Furthermore, Feng et al. (2017) adopt a semigroup approach to investigate the dynamics of SGD and online PCA. Malladi et al. (2022) derive the SME approximations for the adaptive stochastic algorithms including RMSprop and Adam, additionally, they provide efficient experimental verification of the validity of square root scaling rules arising from the SMEs.

One noteworthy observation is the association between the flatness of minima and improved generalization ability (Li et al., 2017; Jastrzebski et al., 2017; 2018). Specifically, SGD is shown to preferentially select flat minima, especially under conditions of large learning rates and small batch sizes (Jastrzebski et al., 2017; 2018; Wu et al., 2018). Papyan (2018; 2019) attribute such enhancement of flatness by SGD to the similarity between covariance of the noise and Hessian of the loss function. Furthermore, Zhu et al. (2018); Wu et al. (2022) unveil the Hessian-variance alignment property of SGD noise, shedding light on the role of SGD in escaping from sharper minima and locating flatter minima.

3 Preliminary

In this section, we present the notations and definitions utilized in our theoretical analysis. *We remark that our experimental settings are more general than the counterparts in the theoretical analysis.*

3.1 Notations

We set a special vector $(1, 1, 1, \dots, 1)^\intercal$ by $\mathbf{1} := (1, 1, 1, \dots, 1)^\intercal$ whose dimension varies. We set n for the number of input samples and m for the width of the NN. We let $[n] = \{1, 2, \dots, n\}$. We denote $\otimes$ as the Kronecker tensor product, and $\langle \cdot, \cdot \rangle$ for standard inner product between two vectors. We denote vector L^2 norm as $\|\cdot\|_2$, vector or function L_∞ norm as $\|\cdot\|_\infty$. We also denote $\mathrm{Tr}(\cdot)$ as the trace of a square matrix, $\boldsymbol{I}_d$ as the identity matrix of size $d \times d$, and $\|\cdot\|_{\mathrm{F}}$ signifies the Frobenius norm of a matrix. Finally, we denote the set of continuous functions $f(\cdot) : \mathbb{R}^D \to \mathbb{R}$ possessing continuous derivatives of order up to and including r by $\mathcal{C}^r(\mathbb{R}^D)$, the space of bounded measurable functions by $\mathcal{B}_b(\mathbb{R}^D)$, and the space of bounded continuous functions by $\mathcal{C}_b(\mathbb{R}^D)$.

3.2 Two-layer neural networks and loss function

We consider the empirical risk minimization problem given by the quadratic loss:

$$\min_{\boldsymbol{\theta}} R_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i \right)^2, \tag{1}$$

where $\mathcal{S} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is the training sample, $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the prediction function, $\boldsymbol{\theta}$ are the parameters, and their dependence is modeled by a two-layer NN with m hidden neurons

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) := \sum_{r=1}^{m} a_r \sigma(\boldsymbol{w}_r^\intercal \boldsymbol{x}), \tag{2}$$

where $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\theta}_a, \boldsymbol{\theta}_{\boldsymbol{w}}) \in \mathbb{R}^D$, where $D := m(d+1)$ throughout this paper. We remark that $\boldsymbol{\theta}$ is the set of parameters with $\boldsymbol{\theta}_a = \mathrm{vec}(\{a_r\}_{r=1}^m)$, $\boldsymbol{\theta}_{\boldsymbol{w}} = \mathrm{vec}(\{\boldsymbol{w}_r\}_{r=1}^m)$, and we impose hereafther that the activation function $\sigma(\cdot)$ to be continuously differentiable up to order 6, i.e., $\sigma \in \mathcal{C}^6(\mathbb{R})$. More precisely, $\boldsymbol{\theta} = \mathrm{vec}(\{\boldsymbol{q}_r\}_{r=1}^m)$, where for each $r \in [m]$, $\boldsymbol{q}_r := (a_r, \boldsymbol{w}_r^\intercal)^\intercal$, and the bias term b_r can be incorporated by expanding $\boldsymbol{x}$ and $\boldsymbol{w}_r$ to $(\boldsymbol{x}^\intercal, 1)^\intercal$ and $(\boldsymbol{w}_r^\intercal, b_r)^\intercal$.

3.3 Dropout

For a fixed learning rate $\eta > 0$, then at the N-th iteration where $t_N := N\eta$, a scaling vector $\boldsymbol{\delta}_N \in \mathbb{R}^m$ is sampled with independent random coordinates: For each $k \in [m]$,

$$(\boldsymbol{\delta}_N)_k = \begin{cases} \frac{1}{p} & \text{with probability } p, \\ 0 & \text{with probability } 1-p, \end{cases} \tag{3}$$

and we observe that $\{\boldsymbol{\delta}_N\}_{N\geq 1}$ is an i.i.d. Bernoulli sequence with $\mathbb{E}\boldsymbol{\delta}_N = \mathbf{1}$. With slight abuse of notations, the σ-fields $\mathcal{F}_N := \{\sigma(\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \cdots \boldsymbol{\delta}_N)\}$ forms a natural filtration. We then apply dropout to the two-layer NNs by computing

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\delta}) := \sum_{r=1}^{m} (\boldsymbol{\delta})_r a_r \sigma(\boldsymbol{w}_r^\intercal \boldsymbol{x}), \tag{4}$$

and we denote the empirical risk associated with dropout by

$$R_{\mathcal{S}}^{\mathrm{drop}}(\boldsymbol{\theta}; \boldsymbol{\delta}) := \frac{1}{2n} \sum_{i=1}^{n} \left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i; \boldsymbol{\delta}) - y_i\right)^2 = \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{r=1}^{m} (\boldsymbol{\delta})_r a_r \sigma(\boldsymbol{w}_r^\intercal \boldsymbol{x}_i) - y_i \right)^2. \tag{5}$$

