Almost sure convergence of dropout algorithms for neural networks

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Almost Sure Convergence of Dropout Algorithms for Neural Networks

Albert Senen–Cerda [1]   Jaron Sanders [1]

Abstract

We investigate the convergence and convergence rate of stochastic training algorithms for Neural Networks (NNs) that, over the years, have spawned from *Dropout* (Hinton et al., 2012). Modeling that neurons in the brain may not fire, dropout algorithms consist in practice of multiplying the weight matrices of a NN component-wise by independently drawn random matrices with $\{0, 1\}$-valued entries during each iteration of the Feedforward–Backpropagation algorithm. This paper presents a probability theoretical proof that for any NN topology and differentiable polynomially bounded activation functions, if we project the NN's weights into a compact set and use a dropout algorithm, then the weights converge to a unique stationary set of a projected system of Ordinary Differential Equations (ODEs). We also establish an upper bound on the rate of convergence of Gradient Descent (GD) on the limiting ODEs of dropout algorithms for arborescences (a class of trees) of arbitrary depth and with linear activation functions.

1. Introduction

Machine learning and especially NNs have found ample use in present-day big data applications. Even though the models as well as the training algorithms for NNs have been known since the 1980s, a full mathematical understanding is missing. Key questions include the surprising success of Stochastic Gradient Descent (SGD) at finding good local minima on a nonconvex risk function (Bhojanapalli et al., 2016), and why overparameterized NNs perform well in spite of the concern of overfitting (Gunasekar et al., 2017).

Several stochastic training algorithms for NNs to avoid

overfitting have spawned from the introduction of *Dropout* (Hinton et al., 2012). Modeling how neurons in the brain may not fire, dropout algorithms consist in practice of multiplying weight matrices of the NN component-wise by independently drawn random matrices with $\{0, 1\}$-valued entries in each iteration of the Feedforward–Backpropagation (FB) algorithm. The elements of these random masking matrices indicate whether each individual weight is (0), or is not masked (1) during a training step, see also Figure 1. Mathematically, this turns the FB algorithm into a step of a SGD algorithm in which the primary source of randomness is the stochastic NN's configuration. Under mild independence assumptions, dropout algorithms can be understood to minimize a risk function averaged over all possible NN configurations, see e.g. (Baldi & Sadowski, 2013). These stochastic training algorithms are thus forms of ensemble training, and intuitively, this explains why there is regularization when using them.

Dropout algorithms are interesting because they lie on the intersection of percolation theory and stochastic optimization. Percolation theory studies the properties of connected components in random graphs, with the canonical example being *bond percolation* (Broadbent & Hammersley, 1957). The two-dimensional bond percolation problem is as follows: consider a lattice of $L \times L$ vertices, randomly remove some of the edges, and ask what is the probability that there exists a path from e.g. left to right. Dropout algorithms' connections to bond percolation become immediately clear when we consider that for an iteration of a dropout algorithm to contribute a potentially useful step towards a minimum of its risk function, there must be a path from input to output in the NN after having applied the random masks. Knowing of the connection to bond percolation, for the same number of iterations, one may therefore at first glance expect that dropout performs worse than a routine implementation of the FB algorithm, but in fact, dropout algorithms usually perform well due to their regularization properties (Hinton et al., 2012; Srivastava et al., 2014). From the point of view of bond percolation, however, this should still come at the cost of convergence rates of dropout algorithms. The convergence rate should depend on the configuration of the NN and the dropout algorithm's mask variables. Exactly how this dependence is

[1]Department of Mathematics & Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands. Correspondence to: Albert Senen-Cerda <a.senen.cerda@tue.nl>, Jaron Sanders <jaron.sanders@tue.nl>.
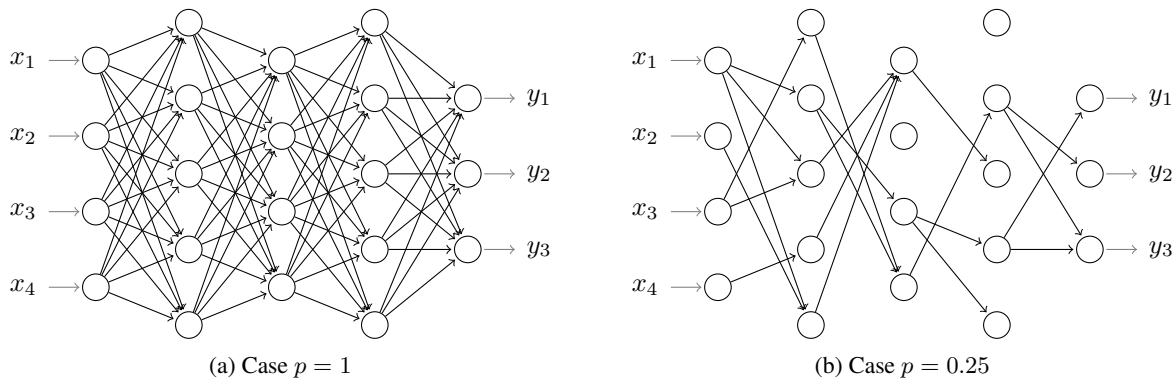
(a) Case $p = 1$ (b) Case $p = 0.25$

Figure 1. Dropconnect's training step (Wan et al., 2013) in a $L = 4$ NN. In this algorithm, every time a sample is provided to a NN, a random NN is first generated by flipping a biased coin for each edge that shows head with probability $p \in (0, 1]$. The output of this random NN is then used to update all weights using the FB algorithm. This effectively implements a SGD on random NNs.

is unclear, and in fact bounds for specific convergence rates of dropout algorithms are unknown. For a full, comprehensive study of the convergence rates of dropout algorithms, one needs to combine percolation theory with stochastic optimization, and this leads to a challenging analysis.

This paper presents a first convergence analysis of dropout through two results. Our first result is a formal probability theoretical proof that for any (fully connected) NN topology and with differentiable polynomially bounded activation functions, if we project a dropout algorithm's weights onto a compact convex set, then the weights converge to a unique stationary set of a projected system of ODEs. This result gives us the formal guarantee that the dropout algorithm is well-behaved for a wide range of NNs and activation functions, and will at least asymptotically (meaning after sufficiently many iterations) not suffer from problems of percolative nature. Pragmatically the projection assumption is furthermore mild and is used since the compact set can be chosen arbitrarily and thus as large as one would like, and the truncation of large variables in computer algorithms is common especially in light of memory constraints. It must be noted, however, that the projection assumption may induce artificial stationary points on the boundary of the compact set, although we expect convergence to such points to be unlikely if the compact set is chosen sufficiently large. Identifying the probability with which projected dropout converges to such an artificial stationary point by generalizing techniques from e.g. (Dupuis & Kushner, 1985; 1989; Buche & Kushner, 2002), would be interesting future work.

While general, our first result lacks specificity: for example, it only characterizes the limit points implicitly, and it does not establish the rate of convergence of dropout. Our second result does establish bounds on convergence rates, but consequentially rely on stronger structural assumptions. Studying more restrictive NN configurations such as lines

(Shamir, 2018), and full linear L-layer NNs (Arora et al., 2018) is however common within the scientific literature on the convergence of GD in NNs. Even without a dropout algorithm this analysis is already a substantial theoretical challenge since the optimization landscape is highly non-convex. The convergence rates of SGD are not fully understood for many NNs. Concretely, our second result is an explicit upper bound on the rates of convergence of regular GD on the limiting ODEs of dropout algorithms for arborescences (a class of trees), of arbitrary depth with linear activation functions. While GD on a limiting ODE is not exactly a dropout algorithm—it is deterministic and not stochastic—analyzing its convergence rate is a major and necessary step towards analyzing the convergence rate of dropout algorithms. This result furthermore does not rely on a projection assumption, and the global minimizer can be explicitly characterized. Such explicit characterization of global minima is generally nontrivial when using dropout algorithms due to their regularization effects (Mianjy et al., 2018; Mianjy & Arora, 2019). The reasons we restrict to arborescences for now is that these (i) are subject to bond percolation, note for example that in a line configuration the probability of there being a path from input to output is exponentially small in the number of layers and this will negatively impact the convergence rate; and (ii) we can explicitly tie the upper bound for the convergence rate to structural properties of the arborescence such as depth and number of paths. Besides giving insight into how NNs are trained when using dropout algorithm, our results hint at what a NN configuration should look like in order to improve convergence rates for dropout algorithms.

1.1. Literature overview

The first description of a dropout algorithm was in (Hinton et al., 2012). Diverse variants of the algorithm have appeared since, including versions in which edges are

dropped (Wan et al., 2013), groups of edges are dropped from the input layer (DeVries & Taylor, 2017), the removal probabilities change adaptively (Ba & Frey, 2013; Li et al., 2016); and that are suitable for recurrent NNs (Zaremba et al., 2014; Semeniuta et al., 2016). The performance of the original algorithm has been investigated on datasets (Hinton et al., 2012; Srivastava et al., 2014), and dropout algorithms have found application in e.g. image classification (Krizhevsky et al., 2012), handwriting recognition (Pham et al., 2014), heart sound classification (Kay & Agarwal, 2016), and drug discovery in cancer research (Urban et al., 2018).

Theoretical studies of dropout algorithms have focused on their regularization effect. The effect was first noted in (Hinton et al., 2012; Srivastava et al., 2014), and subsequently investigated more in-depth for both linear NNs as well as nonlinear NNs in (Baldi & Sadowski, 2013; Wager et al., 2013; Baldi & Sadowski, 2014). Within the context of matrix factorization, it was then shown that dropout's regularization induces an equivalent deterministic optimization problem with regularization on the factors (Cavazza et al., 2017a;b). Characterizations of dropout's risk function and dropout's regularizer for (usually linear) NNs can be found in (Mianjy et al., 2018; Mianjy & Arora, 2019; Pal et al., 2019). There exists however no prior work on whether dropout algorithm as stochastic training algorithms are well-behaved and converge, nor on the impact of the NNs configuration on a dropout algorithm's rate of convergence. It is noteworthy that these questions have been studied within the context of NNs being trained *without* dropout algorithms, see for instance (Arora et al., 2018; Shamir, 2018; Zou et al., 2018).

Dropout can, by construction, be understood as a form of SGD. More generally, dropout algorithms are all stochastic approximation algorithms. The basic stochastic approximations algorithms were first introduced in (Robbins & Monro, 1951; Kiefer et al., 1952), and have been subject to enormous literature due to their ubiquity. For overviews, we refer to (Kushner & Yin, 2003; Borkar, 2009); we rely on the former to prove our first result.

Overview. Section 1 contains our introduction and gives an overview of the related literature on dropout algorithms and some previous convergence results on NNs. In Section 2, we lay out notation, recall the FB algorithm for the reader, and describe the class of dropout algorithms that we study. Sections 3, 4 contain our main results, proof outlines, and discussions thereof. Finally, we conclude in Section 5 with also ideas for future work. The supplementary material contains the details of our proofs.

Notation. Deterministic sequences are indexed with curly brackets in this paper: $\alpha^{\{1\}}, \alpha^{\{2\}}, \cdots$. This is to distinguish from sequences of random variables, which are in-

dexed using square brackets, e.g. $X^{[1]}, X^{[2]}, \cdots$.

Deterministic vectors are written lower case $x \in \mathbb{R}^d$, but an exception is made for random variables (which are always capitalized). Matrices are also always capitalized. For a function $\sigma : \mathbb{R} \to \mathbb{R}$ and a matrix $A \in \mathbb{R}^{a \times b}$, $a, b \geq 1$, we denote $\sigma(A)$ the matrix with σ applied entry-wise to A. The entries of any tensor will be referred to using subscripts, e.g. x_i, $A_{i,j}$, or $T_{i,j,l}$. For any vector $x \in \mathbb{R}^d$, the ℓ_2-norm is defined as $\|x\|_2 \triangleq (\sum_{i=1}^d |x_i|^2)^{1/2}$. For any matrix $A \in \mathbb{R}^{a \times b}$, the Frobenius norm is defined as $\|A\|_F \triangleq (\sum_{i=1}^a \sum_{j=1}^b |A_{i,j}|^2)^{1/2}$. For two matrices A, B, the Hadamard (entrywise) product is denoted by $A \odot B$.

Let $\mathbb{N}_+$ be the strictly positive integers and $\mathbb{N}_0 \triangleq \mathbb{N}_+ \cup \{0\}$. For $l \in \mathbb{N}_+$, we denote $[l] = \{1, \ldots, l\}$. For a function $f \in C^2(\mathbb{R}^n)$, we denote the gradient and Hessian of f with respect to the Euclidean norm $\|\cdot\|_2$ in $\mathbb{R}^n$, by ∇f and $\nabla^2 f$ respectively.

2. Model

2.1. Neural Networks (NNs), and their structure

Let L denote the number of layers in the NN, and let $d_l \in \mathbb{N}_+$ denote the output dimension of layer $l = 1, \ldots, L$. Let $W_{l+1} \in \mathbb{R}^{d_{l+1} \times d_l}$ denote the matrices of weights in between layers l and $l+1$ for $l = 0, 1, \ldots, L-1$. Denote $W = (W_L, \ldots, W_1) \in \mathcal{W}$, with $\mathcal{W} \triangleq \mathbb{R}^{d_L \times d_{L-1}} \times \cdots \times \mathbb{R}^{d_1 \times d_0}$ the set of weights.

Definition 1. *Let σ be an activation function $\sigma : \mathbb{R} \to \mathbb{R}$. A Neural Network (NN) with L layers is given by the class of functions $\Psi_W : \mathbb{R}^{d_0} \to \mathbb{R}^{d_L}$ defined iteratively by*

$$
\begin{aligned}
A_0 &= x, \\
A_{i+1} &= \sigma(W_i A_{i-1}) \quad \forall i \in \{1, \ldots, L-2\}, \\
\Psi_W(x) &= W_L A_{L-1} = A_L.
\end{aligned}
\tag{1}
$$

Canonical activation functions include the linear function $\sigma(t) = t$, the Rectified Linear Unit (ReLU) function $\sigma(t) = \max\{0, t\}$, and the sigmoid function $\sigma(t) = 1/(1 + e^{-t})$. In this paper, we restrict to the case that σ belongs to a class of polynomially bounded differentiable functions.

Definition 2. *The set of polynomially bounded maps with continuous derivatives up to order $r \in \mathbb{N}_0$ is given by*

$$
C_{\text{PB}}^r(\mathbb{R}) = \Big\{ \sigma \in C^r(\mathbb{R}) \big| \forall l = 0, \ldots, r \exists k_l > 0 : \tag{2}
$$
$$
\sup_{x \in \mathbb{R}} |\sigma^{(l)}(x)(1 + x^2)^{-k_l}| < \infty \Big\}.
$$

Note that the linear activation function, as well as the sigmoid activation function, both belong to $C_{PB}^r(\mathbb{R})$ for any $r \in \mathbb{N}_0$. Also, any polynomial activation function $P(x) \in \mathbb{R}[x]$ belongs to $C_{\text{PB}}^{\deg(P)}(\mathbb{R})$. However, the ReLU activation function is not in $C_{\text{PB}}^r(\mathbb{R})$ for any $r \in \mathbb{N}_0$.

2.2. Feedforward–Backpropagation, and SGD

Let $(X, Y) : \Omega \to \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which follows a distribution μ. A NN is typically used to predict output Y given input X. So ideally, the NN is operated using weights in the set $\arg\min_W \mathcal{U}(W)$ where

$$\mathcal{U}(W) \triangleq \int l(\Psi_W(x), y) \, \mathrm{d}\mathbb{P}[(X, Y) = (x, y)] \quad (3)$$

is called the *risk function*, and $l : \mathbb{R}^{d_L} \to [0, \infty)$ is a convex loss function of one's choice. Throughout this article, we will specify the Euclidean ℓ_2-norm $l(x, y) \triangleq \|x - y\|_2^2$ as our loss function of interest. In this paper, we make no distinction between an oracle risk function or empirical risk function. Both situations are covered by (3), which can be seen by choosing the distribution appropriately. What we theoretically do assume is that one has the ability to repeatedly draw independent and identically distributed samples from μ. Hence, the results include the empirical risk case (where μ has finite support) as well as the online learning case as particular cases.

In an attempt to find a critical point in the set $\arg\min_W \mathcal{U}(W)$, the Feedforward–Backpropagation (FB) algorithm is commonly used in NN training. It is a recursive stochastic algorithm and works as follows. Let $\{(Y^{[t]}, X^{[t]})\}_{t \in \mathbb{N}_+}$ be a sequence of independent copies of (X, Y), let $W^{[0]} \in \mathcal{W}$ be an arbitrary nonrandom initialization of the weights. For $i = 1, \ldots, L, r = 1, \ldots, d_{i+1}$, $l = 1, \ldots, d_i$, FB is used iteratively by updating

$$W_{i,r,l}^{[t+1]} = W_{i,r,l}^{[t]} - \alpha^{\{t+1\}} \big(\mathrm{FB}_{W^{[t]}}(X^{[t+1]}, Y^{[t+1]})\big)_{i,r,l}$$
$$(4)$$

for $t = 0, 1, 2$, etc. Here $\{\alpha^{\{t\}}\}_{t \in \mathbb{N}_+}$ denotes a positive, so $\alpha^{\{t\}} \geq 0 \,\forall t \in \mathbb{N}_+$, deterministic step size sequence, and the *FB* step of the algorithm is as follows:

Definition 3. *Assume* $\sigma \in C^1(\mathbb{R})$. *Given weights* $W \in \mathcal{W}$ *and input–output pair* $(x, y) \in \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$, *the tensor* $\mathrm{FB}_W(x, y) \in \mathbb{R}^{d_L \times d_{L-1}} \times \cdots \times \mathbb{R}^{d_1 \times d_0}$ *is calculated iteratively by:*

1. *Computing* $A_1, \ldots, A_L$ *using Definition 1.*
2. *Calculating for* $i = L - 1, \ldots, 1$,

$$R_L = A_L = (y - W_L A_{L-1}) \in \mathbb{R}^{d_L},$$
$$R_i = (W_{i+1}^{\mathrm{T}} R_{i+1}) \odot (\sigma'(W_i A_{i-1})) \in \mathbb{R}^{d_i}. \quad (5)$$

3. *Setting for* $i \in [L]$, $\big(\mathrm{FB}_W(x, y)\big)_i = -2R_i A_{i-1}^{\mathrm{T}}$.

The algorithm in (4) is a step in a SGD algorithm. To see this, note that since $\sigma \in C^1(\mathbb{R})$ by assumption,

$$\big(\mathrm{FB}_W(x, y)\big)_{i,r,l} = \frac{\partial l(\Psi_W(x), y)}{\partial W_{i,r,l}} \quad (6)$$

holds. By substituting (6) into (4), one identifies the SGD algorithm. Under additional assumptions on the distribution of (X, Y) and by linearity of the expectation and gradient operators, one can then furthermore see that $\mathbb{E}[\big(\mathrm{FB}_W(X, Y)\big)_{i,r,l}] = \partial \mathcal{U}(W)/\partial W_{i,r,l} = (\nabla U)_{i,r,l}$, which suggests that $W^{[t]}$ may converge to the crtical set $\{W | \nabla \mathcal{U}(W) = 0\}$. Because (3) is not convex, there is no guarantee that the iterates $W^{[t]}$ in (4) converge to a point in $\arg\min_W \mathcal{U}(W)$. Nonetheless, the surprising success of (4) in NNs is still a key question in the field of machine learning; as a starting point an interested reader may look at e.g. (Gunasekar et al., 2017). Lastly, note that Definition 3 is a computationally efficient manner of calculating $\nabla l(\Psi_W(x), y)$. It is essentially a recursive computation of the partial derivatives which leverages the NN's layered structure together with the chain rule of differentation.

2.3. Dropout algorithms, and their risk functions

Dropout algorithms are stochastic training algorithms to avoid overfitting in NNs. These algorithms work by applying $\{0, 1\}$-valued random matrices as masks in the weights during the FB step. More precisely, we examine the following class of dropout algorithms. Let $(F, X, Y) : \Omega \to \{0, 1\}^{d_L \times d_{L-1}} \times \ldots \times \{0, 1\}^{d_1 \times d_0} \times \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ be a random variable on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Here, we write $F = (F_L, \ldots, F_1)$ and $F_{i+1} \in \{0, 1\}^{d_{i+1} \times d_i}$ for $i = 0, \ldots, L - 1$, similar to how we notate weight matrices. Let $\{(F^{[t]}, X^{[t]}, Y^{[t]})\}_{t \in \mathbb{N}_+}$ be a sequence of independent copies of (F, X, Y). In tensor notation, the weights are updated iteratively by setting

$$W^{[t+1]} = W^{[t]} - \alpha^{\{t+1\}} \Delta^{[t+1]} \quad (7)$$

for $t = 0, 1, 2$, etc, where the random direction

$$\Delta^{[t+1]} \triangleq F^{[t+1]} \odot \mathrm{FB}_{F^{[t+1]} \odot W^{[t]}}(X^{[t+1]}, Y^{[t+1]}). \quad (8)$$

Note in particular that if $F_{i,r,l}^{[t+1]} = 0$ for some i, r, l, then $\Delta_{i,r,l}^{[t]} = 0$. In other words, masked variables are not updated with these dropout algorithms.

The update rule (7) together with (8) describes different variants of dropout algorithms. In canonical *Dropout* (Hinton et al., 2012), for example, $F_{i,r,l'} = F_{i,r,l} \sim$ Bernoulli(p) for any $l, l' \in [d_i]$ with $p = 1/2$. In *Dropconnect* (Wan et al., 2013), $F_{i,r,l} \sim$ Bernoulli(p) for all i, r, l with $p = 1/2$. In *Cutout* (DeVries & Taylor, 2017), $F_{1,r,l} = 0$ whenever $|r - S_1| < c, c \in \mathbb{N}_+$ and $|l - S_2| < c$ with $(S_1, S_2) \sim$ Uniform$([d_1] \times [d_0])$. In fact, the class of dropout algorithms we consider is quite broad. For example, $F^{[t]}$ need not be independent of $(X^{[t]}, Y^{[t]})$, nor does $F_i^{[t]}$ need to have the same distribution as $F_j^{[t]}$ for $i \neq j$.

We call

$$\mathcal{D}(W) \triangleq \int l(\Psi_{f \odot W}(x), y) \, d\mathbb{P}[(F, X, Y) = (f, x, y)] \tag{9}$$

the *dropout algorithm's risk function*. If $F^{[t]}$ is independent of $(X^{[t]}, Y^{[t]})$ for each $t \in \mathbb{N}_0$, and Ω is countable, then the dropout algorithm's risk function simplifies to $\mathcal{D}(W) = \sum_f \mathbb{P}[F = f] \sum_{x,y} l(\Psi_{f \odot W}(x), y) \mathbb{P}[(X, Y) = (x, y)]$, where the sums are over all possible outcomes of the random variables F and (X, Y), respectively.

3. Almost sure convergence of projected dropout algorithms

Our first result pertains to projected dropout algorithms. Let $\mathcal{H} \subseteq \mathcal{W}$ be a convex compact nonempty set and let $P_{\mathcal{H}} : \mathcal{W} \to \mathcal{H}$ be the projection onto $\mathcal{H}$. By compactness and convexity of $\mathcal{H}$, the projection is unique. In a projected dropout algorithm, the weight update in (7) is replaced by

$$W_i^{[t+1]} = P_{\mathcal{H}}(W_i^{[t]} - \alpha^{[t+1]} \Delta_i^{[t+1]}) \text{ for } t \in \mathbb{N}_0. \tag{10}$$

We assume that $\mathcal{H}$ is defined by smooth constraints $\mathcal{H} = \{W \in \mathcal{W} \mid q_i(W) \le 0 \ \forall i \in [l]\}$.

Denote by $\nabla \mathcal{D}|_{\mathcal{H}}(W)$ the gradient of $\mathcal{D}(W)$ restricted to $\mathcal{H}$ and let $T_W \mathcal{W}$ be the tangent space of $\mathcal{W}$ at W. Suppose that $\nabla q_i(W) \ne 0$ whenever $q_i(W) = 0$ and, that these are linearly independent. At any point $W \in \partial \mathcal{H}$, we define the outer normal cone

$$C(W) \triangleq \{v \in T_W \mathcal{W} \mid \nabla q_i(W) v^T \ge 0 \tag{11}$$
$$\text{for } i \in [l] \text{ s.t. } q_i(W) = 0\}$$

We also assume that $C(W)$ is upper semicontinuous, i.e., if $\tilde{W} \in B_{\mathcal{H}}(W, \delta)$, where $B_{\mathcal{H}}(W, \delta)$ is the ball of radius $\delta > 0$ centered at W and intersected with $\mathcal{H}$, then $C(W) = \cap_{\delta > 0}(\cup_{\tilde{W} \in B_{\mathcal{H}}(W, \delta)} C(\tilde{W}))$. Let $\pi(W) \triangleq -t\mathbb{1}[W \in \partial \mathcal{H}]$ with $t \in C(W)$ minimal to resolve the violated constraints of $\mathcal{D}|_{\mathcal{H}}(W)$ at $W \in \partial \mathcal{H}$ so that $\mathcal{D}|_{\mathcal{H}}(W) + \pi(W)$ points inside $\mathcal{H}$. In particular, we have $\pi(W) = -\sum_{i=1}^l \lambda_i(W) \nabla q_i(W) \in -C(W)$ where $\{\lambda_i(W) \ge 0\}_{i=1}^l$ are functions such that $\lambda_i(W) = 0$ if $q_i(W) < 0$.