We remark that the parameters at the N-th step are updated as follows:

$$\boldsymbol{\theta}_N = \boldsymbol{\theta}_{N-1} - \eta \nabla_{\boldsymbol{\theta}} R_{\mathcal{S}}^{\mathrm{drop}}(\boldsymbol{\theta}_{N-1}; \boldsymbol{\delta}_N), \tag{6}$$

where $\boldsymbol{\theta}_0 := \boldsymbol{\theta}(0)$. Finally, we denote hereafter that for all $i \in [n]$,

$$e_i^N := e_i(\boldsymbol{\theta}_{N-1}; \boldsymbol{\delta}_N) := f_{\boldsymbol{\theta}_{N-1}}(\boldsymbol{x}_i; \boldsymbol{\delta}_N) - y_i.$$

4 Stochastic modified equations for dropout

In this section, we approximate the iterative process of dropout (6) in the weak sense (Definition 1).

4.1 Modified loss

As the dropout iteration (6) reads

$$\boldsymbol{\theta}_N - \boldsymbol{\theta}_{N-1} = -\eta \nabla_{\boldsymbol{\theta}} R_{\mathcal{S}}^{\mathrm{drop}}(\boldsymbol{\theta}_{N-1}; \boldsymbol{\delta}_N) = -\frac{\eta}{n} \sum_{i=1}^{n} e_i^N \nabla_{\boldsymbol{\theta}} e_i^N.$$

Since $\boldsymbol{\theta} = \mathrm{vec}(\{\boldsymbol{q}_r\}_{r=1}^m) = \mathrm{vec}\left(\{(a_r, \boldsymbol{w}_r)\}_{r=1}^m\right)$, then given $\boldsymbol{\theta}_{N-1}$, for each $k \in [m]$, the expectation of the increment restricted to $\boldsymbol{q}_k$ reads

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}_{N-1}} \left[\sum_{i=1}^{n} e_i^N \nabla_{\boldsymbol{q}_k} e_i^N \right] = \mathbb{E}_{\boldsymbol{\theta}_{N-1}} \left[\sum_{i=1}^{n} e_i^N (\boldsymbol{\delta}_N)_k \nabla_{\boldsymbol{q}_k} \left(a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\right) \right] \\ =& \sum_{i=1}^{n} e_i \nabla_{\boldsymbol{q}_k} \left(a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\right) + \frac{1-p}{p} \sum_{i=1}^{n} a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i) \nabla_{\boldsymbol{q}_k} \left(a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i)\right), \end{aligned}$$

where we denote for simplicity that $e_i := e_i(\boldsymbol{\theta}) := \sum_{r=1}^{m} a_r \sigma(\boldsymbol{w}_r^\intercal \boldsymbol{x}_i) - y_i$. Compared with e_i^N, e_i does not depend on the random variable $\boldsymbol{\delta}_N$. Hence, as we define the *modified loss* $L_{\mathcal{S}}(\cdot) : \mathbb{R}^D \to \mathbb{R}$ for dropout:

$$L_{\mathcal{S}}(\boldsymbol{\theta}) := \frac{1}{2n} \sum_{i=1}^{n} e_i^2 + \frac{1-p}{2np} \sum_{i=1}^{n} \sum_{r=1}^{m} a_r^2 \sigma(\boldsymbol{w}_r^\intercal \boldsymbol{x}_i)^2. \tag{7}$$

We observe that for each $k \in [m]$, the gradient of $L_{\mathcal{S}}$ restricted to $\boldsymbol{q}_k$ reads

$$\nabla_{\boldsymbol{q}_k} L_{\mathcal{S}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} e_i \nabla_{\boldsymbol{q}_k} \left(a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i) \right) + \frac{1-p}{np} \sum_{i=1}^{n} a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i) \nabla_{\boldsymbol{q}_k} \left(a_k \sigma(\boldsymbol{w}_k^\intercal \boldsymbol{x}_i) \right),$$

which indicates that given $\boldsymbol{\theta}_{N-1}$, the conditional expectation of the increment of the parameter at the N-th step reads

$$\boldsymbol{\theta}_N - \boldsymbol{\theta}_{N-1} = -\eta \mathbb{E}_{\boldsymbol{\theta}_{N-1}} \left[\nabla_{\boldsymbol{\theta}} R_{\mathcal{S}}^{\mathrm{drop}} \left(\boldsymbol{\theta}_{N-1}; \boldsymbol{\delta}_N \right) \right] = -\eta \nabla_{\boldsymbol{\theta}} L_{\mathcal{S}}(\boldsymbol{\theta}) \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{N-1}}.$$

Then in the sense of expectations, $\{\boldsymbol{\theta}_N\}_{N \geq 0}$ follows close to the gradient descent (GD) trajectory of $L_{\mathcal{S}}(\boldsymbol{\theta})$ with fixed learning rate η. In the above procedure, we focus on the drift term of dropout and disregard its fluctuation term as we merely consider the first conditional moment of the parameter increment. Please refer to Appendix G.1 for the detailed derivation of $L_{\mathcal{S}}$.

4.2 Stochastic modified equations

In pursuit of a more comprehensive understanding of the dynamics of dropout, we integrate the fluctuation term of dropout into our analysis. Firstly, as shown above, we observe that given $\boldsymbol{\theta}_{N-1}$,

$$\boldsymbol{\theta}_N - \boldsymbol{\theta}_{N-1} = -\eta \nabla_{\boldsymbol{\theta}} L_{\mathcal{S}}(\boldsymbol{\theta}) \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{N-1}} + \sqrt{\eta} \boldsymbol{V}(\boldsymbol{\theta}_{N-1}), \tag{8}$$

where $L_{\mathcal{S}}(\cdot) : \mathbb{R}^D \to \mathbb{R}$ is the modified loss defined in (7), and $\boldsymbol{V}(\cdot) : \mathbb{R}^D \to \mathbb{R}^D$ represents the fluctuation term of dropout. When given $\boldsymbol{\theta}_{N-1}$, $\boldsymbol{V}(\boldsymbol{\theta}_{N-1})$ has mean $\mathbf{0}$ and covariance $\eta \boldsymbol{\Sigma}(\boldsymbol{\theta}_{N-1})$, where $\boldsymbol{\Sigma}(\cdot) : \mathbb{R}^D \to \mathbb{R}^{D \times D}$, whose expression is deferred to Section 5.1.

Consider the stochastic differential equation (SDE),

$$\mathrm{d}\boldsymbol{\Theta}_t = \boldsymbol{b}\left(\boldsymbol{\Theta}_t\right) \mathrm{d}t + \boldsymbol{\sigma}\left(\boldsymbol{\Theta}_t\right) \mathrm{d}\boldsymbol{W}_t, \quad \boldsymbol{\Theta}_0 = \boldsymbol{\Theta}(0), \tag{9}$$

where $\boldsymbol{W}_t$ is a standard D-dimensional Brownian motion. Its Euler–Maruyama discretization with step size $\eta > 0$ at the N-th step reads

$$\boldsymbol{\Theta}_{\eta N} = \boldsymbol{\Theta}_{\eta(N-1)} + \eta \boldsymbol{b}\left(\boldsymbol{\Theta}_{\eta(N-1)}\right) + \sqrt{\eta} \boldsymbol{\sigma}\left(\boldsymbol{\Theta}_{\eta(N-1)}\right) \boldsymbol{Z}_N,$$

where $\boldsymbol{Z}_N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_D)$ and $\boldsymbol{\Theta}_0 = \boldsymbol{\Theta}(0)$. Thus, if we set

$$\begin{aligned} \boldsymbol{b}\left(\boldsymbol{\Theta}\right) &:= -\nabla_{\boldsymbol{\Theta}} L_{\mathcal{S}}(\boldsymbol{\Theta}), \\ \boldsymbol{\sigma}\left(\boldsymbol{\Theta}\right) &:= \sqrt{\eta} \left(\boldsymbol{\Sigma}\left(\boldsymbol{\Theta}\right)\right)^{\frac{1}{2}}, \\ \boldsymbol{\Theta}_0 &:= \boldsymbol{\theta}_0, \end{aligned} \tag{10}$$

then we would expect (9) to be a "good" approximation of (8) with time identification $t = \eta N$. Based on the previous work (Li et al., 2017), we use approximations in the *weak* sense (Kloeden and Platen, 2011, Section 9.7) since the path of dropout and the corresponding SDE are driven by noises sampled in different spaces.

To compare different discrete time approximations, we need to take the rate of weak convergence into consideration, and we also need to choose an appropriate class of functions as the space of test functions. We introduce the following set of smooth functions:

$$\mathcal{C}_b^M\left(\mathbb{R}^D\right) = \left\{ f \in \mathcal{C}^M\left(\mathbb{R}^D\right) \,\middle|\, \|f\|_{\mathcal{C}^M} := \sum_{|\beta| \leq M} \left\| \mathrm{D}^\beta f \right\|_\infty < \infty \right\}, \tag{11}$$

where D is the usual differential operator. We remark that $\mathcal{C}_b^M(\mathbb{R}^D)$ is a subset of $\mathcal{G}(\mathbb{R}^D)$, the class of functions with polynomial growth, which is chosen to be the space of test functions in previous works (Li et al., 2017; Kloeden and Platen, 2011). To ensure validity of our analysis, we assume that

Assumption 1. *There exists $T^* > 0$, such that for any $t \in [0, T^*]$, there exists a unique t-continuous solution $\boldsymbol{\Theta}_t$ to SDE* (9)*. Furthermore, for each $l \in [3]$, there exists $C(T^*, \boldsymbol{\Theta}_0) > 0$, such that*

$$\sup_{0 \le s \le T^*} \mathbb{E}\left(\|\boldsymbol{\Theta}_s(\cdot)\|_2^{2l}\right) \le C(T^*, \boldsymbol{\Theta}_0). \tag{12}$$

Moreover, for the dropout iterations (6)*, let $0 < \eta < 1$, $T > 0$ and set $N_{T,\eta} := \lfloor \frac{T}{\eta} \rfloor$. There exists $\eta_0 > 0$, such that given any learning rate $\eta \le \eta_0$, then for all $N \in [0 : N_{T^*,\eta}]$ and for each $l \in [3]$, there exists $C(T^*, \boldsymbol{\theta}_0, \eta_0) > 0$, such that*

$$\sup_{0 \le N \le [N_{T^*,\eta}]} \mathbb{E}\left(\|\boldsymbol{\theta}_N\|_2^{2l}\right) \le C(T^*, \boldsymbol{\theta}_0, \eta_0). \tag{13}$$

We remark that local existence of the solution to SDE and estimates of all $2l$-moments of the solution to SDE can be guaranteed for smooth coefficients and sufficiently small time $T^* > 0$. Moreover, as the constants $C(T^*, \boldsymbol{\Theta}_0)$ and $C(T^*, \boldsymbol{\theta}_0, \eta_0)$ are exponential in time, the $2l$-moments of the solution might blow up for large enough T^*, which is unavoidable since we are unable to impose the uniform Lipschitz condition on $\nabla \mathcal{L}_{\mathcal{S}}$ and $\boldsymbol{\Sigma}$. However, our empirical findings suggest that the SME still possess the desired approximation ability to dropout even for a large learning rate, as shown in Fig. 1 (a). We also remark that if $\mathcal{G}(\mathbb{R}^D)$ is chosen to be the test functions in Li et al. (2019), then similar relations to (12) and (13) shall be imposed, except that in our cases, we only require the second, fourth and sixth moments to be uniformly bounded.