Finally, define the set of stationary points $S_{\mathcal{H}} \triangleq \{W \in \mathcal{H} \mid \nabla \mathcal{D}|_{\mathcal{H}}(W) + \pi(W) = 0\}$. The set $S_{\mathcal{H}}$ can be divided into disjoint compact and connected subsets $S_1, \ldots, S_r, \ldots$.

We are now in position to state our first result:

Proposition 1. *Assume that: (N1)* $\sigma \in C_{\mathrm{PB}}^2(\mathbb{R})$, *(N2)* $\mathbb{E}[\|Y\|_2^m \|X\|_2^n] < \infty \ \forall m \in \{0, 1, 2\}, n \in \mathbb{N}_0$, *(N3)*

the random variables $\{(F^{[s]}, X^{[s]}, Y^{[s]})\}_{s \in \mathbb{N}_+}$ *are independent copies of* (F, X, Y), *and*

$$(A2.4) \quad \sum_{t=1}^{\infty} \alpha^{\{t\}} = \infty, \quad \sum_{t=1}^{\infty} (\alpha^{\{t\}})^2 < \infty. \tag{12}$$

Let $\{W^{[t]}\}_{t \in \mathbb{N}_0}$ *be the sequence of random variables generated by* (10) *with* (8). *Then, there is a set N of probability zero such that for* $\omega \notin N$, $\{W^{[t]}(\omega)\}$ *converges to a limit set of the projected ODE*

$$\frac{dW}{dt} = \nabla_W \mathcal{D}|_{\mathcal{H}}(W) + \pi(W). \tag{13}$$

Moreover, if (N4) $\sigma \in C_{\mathrm{PB}}^r(\mathbb{R})$, *with* $\dim(\mathcal{W}) \le r$, *(N5)* $\nabla_W \mathcal{D}|_{\mathcal{H}}(W) + \pi(W) \ne 0$ *whenever* $\nabla_W \mathcal{D}|_{\mathcal{H}}(W) \ne 0$, *then for almost all* $\omega \in \Omega$, $\{W^{[t]}(\omega)\}_{t \in \mathbb{N}}$ *converges to a unique point in* $\{W \in \mathcal{H} \mid \nabla \mathcal{D}|_{\mathcal{H}}(W) = 0\}$.

Proof outline. The proof of Proposition 1 relies on the framework of stochastic approximation in (Kushner & Yin, 2003). Specifically, Proposition 1 follows from Theorem 2.3 on p. 127 if we can show that its conditions (A2.1)–(A2.6) on p. 126 are satisfied. We verify these conditions in Lemmas 1–3, which we explain next. For the derivations, see Appendix A.

First we assume conditions (N1), (N2), (N3) and we prove that the variance of the random update direction in (8) is finite, which verifies (A2.1).

Lemma 1. *Assume (N1)–(N3) from Proposition 1. Then* $\sup_{t \in \mathbb{N}} \mathbb{E}[\|\Delta_i^{[t+1]}\|_F^2] < \infty$ *for* $i = 0, 1, \ldots, L$.

We prove next that if $\sigma \in C_{PB}^r(\mathbb{R})$, then the random update direction in (8), conditional on all prior updates, has conditional expectation $\nabla \mathcal{D}(W^{[t]})$. Lemma 2 verifies (A2.2), (A2.3), and (A2.5):

Lemma 2. *Assume (N2)–(N4) from Proposition 1. Then* $\mathbb{E}[\Delta^{[t+1]} | \mathcal{F}_t] = \nabla \mathcal{D}(W^{[t]})$. *Furthermore,* $\nabla \mathcal{D} : \mathcal{W} \to \mathcal{W}$ *is* $r - 1$ *times continuously differentiable.*

From these conditions the first part of Proposition 1 follows. To prove the second part of Proposition 1, we have to prove that the set of stationary points $S_{\mathcal{H}}$ is well-behaved in the sense that $\mathcal{D}|_{S_i}(W)$ is constant. If an objective function is sufficiently differentiable, this is guaranteed by the Morse–Sard Theorem (Morse, 1939; Sard, 1942). In the present case however we must take into account the possibility of an intersection of the set of stationary points with the boundary $\partial \mathcal{H}$. Assuming (N4) and (N5) provides sufficient conditions.

Lemma 3. *If (N2)–(N5) hold, then* $\mathcal{D}(W)$ *is constant on each* S_i.

Discussion. The previous theorem guarantees that the class of dropout algorithms we consider converges. In

particular, in Proposition 1 the dependence structure of (F, X, Y) as random variables is not restricted and includes commonly used dropout algorithms such as those used in (Hinton et al., 2012; Wan et al., 2013). Furthermore, we allow for a general class of differentiable activation functions. Proposition 1 includes also online and offline learning, depending on which distribution we sample (X, Y) from.

Examining Proposition 1 critically, note that it does not give insight into the convergence rate or the precise stationary point of $\mathcal{D}(W)$ that the iterates converge to. Hence, we consider Proposition 1 as a first step to understand the convergence properties of dropout algorithms. Also, a caveat of the projection in (10) is that if we drop assumption (N5) from Proposition 1, then spurious stationary points may appear in case $\nabla \mathcal{D}(W) \in C(W)$ for some $W \in \partial \mathcal{H}$ or some critical points lie outside $\mathcal{H}$. However, we may be able to avoid this issue by using a stochastic projection set $\mathcal{H}^{[t]}$ (Chen, 2006; Borkar, 2009) or, in practice, just take $\mathcal{H}$ large enough in order to avoid spurious stationary points.

Since the class $C^r_{\mathrm{PB}}(\mathbb{R})$ contains polynomials of any degree, we can also approximate the case where σ is continuous and piecewise smooth, like for example $\mathrm{ReLU}(x) = \max(0, x)$, in the case that the data (X, Y) has compact support. Then, (W, F, X, Y) lie in a compact set so by Weierstrass approximation theorem we can find a sequence $\{\sigma_n\}_n \subset C^r_{\mathrm{PB}}(\mathbb{R})$ such that $\sigma_n \to \sigma$ uniformly in some compact set $K_n \subset \mathbb{R}$ where K_n does not include the discontinuities of σ'. Then $|(\Psi_n)_W(x) - \Psi_W(x)| \to 0$ uniformly in $\mathcal{H} \times \mathrm{supp}(X)$ and $|\nabla(\Psi_n)_W(x) - \nabla \Psi_W(x)| \to 0$ uniformly in $\mathcal{H} \backslash \mathcal{K}_n$, where $\mathcal{K}_n$ is such that we avoid the discontinuities of $\nabla \Psi_W(x)$. In particular the minima of $\mathcal{D}_n(W)$ will be close to the global minima of $\mathcal{D}(W)$. We expect that then, an asymptotic analysis in the case that σ is not differentiable can be carried out with Projected Stochastic Subgradient Descent, which we leave for follow-up research.

4. Convergence rate of GD on $\mathcal{D}(W)$ for arborescences with linear activation

Our second result in this paper pertains to the following regular GD algorithm on a dropout algorithm's risk function:

$$W^{\{t+1\}} = W^{\{t\}} - \alpha \nabla \mathcal{D}(W^{\{t\}}) \quad \text{for} \quad t \in \mathbb{N}_0. \quad (14)$$

Here, we keep the step size $\alpha > 0$ fixed. Note that this algorithm generates a deterministic sequence $\{W^{\{t\}}\}_{t \in \mathbb{N}_0}$ as opposed to a sequence of random variables $\{W^{[t]}\}_{t \in \mathbb{N}_0}$ as generated by (7), (8).

We first give an explicit characterization of a dropout algorithm's risk function (9) in terms of paths in a graph that holds for NNs with linear activation functions. Consider a fixed, directed *base graph* $G = (\mathcal{E}, \mathcal{V})$ without cycles and

in which all paths have length L, which describes a NN's structure as follows. Each vertex $v \in \mathcal{V}$ represents a neuron of the NN, and each directed edge $e = (u, v) \in \mathcal{E}$ indicates that neuron u's output is input to neuron v. Let $\mathcal{G}$ be the set of all subgraphs of the base graph G, and let $\mathcal{E}(g)$ be the set of edges of subgraph $g \in \mathcal{G}$. Let $\Gamma^j_i(g; e)$ be defined as the set of all length-L paths in graph g that start at vertex i, traverse edge e, and end at vertex j. Whenever one of the latter three conditions is not needed, the subscript, argument, or superscript is dropped from the notation, respectively. Note that to each edge $e \in \mathcal{E}$ in the NN, a weight $W_e \in \mathbb{R}$ and a mask variable $F_e \in \{0, 1\}$ are associated. We can write $\mathcal{W} = \mathbb{R}^{|\mathcal{E}|}$ also. For every path $\gamma \triangleq (\gamma_1, \dots, \gamma_L) \in \Gamma(g)$, we write $P_\gamma \triangleq \prod_{e \in \gamma} W_e$ and $F_\gamma \triangleq \prod_{e \in \gamma} F_e$ for notational convenience. Finally, let $G_F \triangleq (\mathcal{E}_F, \mathcal{V})$ be the random subgraph of base graph G that has edge set $\mathcal{E}_F \triangleq \{e \in \mathcal{E} | F_e = 1\}$. We denote $\mu_g \triangleq \mathbb{P}[G_F = g]$, and $\eta_\gamma \triangleq \sum_{\{g \in \mathcal{G} | \gamma \in \Gamma(g)\}} \mu_g$. The next lemma now holds, whose proof is in Appendix B.1:

Lemma 4. *Assume (N6') that the base graph G is a fixed, directed graph without cycles in which all paths have length L, (N7) that $\sigma(t) = t$, and (N8) that F is independent of (X, Y). Then*

$$\mathcal{D}(W) = \sum_{g \in \mathcal{G}} \mu_g \mathbb{E}\Big[\sum_{e=1}^{d_L}\Big(Y_e - \sum_{\gamma \in \Gamma^e(g)} P_\gamma X_{\gamma_0}\Big)^2\Big]. \quad (15)$$

Moreover $\mathcal{D}(W) = \mathcal{J}(W) + \mathcal{R}(W)$, where

$$\mathcal{J}(W) = \sum_{\gamma \in \Gamma(G)} \eta_\gamma \mathbb{E}[(Y_{\gamma_L} - P_\gamma X_{\gamma_0})^2], \quad (16)$$

$$\mathcal{R}(W) = -\sum_{g \in \mathcal{G}} \mu_g \mathbb{E}\Big[\sum_{e=1}^{d_L} \sum_{\gamma \in \Gamma^e(g)} \Big(\Big(1 - \frac{1}{|\Gamma^e(g)|}\Big)Y_e^2$$
$$- P_\gamma X_{\gamma_0} \sum_{\delta \in \Gamma^e(g) \backslash \{\gamma\}} P_\delta X_{\delta_0}\Big)\Big]. \quad (17)$$

For example in the case of *Dropconnect* (Wan et al., 2013), where the masking variables $\{F_e\}_{e \in \mathcal{E}}$ are independently and identically distributed $\mathrm{Bernoulli}(p)$ random variables, Lemma 4 holds with $\mu_g = p^{|\mathcal{E}(g)|}(1-p)^{|\mathcal{E}(G)| - |\mathcal{E}(g)|}$. Also note that if $|\Gamma^i(g)| = 1 \forall g \in \mathcal{G}, i \in [d]$, such as when G is an arborescence, then $\Gamma^{\gamma_L}(g) = \{\gamma\} \forall g \in \mathcal{G}, \gamma \in \Gamma(g)$ and consequently $\mathcal{R}(W) = 0$.

We now focus on a base graph that is an arborescence, see Figure 2. We can then explicitly compute an upper bound to the convergence rate of (14) in case $\sigma(z) = z$. To that end, we first prove the following specification of Lemma 1.

Corollary 1. *Assume (N6) that the base graph G is an arborescence of depth L, and (N7)–(N8) from Lemma 4. Then $\mathcal{D}(W) = \mathcal{I}(W) + \mathcal{D}(W^{\mathrm{opt}})$, where $\mathcal{I}(W) \triangleq \sum_{\gamma \in \Gamma(G)} \nu_\gamma (z_\gamma - P_\gamma)^2$, $\mathcal{D}(W^{\mathrm{opt}}) =$*
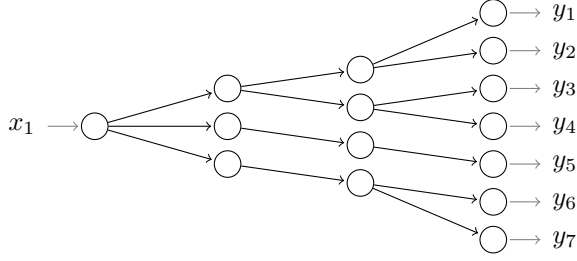
Figure 2. An example arborescence of depth $L = 3$.

$\sum_{\gamma \in \Gamma(G)} \eta_\gamma (\mathbb{E}[Y_{\gamma_L}^2] - \mathbb{E}[Y_{\gamma_L} X_{\gamma_0}]^2 / \mathbb{E}[X_{\gamma_0}^2])$, *and* $\nu_\gamma \triangleq \eta_\gamma \mathbb{E}[X_{\gamma_0}^2]$, $z_\gamma \triangleq \mathbb{E}[Y_{\gamma_L} X_{\gamma_0}]/\mathbb{E}[X_{\gamma_0}^2]$ *for* $\gamma \in \Gamma(G)$.

To derive an upper bound on the convergence rate, we use that for the system of [ODEs](#) $dW/dt = -\nabla \mathcal{D}(W)$, there are conserved quantities. Specifically, let $\mathcal{L}(g; f)$ denote the leafs of the subtree of $g \in \mathcal{G}$ rooted at $f \in \mathcal{E}(g)$ and $\mathcal{L}(G) \triangleq \cup_{f \in \mathcal{E}} \mathcal{L}(G; f)$. Define

$$C_f = C_f(W) \triangleq W_f^2 - \sum_{l \in \mathcal{L}(G;f)} W_l^2 \text{ for } f \in \mathcal{E} \backslash \mathcal{L}(G), \tag{18}$$

for $W \in \mathcal{W}$. We also define $C_{\min} \triangleq \min_{e \in \mathcal{E} \backslash \mathcal{L}(G)} C_e$, and the sequence $C_e^{\{t\}} = C_e(W^{\{t\}})$ for $t \in \mathbb{N}_+$ which we require later. For the function C_f in (18), we can prove Lemma 5, the proof of which is in Appendix [B.2](#).

Lemma 5. *Assume (N2) from Proposition [1](#), and (N6')–(N8) from Lemma [4](#). Then under the negative gradient flow* $dW/dt = -\nabla \mathcal{D}(W)$, $dC_f/dt = 0$ *for all* $f \in \mathcal{E} \backslash \mathcal{L}(G)$.

We are almost in position to state our second result, and will still benefit from more notation. We define $\|\nu\|_1 \triangleq \sum_{\gamma \in \Gamma(G)} \nu_\gamma$ and $\nu_{\max} \triangleq \min_{\gamma \in \Gamma(G)} \nu_\gamma$. For $0 < \delta < M$ we define $\mathcal{S} \triangleq \{W \in \mathcal{W} | M > |W_f| > \delta > 0 \ \forall f \in \mathcal{E}(G) \backslash \mathcal{L}(G); M > |W_f| \ \forall f \in \mathcal{L}(G)\}$. We also define the intervals $I_f \triangleq [C_f^{\{0\}}/2, 3C_f^{\{0\}}/2]$ for $f \in \mathcal{E} \backslash \mathcal{L}(G)$ as well as the set $I \triangleq \times_{f \in \mathcal{E} \backslash \mathcal{L}(G)} I_f \subseteq \mathbb{R}^{|\mathcal{E}| - |\mathcal{L}(G)|}$. Let

$$B(\epsilon, I) \triangleq \{W \in \mathcal{W} | \mathcal{I}(W) \le \epsilon, \tag{19}$$
$$W_f^2 - \sum_{l \in \mathcal{L}(G;f)} W_l^2 \in I_f \text{ for } f \in \mathcal{E} \backslash \mathcal{L}(G)\}.$$

The following proposition now holds, and its proof can be found in Appendix [B.6](#).

Proposition 2. *Assume (N2) from Proposition [1](#), (N6) from Corollary [1](#), (N7)–(N8) from Lemma [4](#), (N9) that* $W^{\{0\}} \in \mathcal{S} \cap B(\epsilon, I)$ *and* $M^L \ge |z_\gamma|$ *for all* $\gamma \in \Gamma(G)$, *and (N10) that* $\frac{1}{2} C_{\min}(W^{\{0\}}) > \delta^2$. *Then if*

$$\alpha \le \min\Big(\nu_{\min} \frac{e^{-1/2}(C_{\min}^{\{0\}})^L}{8 \|\nu\|_1 (2L-1) M^{2(L-1)} \mathcal{I}(W^{\{0\}})}, \tag{20}$$

$$\frac{1}{12\nu_{\max} |\Gamma(G)| M^{2(L-1)}}, \frac{1}{2\nu_{\min}(C_{\min}^{\{0\}})^{L-1}}\Big),$$

the iterates of (14) *will satisfy*

$$\mathcal{D}(W^{\{t\}}) - \mathcal{D}(W^{\text{opt}}) \le (\mathcal{D}(W^{\{t\}}) - \mathcal{D}(W^{\text{opt}}))e^{-\frac{\alpha\tau}{2}t}. \tag{21}$$

where $\tau = 4\nu_{\min}e^{-1/2}(C_{\min}^{\{0\}})^{L-1}$.

Assumptions (N9)–(N10) are satisfied, for example, when initializing $M > W_e^{\{0\}} > \sqrt{2}\delta$ for $e \in \mathcal{E} \backslash \mathcal{L}(G)$ and setting $|W_l| \le \delta/\sqrt{|\mathcal{L}(G)|}$ for all $l \in \mathcal{L}(G)$ and $\epsilon = \mathcal{I}(W^{\{0\}})$.

For additional insight, we provide the following specification of Proposition [2](#) for the case of *Dropconnect* ([Wan et al., 2013](#)). The proof of Corollary [2](#) is deferred to Appendix [B.7](#).

Corollary 2. *Under the assumptions of Proposition [2](#), if additionally* $\{F_e\}_{e \in \mathcal{E}}$ *are independent and identically distributed* Bernoulli(p) *random variables, then* $\nu_{\min} = \nu_{\max} = \mathbb{E}[X^2]p^L$ *and* $\nu_{\min}/\|\nu\|_1 = 1/d_L$. *If* α *satisfies* (20) *the iterates in* (14) *satisfy* (21) *with* $\alpha\tau = O((p^L(C_{\min}^{\{0\}})^{2L})/(L(d_L)^2(M^2)^{2L}))$.

Proof outline. The proof of Proposition [2](#) is by double induction on the statements $A(t) \equiv \{\mathcal{I}(W^{\{s\}}) \le \mathcal{I}(W^{\{s-1\}})e^{-2\nu_{\min}\kappa\alpha}, \forall s \in [t]\}$ and $B(t) \equiv \{W^{\{s\}} \in K, \forall s \in [t]\}$ where $\kappa > 0$ is a free parameter. Concretely, we prove that there exist α and κ such that $A(t) \cap B(t) \Rightarrow B(t+1)$ and $A(t) \cap B(t+1) \Rightarrow A(t+1)$. Appendix [B.6](#) describes in detail how the upcoming Lemmas [6](#)–[8](#) provide sufficient conditions for the induction step. There we also maximize the upper bound on the convergence rate over κ, which gives the rate in (20).

Lemma [6](#) implies that $B(\epsilon, I)$ is compact and that $\mathcal{D}(W)$ is β-smooth on the compact set $K = \mathcal{S} \cap B(\epsilon, I)$, i.e., $\mathcal{D}(W') - \mathcal{D}(W) \le \nabla \mathcal{D}(W)^{\mathrm{T}}(W' - W) + \beta \|W' - W\|_2^2$ for $W, W' \in K$. Its proof is deferred to Appendix [B.3](#). Here, with a minor abuse of notation, we define also

$$B(\epsilon, \{C_f\}_{f \in \mathcal{E} \backslash \mathcal{L}(G)}) \triangleq \{W \in \mathcal{W} | \mathcal{I}(W) \le \epsilon, \tag{22}$$
$$W_f^2 - \sum_{l \in \mathcal{L}(G;f)} W_{\gamma^l}^2 = C_f\}$$

where $\{\gamma^l\} \triangleq \Gamma^l(G)$ for $l \in \mathcal{L}(G)$ if G is an arborescence.

Lemma 6. *Assume (N2) from Proposition [1](#), and (N6) from Corollary [1](#). Then:*

(i) *If* $\epsilon > 0$ *and* $|C_f| < \infty$ *for* $f \in \mathcal{E} \backslash \mathcal{L}(G)$, *then the set* $B(\epsilon, \{C_f\}_{f \in \mathcal{E} \backslash \mathcal{L}})$ *is compact.*

(ii) *If* $\max_{\gamma \in \Gamma(G)} |z_\gamma| \le M^L$, *then the function* $\mathcal{I}(W)$ *is* β-smooth in $\mathcal{S}$ with $\beta = 6\nu_{\max} |\Gamma(G)| M^{2(L-1)}$.

Next, Lemma 7 gives a lower bound on the curvature of $\mathcal{D}(W)$ on K in the direction of $\nabla \mathcal{D}(W)$, in the form of a Polyak–Łojasiewicz (PL)-inequality (Karimi et al., 2016). Its proof is in Appendix B.4.

Lemma 7. *Assume (N2) from Proposition 1, and (N6) from Corollary 1. If $W^{\{t\}} \in \mathcal{S} \cap B(\epsilon, I)$, then $\|\nabla \mathcal{D}(W^{\{t\}})\|_2^2 \geq 4\nu_{\min}(C_{\min}^{\{t\}})^{(L-1)}\big(\mathcal{D}(W^{\{t\}}) - \mathcal{D}(W^{\text{opt}})\big)$.*

Lemma 8 proves that the conserved quantities of the gradient flow remain bounded under the GD algorithm in (14). This lemma allows us to keep track of the iterates in the compact set $K = \mathcal{S} \cap B(\epsilon, I)$ by relating them to conserved quantities and exploiting the fact that under GD, $|C_f^{\{t+1\}} - C_f^{\{t\}}|$ has order $O(\alpha^2)$. Appendix B.4 contains its proof.

Lemma 8. *Assume (N2) from Proposition 1, and (N6) from Corollary 1. If $W^{\{t\}} \in \mathcal{S}$, and $C_f^{\{t\}} > 0$ for all $f \in \mathcal{E} \backslash \mathcal{L}(G)$, then $4\alpha^2 \|\nu\|_1 M^{2(L-1)}\big(\mathcal{D}(W^{\{t\}}) - \mathcal{D}(W^{\text{opt}})\big) \geq |C_f^{\{t+1\}} - C_f^{\{t\}}|$.*

Discussion. A dropout algorithm's regularization properties depend on the base graph G as well as the distribution on F. The regularization contribution to a dropout's risk function is explicitly given in our path representation in Lemma 4 by the term $\mathcal{R}(W)$. Note that if the base graph is an arborescence, then $\mathcal{R}(W) = 0$. This implies that the minimum of $\mathcal{D}(W)$ satisfies $z_\gamma = P_\gamma$ for all $\gamma \in \Gamma(G)$ in Corollary 1. We note here also that when we consider an *anti-arborescence*, the term $\mathcal{R}(W)$ does not vanish. This suggests that when the input vector has a higher dimension than the output—when information gets compressed—dropout algorithms can have increased regularization.

Observe in Corollary 2 that in the case of *Dropconnect* (Wan et al., 2013), the convergence rate depends on p^L and $(C_{\min}^{\{0\}}/M^2)^{2L}$ where $C_{\min}^{\{0\}}/M^2 < 1$. First, this shows the increased difficulty of training NNs as they become deeper, also seen in other convergence results e.g. in (Shamir, 2018) and (Arora et al., 2018). The exponential dependence in L is moreover tight when using GD and is intrinsic to the method (Shamir, 2018). Second, note that p^L can be understood as the probability that $F_\gamma = 1$ for any fixed γ when using *Dropconnect*. This term indicates that as NNs become deeper, and the probability of masking an edge is increased, the convergence rate of GD with dropout will decrease exponentially depending on the probability p.

The bound on the convergence rate in Corollary 2 for *Dropconnect* is strictly decreasing in p. However, this may not be true for other NN configurations, where $\mathcal{R}(W) \neq 0$ may induce a more complex dependence of the convergence rate on p. In fact, there could be some optimal p^* which depends on the data. We expect that we can lift our results to the stochastic dropout algorithm, in cases where in each iteration of the FB algorithm with dropout, only a subset of edges are updated. Finally, given a NN topology and dropout algorithm with regularization term $\mathcal{R}(W)$, there may be *a different* NN topology and dropout algorithm (i.e., distribution on F) with regularization $\hat{\mathcal{R}}(W)$ that minimizes the runtime of the algorithm while *also* satisfying $|\min_W \hat{\mathcal{R}}(W) - \min_W \mathcal{R}(W)| < \epsilon$ for fair comparison.