Definition 1. *The SDE* (9) *is an order α weak approximation to the dropout* (6)*, if for every $g \in \mathcal{C}_b^M\left(\mathbb{R}^D\right)$, there exists $C > 0$ and $\eta_0 > 0$, such that given any $\eta \le \eta_0$ and $T \le T^*$, then for all $N \in [N_{T,\eta}]$,*

$$|\mathbb{E}g(\boldsymbol{\Theta}_{\eta N}) - \mathbb{E}g(\boldsymbol{\theta}_N)| \le C(T^*, g, \eta_0)\eta^\alpha. \tag{14}$$

We now state formally our approximation results.

Theorem 1. *Fix time $T \le T^*$ and learning rate $\eta > 0$. If $\sigma \in \mathcal{C}^6(\mathbb{R})$, then for all $t \in [0, T]$, the stochastic processes $\boldsymbol{\Theta}_t$ satisfying*

$$\mathrm{d}\boldsymbol{\Theta}_t = \boldsymbol{b}_1\left(\boldsymbol{\Theta}_t\right)\mathrm{d}t + \boldsymbol{\sigma}_1\left(\boldsymbol{\Theta}_t\right)\mathrm{d}\boldsymbol{W}_t, \tag{15}$$

is an order-1 approximation of dropout (6)*, where*

$$\begin{aligned}\boldsymbol{b}_1(\boldsymbol{\Theta}) &= -\nabla_{\boldsymbol{\Theta}} L_{\mathcal{S}}(\boldsymbol{\Theta}),\\ \boldsymbol{\sigma}_1(\boldsymbol{\Theta}) &= \sqrt{\eta}\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Theta}\right)\right)^{\frac{1}{2}},\end{aligned}$$

and the expression of $L_{\mathcal{S}}(\cdot)$ is located in (7)*, and the expression of $\boldsymbol{\Sigma}(\cdot)$ can be found in Appendix J. Moreover, if $\sigma \in \mathcal{C}^6(\mathbb{R})$, then for all $t \in [0, T]$, the stochastic processes $\boldsymbol{\Theta}_t$ satisfying*

$$\mathrm{d}\boldsymbol{\Theta}_t = \boldsymbol{b}_2\left(\boldsymbol{\Theta}_t\right)\mathrm{d}t + \boldsymbol{\sigma}_2\left(\boldsymbol{\Theta}_t\right)\mathrm{d}\boldsymbol{W}_t, \tag{16}$$

is an order-2 approximation of dropout (6)*, where*

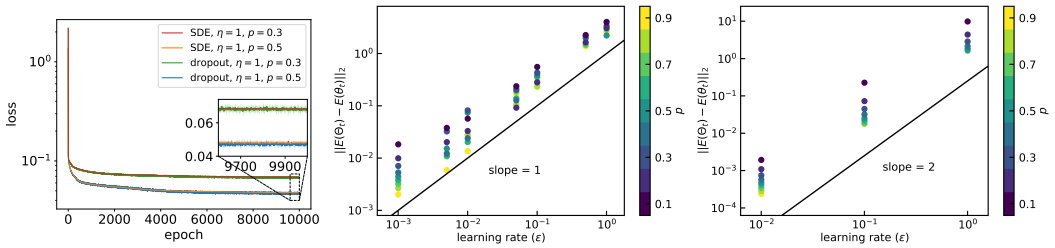
$$\begin{aligned}\boldsymbol{b}_2(\boldsymbol{\Theta}) &= -\nabla_{\boldsymbol{\Theta}}\left(L_{\mathcal{S}}(\boldsymbol{\Theta}) + \frac{\eta}{4}\left\|\nabla_{\boldsymbol{\Theta}} L_{\mathcal{S}}(\boldsymbol{\Theta})\right\|_2^2\right),\\ \boldsymbol{\sigma}_2(\boldsymbol{\Theta}) &= \sqrt{\eta}\left(\boldsymbol{\Sigma}\left(\boldsymbol{\Theta}\right)\right)^{\frac{1}{2}}.\end{aligned}$$

It is noteworthy that our findings reproduce the explicit regularization effect attributed to dropout (Wei et al., 2020; Zhang and Xu, 2022). This regularization effect modifies the expected training objective from the empirical risk $R_{\mathcal{S}}(\boldsymbol{\theta})$ to $L_{\mathcal{S}}(\boldsymbol{\theta})$, and it stems from the inherent stochastic nature of dropout. Unlike SGD, where the noise arises from the stochasticity involved in the selection of training samples, dropout introduces noise by means of the stochastic removal of parameters.

4.3 Numerical simulation of stochastic modified equations

In this subsection, we conduct an empirical validation of the effectiveness of SMEs for dropout. This validation is conducted through an exploration of the resemblance between the numerical simulation of the SME and the real-time training process of dropout. For the numerical simulation of the SME, unless otherwise specified, we employ the Euler-Maruyama method to approximate its dynamic

evolution by the order-1 approximation. It is worth noting that the noise term $\boldsymbol{\sigma}(\boldsymbol{\theta})$ in Equ. (10) involves the computation of the square root of the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Consequently, the size of the covariance matrix significantly affects the speed and accuracy of the numerical simulation process. To mitigate the computational demands associated with the covariance matrix, we resize the MNIST data to 7×7, thereby reducing the number of network parameters involved in the simulations. Additional details of the experimental setup can be found in Appendix A.