5. Conclusion

This paper presented formal proof that a class of dropout algorithms for neural networks, when projected to a compact set, converge almost surely to a unique stationary set of a projected system of ODEs. The result gives formal guarantee that these dropout algorithms are well-behaved for a wide range of NNs and activation functions, and will at least asymptotically not suffer from percolative nature. Additionally, we established an upper bound on the rate of convergence of regular GD on the limiting ODE of dropout algorithms for arborescences of arbitrary depth with linear activation functions. While GD on the limiting ODE is not strictly a dropout algorithm, the result is a major and necessary step towards analyzing the convergence rate of actual ones. Besides providing insight into the optimization of NN using dropout algorithms, our results may be used to indicate what a NN configuration should look like in order to adjust the convergence rates for dropout.

As to future work, we see multiple directions:

– Theoretically, one may be interested in dropping the projection assumption. Proving a convergence result would then become substantially more challenging, because a notion of compactness must be proven first. We do not expect that we can exploit conserved quantities since we may not have such a symmetric optimization landscape. An alternative approach may be to use overparameterized networks as in (Zou et al., 2018). Then, we can fit the data with no training error and we expect to converge when initializing close to a global minimum. However, the regularization of a dropout algorithm may prevent the achievement zero training loss and hence, a first characterization of the global minima may be needed before obtaining a convergence rate.

– If Assumption (N5) is weakened, then spurious stationary points may occur on the boundary of $\mathcal{H}$. We expect convergence to such points to be unlikely if the compact set's size is chosen sufficiently large, but we have not proven such claim. Identifying the probability with which a projected dropout algorithm converges to such an artificial stationary point by generalizing techniques from e.g. (Dupuis & Kushner, 1985; 1989; Buche & Kushner, 2002), would be interesting future work, and an excursion into state-of-the-art mathematics of stochastic approximation.

– On arborescences, there is no regularization by Dropout algorithms. It would therefore be valuable to investigate the convergence rate of GD on the limiting ODE of dropout algorithms also on other graphs. The more paths exist within the base graph that overlap one another, the more strongly the weights depend on one another throughout the iterations. These dependencies complicate the analysis considerably. Also in case we have a *compressing* NN, the regularization will not vanish and the optimization landscape will become more complex.

References

Arora, S., Cohen, N., Golowich, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

Ba, J. and Frey, B. Adaptive dropout for training deep neural networks. In *Advances in neural information processing systems*, pp. 3084–3092, 2013.

Baldi, P. and Sadowski, P. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.

Baldi, P. and Sadowski, P. J. Understanding Dropout. In *Advances in neural information processing systems*, pp. 2814–2822, 2013.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pp. 3873–3881, 2016.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Broadbent, S. R. and Hammersley, J. M. Percolation processes: I. crystals and mazes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 629–641. Cambridge University Press, 1957.

Buche, R. and Kushner, H. J. Rate of convergence for constrained stochastic approximation algorithms. *SIAM journal on control and optimization*, 40(4):1011–1041, 2002.

Cavazza, J., Lane, C., Haeffele, B. D., Murino, V., and Vidal, R. An analysis of dropout for matrix factorization. *arXiv preprint arXiv:1710.03487*, 2017a.

Cavazza, J., Morerio, P., Haeffele, B., Lane, C., Murino, V., and Vidal, R. Dropout as a low-rank regularizer for matrix factorization. *arXiv preprint arXiv:1710.05092*, 2017b.

Chen, H.-F. *Stochastic approximation and its applications*, volume 64. Springer Science & Business Media, 2006.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Dupuis, P. and Kushner, H. J. Stochastic approximations via large deviations: Asymptotic properties. *SIAM journal on control and optimization*, 23(5):675–696, 1985.

Dupuis, P. and Kushner, H. J. Stochastic approximation and large deviations: Upper bounds and wp 1 convergence. *SIAM Journal on Control and Optimization*, 27(5):1108–1135, 1989.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

Hajłasz, P. Whitney's example by way of assouad's embedding. *Proceedings of the American Mathematical Society*, 131(11):3463–3467, 2003.

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

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Kay, E. and Agarwal, A. Dropconnected neural network trained with diverse features for classifying heart sounds. In *2016 Computing in Cardiology Conference (CinC)*, pp. 617–620. IEEE, 2016.

Kiefer, J., Wolfowitz, J., et al. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

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

Li, Z., Gong, B., and Yang, T. Improved dropout for shallow and deep learning. In *Advances in neural information processing systems*, pp. 2523–2531, 2016.

Mianjy, P. and Arora, R. On dropout and nuclear norm regularization. *arXiv preprint arXiv:1905.11887*, 2019.

Mianjy, P., Arora, R., and Vidal, R. On the implicit bias of dropout. *arXiv preprint arXiv:1806.09777*, 2018.

Morse, A. P. The behavior of a function on its critical set. *Annals of Mathematics*, pp. 62–70, 1939.

Pal, A., Lane, C., Vidal, R., and Haeffele, B. D. On the regularization properties of structured dropout. *arXiv preprint arXiv:1910.14186*, 2019.

Pham, V., Bluche, T., Kermorvant, C., and Louradour, J. Dropout improves recurrent neural networks for handwriting recognition. In *2014 14th International Conference on Frontiers in Handwriting Recognition*, pp. 285–290. IEEE, 2014.

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

Sard, A. The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society*, 48(12):883–890, 1942.

Semeniuta, S., Severyn, A., and Barth, E. Recurrent dropout without memory loss. *arXiv preprint arXiv:1603.05118*, 2016.

Shamir, O. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. *arXiv preprint arXiv:1809.08587*, 2018.

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

Urban, G., Bache, K., Phan, D. T., Sobrino, A., Shmakov, A. K., Hachey, S. J., Hughes, C. C., and Baldi, P. Deep learning for drug discovery and cancer research: Automated analysis of vascularization images. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3):1029–1035, 2018.

Wager, S., Wang, S., and Liang, P. S. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pp. 351–359, 2013.

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

Zaremba, W., Sutskever, I., and Vinyals, O. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

A. Proof of Proposition 1

Proposition 1 can be proven using the framework of (Kushner & Yin, 2003). Specifically, Proposition 1 follows from Theorem 2.3 in (Kushner & Yin, 2003). We will show that conditions (A2.1)–(A2.6) hold for Dropout SGD.

Preliminaries

We need to carefully track all sequences of random variables throughout this proof, so we repeat the definition of the class of dropout algorithms we consider here for your convenience.

Definition 4 (Dropout algorithms). *During its $(t+1)$-st* feedforward step*, the algorithm iteratively calculates*

$$A_0^{[t+1]} = X^{[t+1]},$$
$$A_i^{[t+1]} = \sigma((W_i^{[t]} \odot F_i^{[t+1]})A_{i-1}^{[t+1]}) \tag{23}$$

for $i = 1, 2, \ldots, L-1$, to output

$$\Psi_{F^{[t+1]} \odot W^{[t]}}(X^{[t+1]}) = (W_L^{[t]} \odot F_L^{[t+1]})A_{L-1}^{[t+1]} = A_L^{[t+1]}. \tag{24}$$

Subsequentially for its $(t+1)$-st backpropagation step *the algorithm calculates*

$$R_L^{[t+1]} = (Y^{[t+1]} - (W_L^{[t]} \odot F_L^{[t+1]})A_{L-1}^{[t+1]}) \in \mathbb{R}^{d_L},$$
$$R_j^{[t+1]} = ((W_{j+1}^{[t]} \odot F_{j+1}^{[t+1]})^T R_{j+1}^{[t+1]}) \odot (\sigma'((W_j^{[t]} \odot F_j^{[t+1]})A_{j-1}^{[t+1]})) \in \mathbb{R}^{d_i}, \tag{25}$$

iteratively for $j = L-1, \ldots, 1$. The algorithm then calculates

$$\Delta_i^{[t+1]} = -2F_i^{[t+1]} \odot (R_i^{[t+1]}(A_{i-1}^{[t+1]})^\mathrm{T}) \tag{26}$$

for $i = 1, \ldots, L$, and finally updates all weights according to (4).

We also start by proving a few useful bounds pertaining to the Frobenius norm, which we will later iterate.

Lemma 9. *For any matrix $A \in \mathbb{R}^{m \times n}$ and $1 \le k < \infty$, it holds that $\sum_{i,j}(1 + A_{ij}^2)^k \le nm(1 + \|A\|_\mathrm{F})^{2k}$. For any two matrices $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$ and $0 \le k < \infty$, it holds that $(1 + \|AB\|_\mathrm{F})^k \le (1 + \|A\|_\mathrm{F})^k(1 + \|B\|_\mathrm{F})^k$. For any two matrices $A, B \in \mathbb{R}^{n \times m}$, it holds that $\|A \odot B\|_\mathrm{F} \le \|A\|_\mathrm{F} \|B\|_\mathrm{F}$.*

Proof. Recall Minkowski's inequality for sequences; that is $\left(\sum_i |x_i + y_i|^k\right)^{1/k} \le \left(\sum_i |x_i|^k\right)^{1/k} + \left(\sum_i |y_i|^k\right)^{1/k}$, which holds for $1 \le k < \infty$. It (i) implies that for any matrix $A \in \mathbb{R}^{n \times m}$ and $1 \le k < \infty$, that

$$\sum_{i,j}(1 + A_{ij}^2)^k \overset{\text{(i)}}{\le} \left((nm)^{1/k} + \left(\sum_{i,j}|A_{i,j}^2|^k\right)^{1/k}\right)^k \overset{\text{(ii)}}{\le} nm\left(1 + \left(\sum_{i,j}|A_{i,j}^2|^k\right)^{1/k}\right)^k \tag{27}$$

where (ii) we have used that the function z^k is nondecreasing in $z \ge 0$ whenever $k \ge 0$. Because (iii) for the ℓ_k-norm for sequences it holds that $\|x\|_{2k}^2 \le \|x\|_2^2$ whenever $1 \le k < \infty$, we obtain

$$\sum_{i,j}(1 + A_{ij}^2)^k \overset{\text{(iii)}}{\le} nm(1 + \|A\|_\mathrm{F}^2)^k \overset{\text{(iv)}}{\le} nm(1 + \|A\|_\mathrm{F})^{2k} \tag{28}$$

where (iv) we have used that the function $(1 + z^2)^k \le (1 + z)^{2k}$ for all $z \ge 0$ whenever $k \ge 0$. This proves the first inequality.

The second inequality is an immediate consequence of the submultiplicativity property of the Frobenius norm and its positivity, i.e.,

$$1 + \|AB\|_\mathrm{F} \le 1 + \|A\|_\mathrm{F}\|B\|_\mathrm{F} \le 1 + \|A\|_\mathrm{F} + \|B\|_\mathrm{F} + \|A\|_\mathrm{F}\|B\|_\mathrm{F}. \tag{29}$$

Raising to the k-th power left and right finishes its proof.

The third inequality follows from strict positivity of the summands:

$$\|A \odot B\|_\mathrm{F}^2 = \sum_{i,j} A_{ij}^2 B_{ij}^2 \le \left(\sum_{i,j} A_{ij}^2\right)\left(\sum_{i,j} B_{ij}^2\right) = \|A\|_\mathrm{F}^2 \|B\|_\mathrm{F}^2. \tag{30}$$

Each of the inequalities has now been shown. $\qquad\square$

A.1. Boundedness of $\Delta^{[t+1]}$ in expectation – Proof of Lemma 1

The idea is to expand the terms in $\Delta_i^{[t+1]}$ defined in Definition 4 recursively, and identify a polynomial in variables $\{\|Y\|_2^n \|X\|_2^m\}_{m \in \mathbb{N}_0}$ and $n = 0, 1, 2$.

First, we will prove two bounds on the activation function applied to an arbitrary matrix A. Recall that $\sigma \in C_{PB}^2(\mathbb{R})$ by assumption (N1). There thus (i) exists some $C_0, k_0 > 0$ such that $|\sigma(z)| \leq C_0(1 + z^2)^{k_0}$ for all $z \in \mathbb{R}$, and there exists some $C_1, k_1 > 0$ such that $|\sigma'(z)| \leq C_1(1 + z^2)^{k_1}$ for all $z \in \mathbb{R}$. Let $k = \max\{1, k_0, k_1\}$. Then

$$\|\sigma(A)\|_{\mathrm{F}}^2 = \sum_{i,j} |\sigma(A_{ij})|^2 \overset{\text{(i)}}{\leq} C_0 \sum_{i,j}(1 + A_{ij}^2)^k \overset{(28)}{\leq} C_2(1 + \|A\|_{\mathrm{F}})^{2k} \tag{31}$$

for some constant $C_2 > 0$. Similarly there exists some $C_3 > 0$ such that $\|\sigma'(A)\|_F \leq C_3(1 + \|A\|_{\mathrm{F}})^k$. Note furthermore that for all $l \geq 0$, (ii) by submultiplicativity of the Frobenius norm,

$$(1 + \|A\sigma(B)\|_{\mathrm{F}})^l \overset{\text{(ii)}}{\leq} (1 + \|A\|_{\mathrm{F}}\|\sigma(B)\|_{\mathrm{F}})^l \overset{(31)}{\leq} \left(1 + C_2^{1/2}\|A\|_{\mathrm{F}}(1 + \|B\|_{\mathrm{F}})^k\right)^l \leq C_4(1 + \|A\|_{\mathrm{F}})^l(1 + \|B\|_{\mathrm{F}})^{kl} \tag{32}$$

for $C_4 = \max\{1, C_2^{l/2}\} > 0$. Again, a similar bound holds for σ'.

Next, note that we have by (i) submultiplicativity and (30) that

$$\|\Delta_i^{[t+1]}\|_{\mathrm{F}} = \|F_i^{[t+1]} \odot (R_i^{[t+1]}(A_{i-1}^{[t+1]})^{\mathrm{T}})\|_{\mathrm{F}} \overset{\text{(i)}}{\leq} \|F_i^{[t+1]}\|_{\mathrm{F}}\|R_i^{[t+1]}\|_{\mathrm{F}}\|A_{i-1}^{[t+1]}\|_{\mathrm{F}}. \tag{33}$$

The first term is bounded with probability one: $F_{i,r,l}^{[t]} \in \{0, 1\}$ for all i, r, l, t. For the second term, consider the following bound:

$$\|R_i^{[t+1]}\|_{\mathrm{F}} \overset{(25)}{=} \|(W_{i+1}^{[t]} \odot F_{i+1}^{[t+1]})^{\mathrm{T}} R_{i+1}^{[t+1]} \odot \sigma'((W_i^{[t]} \odot F_i^{[t+1]})A_{i-1}^{[t+1]})\|_{\mathrm{F}}$$

$$\overset{(30)}{\leq} \|W_{i+1}^{[t]} \odot F_{i+1}^{[t+1]}\|_{\mathrm{F}}\|\sigma'((W_i^{[t]} \odot F_i^{[t+1]})A_{i-1}^{[t+1]})\|_{\mathrm{F}}\|R_{i+1}^{[t+1]}\|_{\mathrm{F}} \tag{34}$$

for $1 \leq i \leq L$, where we have also used the submultiplicative property. For the third term, consider the next bound: (i) recursing (32) with $A = I$ and $B = (W_j^{[t]} \odot F_j^{[t+1]})A_{j-1}^{[t+1]}$ etc, we obtain that there exists some $C_5 > 0$, say, so that

$$\|A_j^{[t+1]}\|_{\mathrm{F}} \overset{(23)}{=} \|\sigma((W_j^{[t]} \odot F_j^{[t+1]})A_{j-1}^{[t+1]})\|_{\mathrm{F}} \overset{(31)}{\leq} C_2(1 + \|(W_j^{[t]} \odot F_j^{[t+1]})A_{j-1}^{[t+1]}\|_{\mathrm{F}})^k$$

$$\overset{(29)}{\leq} C_2(1 + \|W_j^{[t]} \odot F_j^{[t+1]}\|_{\mathrm{F}})^k(1 + \|A_{j-1}^{[t+1]}\|_{\mathrm{F}})^k \overset{\text{(i)}}{\leq} C_5(1 + \|X^{[t+1]}\|_2)^{k^j} \prod_{l=1}^{j-1}(1 + \|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}})^{k^{j-l}} \tag{35}$$

for $j = 1, 2, \ldots, L - 1$. Similar to the derivation in (35), we obtain instead with σ' that there exists some $C_6 > 0$ such that

$$\|\sigma'((W_j^{[t]} \odot F_j^{[t+1]})A_{j-1}^{[t+1]})\|_{\mathrm{F}} \leq C_6(1 + \|X^{[t+1]}\|_2)^{k^j} \prod_{l=1}^{j-1}(1 + \|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}})^{k^{j-l}}. \tag{36}$$

Recall that $\|\Delta_i^{[t+1]}\|_{\mathrm{F}} \leq \|F_i^{[t+1]}\|_{\mathrm{F}}\|R_i^{[t+1]}\|_{\mathrm{F}}\|A_{i-1}^{[t+1]}\|_{\mathrm{F}}$. This, together with using (34) repeatedly for $j = i, \ldots, L - 1$, and (35), (36), yields

$$\|\Delta_i^{[t+1]}\|_{\mathrm{F}} \overset{(34)}{\leq} \|F_i^{[t+1]}\|_{\mathrm{F}}\|R_L^{[t+1]}\|_{\mathrm{F}}\|A_i^{[t+1]}\|_{\mathrm{F}} \prod_{j=i}^{L-1} \|W_{j+1}^{[t]} \odot F_{j+1}^{[t+1]}\|_{\mathrm{F}}\|\sigma'((W_j^{[t]} \odot F_j^{[t+1]})A_{j-1}^{[t+1]})\|_{\mathrm{F}}$$

$$\overset{(35)}{\leq} C_5\|F_i^{[t+1]}\|_{\mathrm{F}}(1 + \|X^{[t+1]}\|_2)^{k^i} \prod_{l=1}^{i-1}(1 + \|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}})^{k^{i-l}}$$

$$\times \|R_L^{[t+1]}\|_{\mathrm{F}} \prod_{j=i}^{L-1} \|W_{j+1}^{[t]} \odot F_{j+1}^{[t+1]}\|_{\mathrm{F}}\|\sigma'((W_j^{[t]} \odot F_j^{[t+1]})A_{j-1}^{[t+1]})\|_{\mathrm{F}}$$

$$\overset{(36)}{\leq} C_7 \|F_i^{[t+1]}\|_{\mathrm{F}} \big(1 + \|X^{[t+1]}\|_2\big)^{k^i} \prod_{l=1}^{i-1} \big(1 + \|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}}\big)^{k^{i-l}}$$

$$\times \|R_L^{[t+1]}\|_{\mathrm{F}} \prod_{j=i}^{L-1} \|W_{j+1}^{[t]} \odot F_{j+1}^{[t+1]}\|_{\mathrm{F}} \big(1 + \|X^{[t+1]}\|_2\big)^{k^j} \prod_{l=1}^{j} \big(1 + \|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}}\big)^{k^{j-l}}$$

$$\leq C_7 \|F_i^{[t+1]}\|_{\mathrm{F}} \|R_L^{[t+1]}\|_{\mathrm{F}} \Big(\prod_{j=i}^{L-1} \|W_{i+1}^{[t]} \odot F_{i+1}^{[t+1]}\|_{\mathrm{F}}\Big) \Big(\prod_{j=i}^{L-1} \big(1 + \|X^{[t+1]}\|_2\big)^{2k^j} \prod_{l=1}^{j} \big(1 + \|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}}\big)^{2k^{j-l}}\Big)$$

$$= C_7 \|F_i^{[t+1]}\|_{\mathrm{F}} \|R_L^{[t+1]}\|_{\mathrm{F}} \Big(\prod_{j=i}^{L-1} \|W_{i+1}^{[t]} \odot F_{i+1}^{[t+1]}\|_{\mathrm{F}}\Big) \big(1 + \|X^{[t+1]}\|_2\big)^{\sum_{j=i}^{L-1} 2k^j} \Big(\prod_{j=i}^{L-1} \prod_{l=1}^{j} \big(1 + \|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}}\big)^{2k^{j-l}}\Big).$$

(37)

Lastly, we bound $\|R_L^{[t+1]}\|_{\mathrm{F}}$. By applying (i) subadditivity of the norm $\|A + B\|_F \leq \|A\|_F + \|B\|_F$ and then using the elementary bound $(a + b)^2 \leq 2(a^2 + b^2)$ as well as the submultiplicativity property, we obtain

$$\|R_L^{[t+1]}\|_{\mathrm{F}} \overset{(25)}{=} \|Y^{[t+1]} - (W_L^{[t]} \odot F_L^{[t+1]}) A_{L-1}^{[t+1]}\|_{\mathrm{F}} \overset{(i)}{\leq} \|Y^{[t+1]}\|_2^2 + \|W_L^{[t]} \odot F_L^{[t+1]}\|_{\mathrm{F}} \|A_{L-1}^{[t+1]}\|_{\mathrm{F}}$$

$$\overset{(35)}{\leq} \|Y^{[t+1]}\|_2 + \|W_L^{[t]} \odot F_L^{[t+1]}\|_{\mathrm{F}} \big(1 + \|X^{[t+1]}\|_2\big)^{k^{L-1}} \prod_{l=1}^{L-1} \big(1 + 2\|W_l^{[t]} \odot F_l^{[t+1]}\|_{\mathrm{F}}\big)^{k^{L-l}}. \quad (38)$$

By combining inequalities (37), (38), and upper bounding the exponent of the term $1 + \|X^{[t+1]}\|_{\mathrm{F}}$ in (38) by $2\sum_{j=1}^{L-1} k^j$, we conclude that

$$\|\Delta_i^{[t+1]}\|_{\mathrm{F}} \leq C_8 \|Y^{[t+1]}\|_2 \big(1 + \|X^{[t+1]}\|_2\big)^{2\sum_{j=1}^{L-1} k^j} \|F_i^{[t+1]}\|_{\mathrm{F}} P_1\big(\|W_1^{[t]} \odot F_1^{[t+1]}\|_{\mathrm{F}}, \dots, \|W_L^{[t]} \odot F_L^{[t+1]}\|_{\mathrm{F}}\big)$$

$$+ C_9 \big(1 + \|X^{[t+1]}\|_2\big)^{2\sum_{j=1}^{L} k^j} \|F_i^{[t+1]}\|_{\mathrm{F}} P_2\big(\|W_1^{[t]} \odot F_1^{[t+1]}\|_{\mathrm{F}}, \dots, \|W_L^{[t]} \odot F_L^{[t+1]}\|_{\mathrm{F}}\big) \quad (39)$$

for $i = 1, \dots, L$ and some constants C_8, C_9 and polynomials $P_1(z_1, \dots, z_L), P_2(z_1, \dots, z_L)$, say, the latter both in L variables. Because of the projection and by definition of $\mathcal{H}$, there exists a constant M such that $\|W_i^{[t]}\|_{\mathrm{F}} \leq M$ with probability one for all $i = 1, \dots, L, t \in \mathbb{N}_+$. Furthermore, $\|F_i^{[t]}\|_{\mathrm{F}} \leq \max_{i=0,\dots,L-1} \sqrt{d_i d_{i+1}}$ with probability one for all $i = 1, \dots, L, t \in \mathbb{N}_+$. These two bounds, together with (39) and the fact that P_1, P_2 are polynomials, as well as the hypothesis that $\mathbb{E}[\|Y\|_2^m \|X\|_2^n] < \infty \, \forall m \in \{0, 1, 2\}, n \in \mathbb{N}_0$, implies the result. $\qquad\square$

A.2. Conditional expectation of $\Delta^{[t+1]}$ – Proof of Lemma 2

Condition (A2.2) is that each of the expectations of random directions $\Delta_i^{[t+1]}$ for $i = 1, \dots, L$ conditional on $\mathcal{F}_t$ can written as a function of the weights. Here, $\mathcal{F}_t$ denotes the smallest σ-algebra generated by $\cup_{s \leq t} \{W^{[0]}, (F^{[s]}, X^{[s]}, Y^{[s]})\}$. For the class of dropout algorithms under consideration, we show in Lemma 2 that this is true with the function being the gradient of dropout algorithm's risk function in (9). Condition (A2.3), the continuity of the derivative, is also one of Lemma 2's consequences. Lastly, condition (A2.5) is guaranteed by Lemma 2, since it essentially proves that $\mathbb{E}[\Delta^{[t+1]}|\mathcal{F}_t] - \nabla\mathcal{D}(W^{[t]}) = 0$.