(a) loss trajectory (b) order-1 approximation (c) order-2 approximation

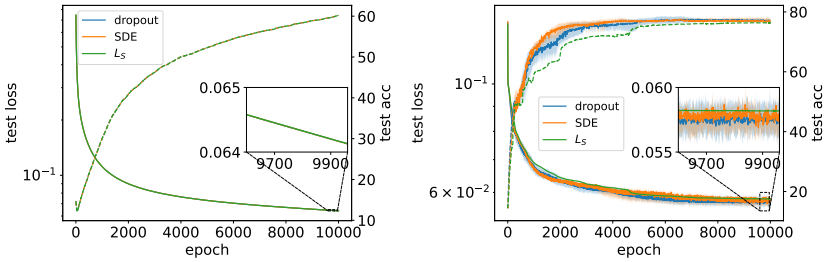
Figure 1: We train two-layer fully connected networks on MNIST. The curves and points are derived from the average results of individual trials, each of which utilized the same initialization distribution. (a) The training loss trajectory obtained by SME simulation or dropout training under four cases of different learning rates and dropout rates. The error bands, portrayed with greater transparency, are derived from the maximum and minimum loss values observed across these six random trials at each training step. (b, c) Convergence-order verification of first-order and second-order SME approximations. Each point represents the value of $\|\mathbb{E}(\boldsymbol{\Theta}_t) - \mathbb{E}(\boldsymbol{\theta}_t)\|_2$ under a given learning rate (abscissa) and dropout rate (color).

Fig. 1 illustrates the close correspondence between the dynamics of the SME and dropout throughout the training process. This similarity is examined from two perspectives: the trajectory of loss functions and the approximation order. In Fig. 2, we emphasize that although dropout introduces a noise component that modifies the loss function from $R_{\mathcal{S}}(\boldsymbol{\theta})$ to $L_{\mathcal{S}}(\boldsymbol{\theta})$, when contrasted with SMEs, the training behavior of gradient descent utilizing $L_{\mathcal{S}}(\boldsymbol{\theta})$ as the loss function is distinct from dropout. This distinction becomes apparent when large learning rates are employed in the optimization process.

To comprehensively assess the similarity between dropout and SME simulations, we first consider four distinct cases, each characterized by various dropout rates and learning rates shown in Fig. 1(a). Fig. 1(a) depicts the evolution of loss values under these distinct settings for both dropout and SME simulations. To ensure the robustness of our analysis, for each configuration, we conduct six independent trials of dropout and SME simulations, all initialized with identical distribution. The displayed curves represent the means of these six random trials. Moreover, the error band, indicated by lighter colors, covers the range between the maximum and minimum loss values obtained from the six trials. The observed alignment of loss trajectories between the SME simulation and dropout, as evident in Fig. 1(a), underscores a prominent resemblance in their respective loss trajectories.

To further evaluate the similarity of their parameters, we verify the approximation orders of different SME simulations. Figs. 1(b, c) numerically verify the approximation orders of the first-order and the second-order approximation equation in Theorem 1 respectively. Each point represents the value of $\|\mathbb{E}(\boldsymbol{\Theta}_t) - \mathbb{E}(\boldsymbol{\theta}_t)\|_2$ under a given learning rate (abscissa) and dropout rate (color). The expectation is obtained by calculating the mean of 10 independent experiments with the same initialization for both dropout and SME simulation. The logarithmic plots clearly illustrate the experimental validation of the theoretical approximation order of SME, for both order-1, shown in Fig. 1(b), and order-2, shown in Fig. 1(c). Additionally, under the same learning rate, larger values of p exhibit enhanced approximation capabilities. This improvement is attributed to the reduction in noise with increasing p, consequently minimizing the impact of noise on the training process.

We also conduct experiments to validate the applicability of the SME approximation in complex networks and SGD settings. In the former, we simulate complex networks through numerical approximation of the drift term, while in the latter, we rely on the fact that SGD noise is unbiased. For a thorough discussion and detailed numerical results, please refer to Appendix C.



(a) $\eta = 0.005, p = 0.9$ (b) $\eta = 1, p = 0.3$

Figure 2: The test loss and test accuracy trajectory obtained by SME simulation, dropout training, and gradient descent training with loss function $L_{\mathcal{S}}(\boldsymbol{\theta})$ under two settings. The curves are derived from the average results of six individual trials, each of which utilized the same initialization distribution. The error bands, portrayed with greater transparency, are derived from the maximum and minimum loss values observed across these six random trials at each training step.

Fig. 2 depicts the test loss and test accuracy associated with $\eta = 0.005$, $p = 0.9$ and $\eta = 1$, $p = 0.3$. In Fig. 2(a), the trajectories of loss and accuracy, generated using three different training methods, exhibit a remarkable degree of concurrence. This phenomenon can be primarily attributed to the utilization of a small learning rate, where the diffusion component significantly diminishes, consequently endowing the drift term $L_{\mathcal{S}}(\boldsymbol{\theta})$ with a dominant influence. In contrast, as illustrated in Fig. 2(b), a notable divergence becomes evident in the loss and accuracy trajectories. This discrepancy arises from the impact of the diffusion term during training, particularly with the application of large learning rates. Notably, in contrast to the training behavior of gradient descent utilizing $L_{\mathcal{S}}(\boldsymbol{\theta})$ as the loss function, the trajectory generated by SME simulation exhibits a closer alignment with the trajectory of dropout.

It is noteworthy that as large learning rate and dropout rate contribute to an increased amplitude of diffusion, methods incorporating noise such as SME and dropout tend to exhibit enhanced generalization performance. As demonstrated in Figure 2(b), the test accuracy attained by $L_{\mathcal{S}}(\boldsymbol{\theta})$ consistently remains below the lower threshold of test accuracy attained by noise-inclusive methods for the major portion of the training duration. In the sequel, we delve into an in-depth exploration of the influence exerted by noise on our learning outcomes.

5 The Effect of Dropout Noise Structure

We begin this section by examining the noise structure of dropout.

5.1 Explicit Form of the Noise Structure of Dropout

In this subsection, we present the expression for the covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Once again, as $\boldsymbol{\theta} = \mathrm{vec}(\{\boldsymbol{q}_r\}_{r=1}^m) = \mathrm{vec}\left(\{(a_r, \boldsymbol{w}_r)\}_{r=1}^m\right)$, then as we denote covariance of $\nabla_{\boldsymbol{\theta}} R_{\mathcal{S}}^{\mathrm{drop}}\left(\boldsymbol{\theta}_{N-1}; \boldsymbol{\delta}_N\right)$ by $\boldsymbol{\Sigma}(\boldsymbol{\theta}_{N-1})$, i.e.,

$$\boldsymbol{\Sigma}_{kr}(\boldsymbol{\theta}_{N-1}) := \mathrm{Cov}\left(\nabla_{\boldsymbol{q}_k} R_{\mathcal{S}}^{\mathrm{drop}}\left(\boldsymbol{\theta}_{N-1}; \boldsymbol{\delta}_N\right), \nabla_{\boldsymbol{q}_r} R_{\mathcal{S}}^{\mathrm{drop}}\left(\boldsymbol{\theta}_{N-1}; \boldsymbol{\delta}_N\right)\right),$$

then

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \cdots & \boldsymbol{\Sigma}_{1m} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} & \cdots & \boldsymbol{\Sigma}_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\Sigma}_{m1} & \boldsymbol{\Sigma}_{m2} & \cdots & \boldsymbol{\Sigma}_{mm} \end{bmatrix}.$$

Such expression of $\boldsymbol{\Sigma}$ arises from the inherent decoupling properties among neurons within the two-layer neural network. Due to space limitation, we defer the detailed expression of $\boldsymbol{\Sigma}_{kr}$ to Appendix J.

5.2 INTUITIVE EXPLANATION FOR THE HESSIAN-VARIANCE ALIGNMENT RELATIONS

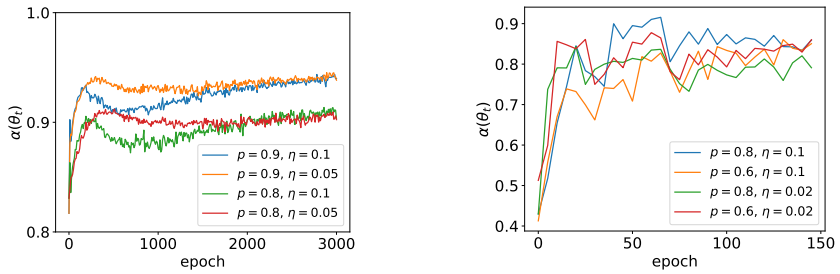
In this subsection, we endeavor to show the structural similarity between the covariance and the Hessian in terms of Hessian-variance alignment relations. Under the assumption that $\boldsymbol{\theta}$ is close to a global minimum, we intuitively derive the structural similarity between the Hessian and the covariance at the final stage of the training process as follows:

$$\begin{aligned}
\boldsymbol{H}(\boldsymbol{\theta}) &\approx \frac{1}{n}\sum_{i=1}^{n}\left[\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)\otimes\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right)+\frac{1-p}{p}\sum_{r=1}^{m}\nabla_{\boldsymbol{q}_r}\left(a_r\sigma(\boldsymbol{w}_r^{\intercal}\boldsymbol{x}_i)\right)\otimes\nabla_{\boldsymbol{q}_r}\left(a_r\sigma(\boldsymbol{w}_r^{\intercal}\boldsymbol{x}_i)\right)\right], \\
\boldsymbol{\Sigma}(\boldsymbol{\theta}) &\approx \frac{1}{n}\sum_{i=1}^{n}\left[l_{i,1}\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\otimes\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)+l_{i,2}\frac{1-p}{p}\sum_{r=1}^{m}\nabla_{\boldsymbol{q}_r}\left(a_r\sigma(\boldsymbol{w}_r^{\intercal}\boldsymbol{x}_i)\right)\otimes\nabla_{\boldsymbol{q}_r}\left(a_r\sigma(\boldsymbol{w}_r^{\intercal}\boldsymbol{x}_i)\right)\right],
\end{aligned} \tag{17}$$

where $\boldsymbol{H}(\boldsymbol{\theta}) := \nabla^2_{\boldsymbol{\theta}} L_{\mathcal{S}}(\boldsymbol{\theta})$, and $l_{i,1} := (e_i)^2 + \frac{1-p}{p}\sum_{r=1}^{m} a_r^2\sigma(\boldsymbol{w}_r^{\intercal}\boldsymbol{x}_i)^2$, $l_{i,2} := (e_i)^2$. A detailed derivation of (17) is provided in Appendix K. With the establishment of structural similarity through the aforementioned intuitive approximations outlined in (17), we proceed to empirically investigate the intricate relationship between the Hessian and the covariance, and details of the experimental settings can be found in Appendix A.

5.3 EXPERIMENTAL RESULTS ON THE HESSIAN-VARIANCE ALIGNMENT RELATIONS

Motivated by the relation (17), we empirically demonstrate the structural similarity between the Hessian and the covariance of dropout, and this demonstration serves to validate the Hessian-variance alignment relation. Based on this relation, the introduction of dropout noise has the potential to expedite the escape of the model from locating sharp minima, thereby effectively enhancing the flatness of the solution. Furthermore, in Appendix B, we also explore another relationship between the Hessian and the covariance known as the inverse variance-flatness relation (Feng and Tu, 2021), which also contributes to aiding the model in avoidance of the sharp minima during its optimization process.