Proof. Let $i \in \{1, \dots, L\}, r \in \{1, \dots, d_{i+1}\}$ and $l \in \{1, \dots, d_i\}$. Recall that $\mathcal{F}_t$ is the smallest σ-algebra generated by $\{W^{[0]}, (F^{[s]}, X^{[s]}, Y^{[s]})\}_{s \leq t}$, and note that $W^{[t]}$ is $\mathcal{F}_t$-measurable. The (i) $\mathcal{F}_t$-measurability of $W^{[t]}$ together with the (ii) hypothesis that the sequences of random variables $\{(F^{[s]}, X^{[s]}, Y^{[s]})\}_{s \in \mathbb{N}_+}$ is i.i.d. implies that

$$\mathbb{E}[\Delta_{i,r,l}^{[t]}|\mathcal{F}_t] \overset{(8)}{=} \mathbb{E}\Big[\big(F_{i,r,l}^{[t+1]} \mathrm{FB}_{F^{[t+1]} \odot W^{[t]}}(X^{[t+1]}, Y^{[t+1]})\big)_{i,r,l} \Big| \mathcal{F}_t\Big]$$

$$\overset{(i,ii)}{=} \int F_{i,r,l} \mathrm{FB}_{F \odot W^{[t]}}(X, Y)_{i,r,l} \, \mathrm{d}\mathbb{P}[F^{[t+1]} = F, X^{[t+1]} = X, Y^{[t+1]} = Y]$$

$$\overset{(6)}{=} \int F_{i,r,l} \frac{\partial l(\Psi_{F \odot W^{[t]}}(X), Y)}{\partial (F_{i,r,l} W_{i,r,l})} \, \mathrm{d}\mathbb{P}[F^{[t+1]} = F, X^{[t+1]} = X, Y^{[t+1]} = Y]$$

$$= \int \frac{\partial l(\Psi_{F \odot W^{[t]}}(X), Y)}{\partial W_{i,r,l}} \, \mathrm{d}\mathbb{P}[F^{[t+1]} = F, X^{[t+1]} = X, Y^{[t+1]} = Y]. \tag{40}$$

Next, we need to check that we can exchange the derivative and expectation. Note that we have the same assumptions $\mathbb{E}[\|Y\|_2^m \|X\|_2^n] < \infty \, \forall m \in \{0, 1, 2\}, n \in \mathbb{N}_+$ as Lemma 1. as well as $\sigma \in C_{\mathrm{PB}}^r(\mathbb{R})$. Therefore, by (39) in Lemma 1 in any compact $\mathcal{K} \subset \mathcal{W}$ we have $|\Delta_{i,r,l}^{[t+1]}|$ is upper bounded by the right hand side of (39) and moreover $\mathbb{E}[\Delta_{i,r,l}^{[t+1]}] \leq C_K$ for some $C_K \leq \infty$ only dependent on K. The interchange is then warranted by the dominated convergence theorem. Hence continuing from (40), we obtain

$$\mathbb{E}[\Delta_{i,r,l}^{[t]} | \mathcal{F}_t] = \frac{\partial}{\partial W_{i,r,l}} \int l(\Psi_{F \odot W^{[t]}}(X), Y) \, \mathrm{d}\mathbb{P}[F^{[t+1]} = F, X^{[t+1]} = X, Y^{[t+1]} = Y] \overset{(9)}{=} \frac{\partial \mathcal{D}(W^{[t]})}{\partial W_{i,r,l}}. \tag{41}$$

If $\sigma \in C_{\mathrm{PB}}^r(\mathbb{R})$, then by the chain rule and upper bounds for any multi-index s on the weights $\mathcal{E}$ a bound similar to (39) holds:

$$|\partial^s l(Y, \Psi_{W \odot F}(X))| \leq \|Y\|_{\mathrm{F}} P_{1,s}(\|W_1\|_{\mathrm{F}}, \dots, \|W_L\|_{\mathrm{F}}, \{\|X\|_2^j\}_{j=1}^{n_{s,1}}) + P_{2,s}(\|W_1\|_{\mathrm{F}}, \dots, \|W_L\|_{\mathrm{F}}, \{\|X\|_2^j\}_{j=1}^{n_{s,2}}) \tag{42}$$

where $P_{1,s}, P_{2,s}$ are polynomials and $n_{s,1}, n_{s,2}$ are the top exponents in the expansion in $\|X\|_{\mathrm{F}}$. Hence, using the assumption $\mathbb{E}[\|Y\|_2^m \|X\|_2^n] < \infty \, \forall m \in \{0, 1, 2\}, n \in \mathbb{N}_+$, we obtain for any $W \in K \subset \mathcal{W}$ a compact set that $\mathbb{E}[|\partial^s l(Y, \Psi_{W \odot F}(X))|] \leq C_K$. In particular we can apply dominated convergence and conclude $\mathcal{D}(W) \in C^{r-1}(\mathcal{W})$ with $\partial^s \mathcal{D}(W) = \mathbb{E}[\partial^s l(Y, \Psi_{W \odot F}(X))]$. $\qquad\square$

A.3. Constant $\mathcal{D}(W)$ on a critical set – Proof of Lemma 3

Verification of (A2.6): We need Sard's theorem to prove Lemma 3, which gives sufficient conditions for condition (A2.6).

Proposition 3. *(Sard, 1942) Let $f : M \to N$ be a $f \in C^r$ map between manifolds with $\dim(M) = m$, $\dim(N) = n$. Let $\mathrm{Crit}(f) = \{x \in M | \nabla f(x) = 0\}$ be the set of critical points of f. If $r > m/n - 1$, then $f(\mathrm{Crit}(f))$ has measure zero.*

With Proposition 3, we can now prove Lemma 3 assuming (N2)–(N5) from Proposition 1.

Proof. By Lemma 2, we have $\mathcal{D}(W) \in C^r(\mathcal{W})$. By assumption (N5) we have that if $W \in \partial \mathcal{H}$ and $\mathcal{D}(W) + \pi(W) = 0$, then $\mathcal{D}(W) = 0$. Furthermore $W \in S_j$ for some j, i.e., the critical points of $\mathcal{D}(W) + \pi(W)$ are $\{W \in \mathcal{W} \mid \nabla \mathcal{D}(W) = 0\} \cap \mathcal{H}$. We apply Sard's theorem (Proposition 3) to $\mathcal{D}(W)$. We have that if $r \geq \dim(\mathcal{W})$, then $\mathcal{D}(S_i) \subseteq \mathbb{R}$ has measure zero. Since S_i is connected there is a continuous path $z_{a,b} : [0, 1] \to S_i$ joining any two points $a, b \in S_i$. By continuity of $\mathcal{D}(W)$ we must have then $\mathcal{D}(a) = \mathcal{D}(b)$, since otherwise we would have $[\mathcal{D}(a), \mathcal{D}(b)] \subseteq \mathcal{D}(S_i)$ which has positive measure in $\mathbb{R}$. Therefore $\mathcal{D}(S_i)$ must be a constant. $\qquad\square$

Note that in the previous lemma the condition $r \geq \dim(\mathcal{W})$ must hold since there are counterexamples when $r < \dim(\mathcal{W})$ (Hajłasz, 2003).

Since Conditions (A2.1)–(A2.6) of Thm. 2.3 on p. 127 in (Kushner & Yin, 2003) are satisfied, the proof of Proposition 1 is now completed.

B. Proof of Proposition 2

The proof uses double induction, which is a common approach for iterative schemes where boundedness of iterates and convergence depend on one another. First, we obtain a path representation for $\mathcal{D}(W)$ in Appendix B.1. Next, we prove that there are conserved quantities in the flow of $\nabla \mathcal{D}(W)$ in Appendix B.2. Then, we prove a bound that guarantees a notion of compactness in Appendix B.3. This is followed by a proof in Appendix B.4 that there is a PL-inequality. In Appendix B.5, we prove that the conserved quantities also remain bounded through GD's iterations. Finally, we perform the double induction in Appendix B.6.

On the exchange of derivative and expectation in this section. We start by noting that whenever we make both Assumption (N2) in Proposition 1 and (N7) in Lemma 4, that then the exchange of derivative and expectation is warranted. This occurs several times throughout this section. We refer to the proof of Lemma 2 for the details.

B.1. Path representation of $\mathcal{D}(W)$ – Proofs of Lemma 4 and Corollary 1

Proof of (15). Recall that $G_F = (\mathcal{E}_F, \mathcal{V})$ is a random subgraph of $G = (\mathcal{E}, \mathcal{V})$ with edge set $\mathcal{E}_F = \{e \in \mathcal{E} | F_e = 1\}$. By (i) the law of total expectation, and by (ii) independence of F and (X, Y) :

$$\mathcal{D}(W) = \mathbb{E}\Big[\sum_{i=1}^{d_L}\big(Y_f - \sum_{\gamma \in \Gamma^i(G)} P_\gamma F_\gamma X_{\gamma_0}\big)^2\Big] \stackrel{(i)}{=} \sum_{g \in \mathcal{G}} \mathbb{E}\Big[\sum_{f=1}^{d_L}\big(Y_f - \sum_{\gamma \in \Gamma^f(G_F)} P_\gamma X_{\gamma_0}\big)^2 \Big|\{G_F = g\}\Big]\mathbb{P}[G_F = g]$$

$$\stackrel{(ii)}{=} \sum_{g \in \mathcal{G}} \mu_g \mathbb{E}\Big[\sum_{f=1}^{d_L}\big(Y_f - \sum_{\gamma \in \Gamma^f(g)} P_\gamma X_{\gamma_0}\big)^2\Big]. \tag{43}$$

Proof of (16). Expand (43) to find

$$\mathcal{D}(W) = \sum_{g \in \mathcal{G}} \mu_g \mathbb{E}\Big[\sum_{f=1}^{d_L}\Big(Y_f^2 - 2Y_f \sum_{\gamma \in \Gamma^f(g)} P_\gamma X_{\gamma_0} + \sum_{\gamma \in \Gamma^f(g)} \sum_{\delta \in \Gamma^f(g)} P_\gamma X_{\gamma_0} P_\delta X_{\delta_0}\Big)\Big]. \tag{44}$$

Setting $\eta_\gamma = \sum_{\{g \in \mathcal{G} | \gamma \in \Gamma(g)\}} \mu_g$, we obtain

$$\mathcal{D}(W) = \sum_{g \in \mathcal{G}} \mu_g \mathbb{E}\Big[\Big(\sum_{f=1}^{d_L} \sum_{\gamma \in \Gamma^f(g)} \Big(\frac{Y_f^2}{|\Gamma^f(g)|} - 2Y_f P_\gamma X_{\gamma_0}\Big) + \sum_{\gamma \in \Gamma(g)} \sum_{\delta \in \Gamma^{\gamma_L}(g)} P_\gamma X_{\gamma_0} P_\delta X_{\delta_0}\Big)\Big]$$

$$= \sum_{\gamma \in \Gamma(G)} \eta_\gamma \mathbb{E}\Big[\big(Y_{\gamma_L} - P_\gamma X_{\gamma_0}\big)^2\Big] - \sum_{g \in \mathcal{G}} \mu_g \mathbb{E}\Big[\sum_{f=1}^{d_L} \sum_{\gamma \in \Gamma^f(g)} \Big(\Big(1 - \frac{1}{|\Gamma^f(g)|}\Big)Y_f^2 - P_\gamma X_{\gamma_0} \sum_{\delta \in \Gamma^f(g) \backslash \{\gamma\}} P_\delta X_{\delta_0}\Big)\Big] \tag{45}$$

after rearranging terms. This completes Lemma 4's proof after identifying $\mathcal{J}(W)$ and $\mathcal{R}(W)$ here as the left and right sum, respectively.

To prove Corollary 1, consider that since for an arborescence $\mathcal{R}(W) = 0$, we can write

$$\sum_{\gamma \in \Gamma(G)} \eta_\gamma \mathbb{E}\Big[\big(Y_{\gamma_L} - P_\gamma X_{\gamma_0}\big)^2\Big] = \sum_{\gamma \in \Gamma(G)} \eta_\gamma \mathbb{E}[X_{\gamma_0}^2]\Big(\frac{\mathbb{E}[Y_{\gamma_L} X_{\gamma_0}]}{\mathbb{E}[X_{\gamma_0}^2]} - P_\gamma\Big)^2 + \sum_{\gamma \in \Gamma(G)} \eta_\gamma \Big(\mathbb{E}[Y_{\gamma_L}^2] - \frac{\mathbb{E}[Y_{\gamma_L} X_{\gamma_0}]^2}{\mathbb{E}[X_{\gamma_0}^2]}\Big)$$

$$\stackrel{(iii)}{=} \mathcal{I}(W) + \mathcal{D}(W^{\text{opt}}). \tag{46}$$

Here, (iii) follows because since $\mathcal{I}(W) \geq 0$ and $\mathcal{I}(W) = 0$ at $z_\gamma = P_\gamma$, what remains must be the optimum.