(a) FNN, MNIST (b) VGG16, CIFAR-10

Figure 3: The cosine similarity $\alpha(\boldsymbol{\theta}_t)$ between the Hessian of the loss function and the covariance of the dropout noise at each training epoch t for different choices of dropout rate and learning rate. (a) The FNN with size 784-50-50-10 is trained on the MNIST dataset using the first 10000 examples as the training dataset. The dropout layer is added after the first hidden layer. (b) The VGG16 is trained on the CFIAR-10 dataset using the full examples as the training dataset. The dropout layers are added after the first two convolutional layers of each block and the first fully-connected layer. The calculation of $\alpha(\boldsymbol{\theta}_t)$ is performed every five epochs.

To investigate the Hessian-Variance alignment relation, we study the cosine similarity quantity $\alpha(\boldsymbol{\theta}_t)$[1] between the covariance matrix $\boldsymbol{\Sigma}_t := \boldsymbol{\Sigma}(\boldsymbol{\theta}_t)$ and the Hessian matrix $\boldsymbol{H}_t := \boldsymbol{H}(\boldsymbol{\theta}_t)$ at each time step t. $\boldsymbol{\Sigma}_t$ is the covariance matrix of $\mathcal{D}_{\text{grad}}$, a collection of gradients calculated with different dropout variables $\boldsymbol{\delta}$ sampled at the tth step, whose detailed definition can be found in Section B.1. On the other hand, $\boldsymbol{H}_t$ is the Hessian of the loss function evaluated at the tth iteration. Then the crucial

[1]This variable is also used in Wu et al. (2022) for studying SGD.

cosine similarity metric $\alpha(\boldsymbol{\theta}_t)$ is formally expressed as:

$$\alpha(\boldsymbol{\theta}_t) = \frac{\mathrm{Tr}(\boldsymbol{H}_t \boldsymbol{\Sigma}_t)}{\|\boldsymbol{H}_t\|_{\mathrm{F}} \|\boldsymbol{\Sigma}_t\|_{\mathrm{F}}} \tag{18}$$

As depicted in Fig. 3, it is evident that throughout the training process, $\alpha(\boldsymbol{\theta}_t)$ consistently attains values surpassing 0.85 in Fig. 3(a) and 0.7 in Fig. 3(b), and these observations hold true across varying learning rates and dropout rates. It's worth noting that, based on 100 samples, the average cosine similarity between the two random matrices based on the selected parameters is only 7.8×10^{-4}. The eigenvalues of the two random matrices are derived from the eigenvalues of the Hessian matrix and covariance matrix of the selected parameters, and the corresponding eigenvectors are sampled from a normal distribution and normalized. The model parameters are derived from the final model represented by the blue line in Fig. 3(a). Consequently, the introduced noise is highly anisotropic in that it aligns well with the Hessian matrix across all directions. We acknowledged that due to computational constraints, this experiment limits the trace calculation to a subset of parameters.

6 Conclusions and discussions

Our main contribution comprises two key aspects. First, we derive the SMEs that provide a weak approximation to the dynamics of the dropout algorithm applied to two-layer NNs. Second, we conduct an empirical inquiry that demonstrates the persistent validity of the Hessian-variance alignment relation throughout the training process of dropout. The Hessian-variance alignment relation has been established to be beneficial for the model to locate flatter minima, thus indicating that dropout acts as an implicit regularizer that enhances the generalization power possessed by the model.

Extension of the SME framework to multi-layer networks and SGD. While our theoretical analysis has predominantly centered around the dropout algorithm applied to two-layer neural networks and GD, it is important to note that the derivation of SMEs is not confined exclusively to two-layer neural networks, GD, or even to the dropout algorithm. For various types of neural networks, the feasibility of constructing such modified equations remains viable, provided that the stochastic algorithm iteratively updates the parameters in a recursive manner, i.e., iterations form a time-homogeneous Markov chain. Furthermore, this applicability holds as long as Taylor's theorem with the Lagrange form of the remainder remains valid for sufficiently small learning rates. It is worth acknowledging that the complexity introduced by multi-layer networks primarily arises from the presence of dropout layers within the activation functions. This introduces a high degree of non-linearity to the loss with respect to the dropout variable, rendering it challenging to explicitly calculate the drift and diffusion components of the SME. We numerically verify the SDE approximation capability of complex network structures and SGD in Appendix C.

The effect of learning rate on dropout. In the small learning rate regime, wherein the noise term exerts slight influence, the loss trajectories of $L_S(\boldsymbol{\theta})$ and $R_S^{\mathrm{drop}}(\boldsymbol{\theta}; \boldsymbol{\delta})$ exhibit a notable degree of congruence. This observation has been affirmed through theoretical and empirical validations. However, it remains imperative to maintain the diffusion term is important if we aspire to gain deeper insights into the nature of dropout algorithms or other stochastic algorithms. As illustrated in Fig. 2(b), in the large learning rate regime, the trajectory derived from the SME simulation aligns more closely with its dropout counterpart, in stark contrast to the trajectory arising from GD training on $L_S(\boldsymbol{\theta})$. Furthermore, SMEs consistently exhibit better generalization capability in comparison to GD. Therefore, a comprehensive analytical framework that duly accommodates both drift and diffusion terms stands as a more informative tool for the insightful analysis of dropout algorithms.

More refined analysis of noise structures. In addition to the Hessian-Variance alignment relation, the structural similarity between the Hessian and the covariance engenders yet another intriguing relationship known as the inverse variance-flatness relation (Feng and Tu, 2021). Different from the Hessian-Variance alignment relation, it focuses more on the similarity of the two feature directions. In Appendix B, an investigation has been conducted to examine the correlation between the noise structure introduced by dropout and the nature of the loss landscape. This relationship also plays a pivotal role in assisting the model to steer clear of sharp minima. The high similarity of the eigenvectors of two matrices is a natural extension of the inverse variance-flatness relation, please refer to Appendix B for detailed validation results. In Appendix D, we compare the effect of noise on the model in three training strategies, dropout, SGD, and parametric noise injection (Orvieto et al., 2023), which all appear to be helpful for flatness.

ACKNOWLEDGMENTS

This work is sponsored by the National Key R&D Program of China Grant No. 2022YFA1008200 (Z. X., T. L.), the National Natural Science Foundation of China Grant No. 92270001 (Z. X.), 12371511 (Z. X.), 12101401 (T. L.), Shanghai Municipal of Science and Technology Major Project No. 2021SHZDZX0102 (Z. X., T. L.), Shanghai Municipal Science and Technology Key Project No. 22JC1401500 (T. L.), and the HPC of School of Mathematical Sciences and the Student Innovation Center, and the Siyuan-1 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University.

REFERENCES

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

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

M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.

D. P. Helmbold, P. M. Long, On the inductive bias of dropout, The Journal of Machine Learning Research 16 (2015) 3403–3454.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. T. P. Tang, On large-batch training for deep learning: Generalization gap and sharp minima, arXiv preprint arXiv:1609.04836 (2016).

Y. Feng, Y. Tu, The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima, Proceedings of the National Academy of Sciences 118 (2021).

Z. Zhu, J. Wu, B. Yu, L. Wu, J. Ma, The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects, arXiv preprint arXiv:1803.00195 (2018).

C. Wei, S. Kakade, T. Ma, The implicit and explicit regularization effects of dropout, in: International Conference on Machine Learning, PMLR, 2020, pp. 10181–10192.

Z. Zhang, Z.-Q. J. Xu, Implicit regularization of dropout, arXiv preprint arXiv:2207.05952 (2022).

Q. Li, C. Tai, E. Weinan, Stochastic modified equations and adaptive stochastic gradient algorithms, in: International Conference on Machine Learning, PMLR, 2017, pp. 2101–2110.

L. Wu, M. Wang, 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 (2022) 4680–4693.

B. Neyshabur, S. Bhojanapalli, D. McAllester, N. Srebro, Exploring generalization in deep learning, arXiv preprint arXiv:1706.08947 (2017).

S. Wager, S. Wang, P. S. Liang, Dropout training as adaptive regularization, Advances in neural information processing systems 26 (2013) 351–359.

D. McAllester, A pac-bayesian tutorial with a dropout bound, arXiv preprint arXiv:1307.2118 (2013).

L. Wan, M. Zeiler, S. Zhang, Y. Lecun, R. Fergus, Regularization of neural networks using dropconnect, in: In Proceedings of the International Conference on Machine learning, Citeseer, 2013.

W. Mou, Y. Zhou, J. Gao, L. Wang, Dropout training, data-dependent regularization, and generalization bounds, in: International conference on machine learning, PMLR, 2018, pp. 3645–3653.

J. Cavazza, P. Morerio, B. Haeffele, C. Lane, V. Murino, R. Vidal, Dropout as a low-rank regularizer for matrix factorization, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 435–444.

P. Mianjy, R. Arora, On convergence and generalization of dropout training, Advances in Neural Information Processing Systems 33 (2020).

R. Arora, P. Bartlett, P. Mianjy, N. Srebro, Dropout: Explicit forms and capacity control, in: International Conference on Machine Learning, PMLR, 2021, pp. 351–361.

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

T. Luo, Z.-Q. J. Xu, Z. Ma, Y. Zhang, Phase diagram for two-layer relu neural networks at infinite-width limit, Journal of Machine Learning Research 22 (2021) 1–47.

H. Zhou, Q. Zhou, T. Luo, Y. Zhang, Z.-Q. J. Xu, Towards understanding the condensation of neural networks at initial training, arXiv preprint arXiv:2105.11686 (2021).

H. Zhou, Q. Zhou, Z. Jin, T. Luo, Y. Zhang, Z.-Q. J. Xu, Empirical phase diagram for three-layer neural networks with infinite width, Advances in Neural Information Processing Systems (2022).

Q. Li, C. Tai, E. Weinan, Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations, The Journal of Machine Learning Research 20 (2019) 1474–1520.

Y. Feng, L. Li, J.-G. Liu, Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations, arXiv preprint arXiv:1712.06509 (2017).

S. Malladi, K. Lyu, A. Panigrahi, S. Arora, On the SDEs and scaling rules for adaptive gradient algorithms, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022. URL: https://openreview.net/forum?id=F2mhzjHkQP.

H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, arXiv preprint arXiv:1712.09913 (2017).

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, A. Storkey, Three factors influencing minima in sgd, arXiv preprint arXiv:1711.04623 (2017).

S. Jastrzebski, Z. Kenton, N. Ballas, A. Fischer, Y. Bengio, A. Storkey, On the relation between the sharpest directions of dnn loss and the sgd step length, arXiv preprint arXiv:1807.05031 (2018).

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

V. Papyan, The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size, arXiv preprint arXiv:1811.07062 (2018).

V. Papyan, Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians, arXiv preprint arXiv:1901.08244 (2019).

P. Kloeden, E. Platen, Numerical Solution of Stochastic Differential Equations, Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg, 2011. URL: https://books.google.com.hk/books?id=BCvtssom1CMC.

A. Orvieto, A. Raj, H. Kersting, F. Bach, Explicit regularization in overparametrized models via noise injection, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 7265–7287.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

R. Shwartz-Ziv, N. Tishby, Opening the black box of deep neural networks via information, arXiv preprint arXiv:1703.00810 (2017).

S. P. Meyn, R. L. Tweedie, Markov chains and stochastic stability, Springer Science & Business Media, 2012.

Y. Feng, L. Li, J.-G. Liu, Semigroups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations, Communications in Mathematical Sciences 16 (2018) 777–789.

M. Hairer, Ergodic theory for stochastic pdes, preprint (2008).

B. Oksendal, Stochastic differential equations: an introduction with applications, Springer Science & Business Media, 2013.