This completes the proofs of Lemma 4 and Corollary 1. $\qquad\square$

B.2. Conserved quantities – Proof of Lemma 5

Proof. For any edge $f \in \mathcal{E}$,

$$W_f \frac{\partial \mathcal{D}}{\partial W_f} \overset{(15)}{=} \sum_{g \in \mathcal{G}} \mu_g \mathbb{E} \Big[\sum_{e=1}^d 2 \big(Y_e - \sum_{\gamma \in \Gamma^e(g)} P_\gamma X_{\gamma_0} \big) \big(\sum_{\delta \in \Gamma^e(g;f)} P_\delta X_{\delta_0} \big) \Big]$$

$$= \sum_{g \in \mathcal{G}} \mu_g \mathbb{E} \Big[\sum_{\delta \in \Gamma(g;f)} 2 \big(Y_{\delta_L} - \sum_{\gamma \in \Gamma^{\delta_L}(g)} P_\gamma X_{\gamma_0} \big) P_\delta X_{\delta_0} \Big]. \tag{47}$$

Note that $\Gamma(g;l) = \Gamma^l(g)$ for any leaf $l \in \mathcal{L}(G)$ and $g \in \mathcal{G}$, and therefore in particular

$$W_l \frac{\partial \mathcal{D}}{\partial W_l} = \sum_{g \in \mathcal{G}} \mu_g \sum_{\delta \in \Gamma^l(g)} \mathbb{E} \Big[2 \big(Y_{\delta_L} - \sum_{\gamma \in \Gamma^{\delta_L}(g)} P_\gamma X_{\gamma_0} \big) P_\delta X_{\delta_L} \Big]. \tag{48}$$

Recall that $\mathcal{L}(G; f)$ is the set of leafs of the subtree of the base graph G rooted at $f \in \mathcal{E}$. By the fact that $\{\Gamma^l(g;f)\}_{l \in \mathcal{L}(G;f)}$ partitions $\Gamma(g;f)$ for any $g \in \mathcal{G}$, viz.,

$$\Gamma(g;f) = \cup_{l \in \mathcal{L}(G;f)} \Gamma^l(g;f), \quad \Gamma^{l_1}(g;f) \cap \Gamma^{l_2}(g;f) = \emptyset \text{ for all } l_1 \neq l_2, g \in \mathcal{G}, \tag{49}$$

it follows that

$$\sum_{l \in \mathcal{L}(G;f)} W_l \frac{\partial \mathcal{D}}{\partial W_l} = W_f \frac{\partial \mathcal{D}}{\partial W_f}. \tag{50}$$

Note in fact that this proof works for *any* base graph G that has no cycles and only length-L paths, so not just an arborescence. This is why we make Assumption (N6') as opposed to the stronger Assumption (N6) in Corollary 1. □

B.3. Compactness, and smoothness – Proof of Lemma 6

In the proof of Lemma 6, we will upper bound the operator norm of the Hessian. Recall that for a symmetric bilinear matrix A we define $\|A\|_{\mathrm{op}} \triangleq \sup_{\|v\|_2 = 1} \|v^T A v\|_2$.

Proof of (i). By continuity of the conditions in (19), the set $B(\epsilon, \{C_f\}_{f \in \mathcal{E} \backslash \mathcal{L}})$ is closed. We need to prove boundedness. Let $W \in B(\epsilon, \{C_f\}_{f \in \mathcal{E} \backslash \mathcal{L}})$, and suppose w.l.o.g. that for some $f^* \in \mathcal{E} \backslash \mathcal{L}$ we have $|W_{f^*}| > Q$, where $Q > \max_{j \in \mathcal{E} \backslash \mathcal{L}, \gamma \in \Gamma(G)} \{|C_j|, |z_\gamma|\}$. We want to find a path $\gamma \in \Gamma(G)$ such that P_γ is large for a contradiction with the assumption that $\mathcal{I}(W) \leq \epsilon$. By (18), we have the inequality $\sum_{l \in \mathcal{L}(G;f^*)} W_l^2 > Q^2 - |C_{f^*}|$ so that for some $l^* \in \mathcal{L}(G; f^*)$ we must have $W_{l^*}^2 > (Q - |C_{f^*}|)/|\mathcal{L}(G; f^*)|$. Consequently, we have by (18) that $|W_e|^2 > (Q^2 - |C_{f^*}|)/|\mathcal{L}(G; f^*)| - |C_e|$ for any edge $e \in \gamma$ in any path $\gamma \in \Gamma^{l^*}(G)$ except for the edge f^* where we have $|W_{f^*}| > Q$ by assumption. In particular, we have the bound $|W_e| > O(Q)$ for any edge $e \in \gamma$ for any path $\gamma \in \Gamma(G; f^*)$. Therefore if we pick $\gamma \in \Gamma(G; f^*)$ we have

$$\epsilon \overset{(19)}{\geq} \mathcal{I}(W) \geq \nu_\gamma (z_\gamma - P_\gamma)^2 \geq \nu_\gamma (|P_\gamma| - |z_\gamma|)^2 > O(Q^{2L}) \tag{51}$$

for sufficiently large Q, which is a contradiction. We must thus have $|W_{f^*}| \leq Q$ for some $Q < \infty$. If on the other hand $|W_l| > Q$ for some $l \in \mathcal{L}(G; f^*)$, by (18) we must also have $(W_{f^*})^2 > Q^2 + C_{f^*} > O(Q^2)$ for sufficiently large Q. This case is, thus, the same as before.

Proof of (ii). Using a regular upper bound to the entries of $\nabla^2 \mathcal{I}(W)$ when $W \in \mathcal{S}$ will suffice. Element-wise, we have

$$(\nabla^2 \mathcal{I}(W))_{i,j} = \begin{cases} 2 \sum_{\delta \in \Gamma(G;i) \cap \Gamma(G;j)} \nu_\delta \big(\frac{P_\delta}{W_i} \frac{P_\delta}{W_j} - \frac{P_\delta}{W_i W_j} (z_\gamma - P_\gamma) \big), & \text{if } i \neq j, \Gamma(G;i) \cap \Gamma(G;j) \neq \emptyset, \\ 2 \sum_{\gamma \in \Gamma(G;i)} \nu_\gamma (\frac{P_\gamma}{W_i})^2 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

Hence, noting that since we have $|W_f| \leq M$ for all $f \in \mathcal{E}$ on $\mathcal{S}$, we can bound $|P_\gamma / W_f| \leq M^{L-1}$, $|z_\gamma| \leq M^L$ and the other terms similarly. We upper bound the number of terms in the sum over $\Gamma(G;i)$ and $\Gamma(G;i) \cap \Gamma(G;j)$ by $|\Gamma(G)|$ and $\nu_\gamma \leq \nu_{\max}$. Adding all terms, we obtain that $6\nu_{\max} |\Gamma(G)| M^{2(L-1)}$ is an upper bound for each of the entries of $\nabla^2 \mathcal{I}(W)$. This gives an upper bound $\|\nabla^2 \mathcal{I}(W)\|_{\mathrm{op}} \leq 6\nu_{\max} |\Gamma(G)| M^{2(L-1)}$ in $\mathcal{S}$. □

B.4. PL-inequality on a compact set – Proof of Lemma 7

Recall the definition of a PL-inequality:

Definition 5. *Let $u \in C^2(K, \mathbb{R})$ where $K \subset \mathbb{R}^n$ is compact and $K \backslash \partial K \neq \emptyset$. Denote by $u^* = \min_{x \in K} u(x)$ and suppose that $u^* \in K \backslash \partial K$. We say that u satisfies a* Polyak–Łojasiewicz (PL) *inequality if there exist a $\tau_K > 0$ depending only on K such that*

$$\|\nabla u(x)\|_2^2 \geq \tau_K (u(x) - u^*) \quad \text{for all} \quad x \in K. \tag{53}$$

A PL-inequality together with β-smoothness on a compact set will imply that $\mathcal{D}(W^{\{t\}}) - \mathcal{D}(W^{\text{opt}})$ decreases. To see this, note that by (i) β-smoothness, and (ii) the update rule

$$
\begin{aligned}
\mathcal{D}(W^{\{t+1\}}) - \mathcal{D}(W^{\{t\}}) &\overset{(i)}{\leq} \nabla \mathcal{D}(W^{\{t\}})^{\mathrm{T}}(W^{\{t+1\}} - W^{\{t\}}) + \beta\|W^{\{t+1\}} - W^{\{t\}}\|_2^2 \\
&\overset{(ii)}{=} \alpha(\beta\alpha - 1)\|\nabla \mathcal{D}(W^{\{t\}})\|_2^2
\end{aligned}
\tag{54}
$$

If furthermore $\alpha \leq 1/(2\beta)$, then also $\beta\alpha - 1 \leq -1/2$. Together with (53), and after rearranging terms, one finds that

$$\mathcal{D}(W^{\{t+1\}}) - \mathcal{D}(W^{\{t\}}) \leq \frac{\alpha\tau_K}{2}(\mathcal{D}(W^{\{t\}}) - \mathcal{D}(W^{\text{opt}})) \quad \text{for all} \quad W \in K. \tag{55}$$

By (iii) $1 + x \leq e^x$ for all $x \in \mathbb{R}$, we obtain (21). What remains is to now actually prove that there is a PL-inequality in some compact set, that the iterates remain in that compact set, and that the function is β-smooth.

Proof of 7. First note that if $l \in \mathcal{L}(G)$ and $\gamma \in \Gamma(G; l)$, the indexes of the weights in the product $|P_\gamma^{\{t\}}/W_l^{\{t\}}|$ belong to the index set $\mathcal{E} \backslash \mathcal{L}(G)$. The proof follows (i) by restricting the sum, and (ii) from the fact that for every path $\gamma \in \Gamma(G)$ in an arborescence G, there is exactly one leaf $l \in \mathcal{L}(G)$ such that $\gamma^l = \gamma$. Thus

$$
\begin{aligned}
\sum_{e \in \mathcal{E}} \left| \frac{\partial}{\partial W_e} \mathcal{I}(W^{\{t\}}) \right|^2 &= 4 \sum_{e \in \mathcal{E}} \left| \sum_{\gamma \in \Gamma(G;e)} \nu_\gamma \frac{P_\gamma^{\{t\}}}{W_e^{\{t\}}} (z_\gamma - P_\gamma^{\{t\}}) \right|^2 \overset{(i)}{\geq} 4 \sum_{l \in \mathcal{L}(G)} \left| \nu_{\gamma^l} \frac{P_{\gamma^l}^{\{t\}}}{W_l^{\{t\}}} (z_{\gamma^l} - P_{\gamma^l}^{\{t\}}) \right|^2 \\
&\overset{(ii)}{=} 4 \sum_{\gamma \in \Gamma(G)} \nu_\gamma^2 \left| \frac{P_\gamma^{\{t\}}}{W_{\gamma_L}^{\{t\}}} (z_\gamma - P_\gamma^{\{t\}}) \right|^2 \overset{(iii)}{\geq} 4\nu_{\min} \Big(\min_{f \in \mathcal{E} \backslash \mathcal{L}(G)} |W_f^{\{t\}}|^2 \Big)^{L-1} \mathcal{I}(W^{\{t\}}),
\end{aligned}
\tag{56}
$$

where in (iii) we have used the bound $|W_i^{\{t\}}| \geq \min_{e \in \mathcal{E} \backslash \mathcal{L}(G)} |W_e^{\{t\}}|$ for all $i \in \mathcal{E} \backslash \mathcal{L}(G)$ and similarly with $\nu_\gamma \geq \nu_{\min}$ for $\gamma \in \Gamma(G)$.

Finally, by (18), we have $\min_{e \in \mathcal{E} \backslash \mathcal{L}(G)} |W_e^{\{t\}}|^2 \geq C_{\min}^{\{t\}}$. This completes the proof. $\qquad \square$

B.5. Conserved quantities remain bounded throughout GD – Proof of Lemma 8

Proof. Pick $f \in \mathcal{E} \backslash \mathcal{L}(G)$. By (i) Corollary 1, and (ii) Lemma 5, we have

$$
\begin{aligned}
C_f^{\{t+1\}} &= (W_f^{\{t+1\}})^2 - \sum_{l \in \mathcal{L}(G;i)} (W_l^{\{t+1\}})^2 \\
&\overset{(14)}{=} \left(W_f^{\{t\}} - \alpha \frac{\partial}{\partial W_f} \mathcal{D}(W^{\{t\}}) \right)^2 - \sum_{l \in \mathcal{L}(G;f)} \left(W_l^{\{t\}} - \alpha \frac{\partial}{\partial W_l} \mathcal{D}(W^{\{t\}}) \right)^2 \\
&\overset{(i)}{=} \left(W_f^{\{t\}} - \alpha \frac{\partial}{\partial W_f} \mathcal{I}(W^{\{t\}}) \right)^2 - \sum_{l \in \mathcal{L}(G;f)} \left(W_l^{\{t\}} - \alpha \frac{\partial}{\partial W_l} \mathcal{I}(W^{\{t\}}) \right)^2 \\
&\overset{(ii)}{=} C_f^{\{t\}} + \alpha^2 \left(\left(\frac{\partial}{\partial W_f} \mathcal{I}(W^{\{t\}}) \right)^2 - \sum_{l \in \mathcal{L}(G;f)} \left(\frac{\partial}{\partial W_l} \mathcal{I}(W^{\{t\}}) \right)^2 \right) \\
&= C_i^{\{t\}} + 4\alpha^2 \left(\left(\sum_{\gamma \in \Gamma(G;f)} \nu_\gamma \frac{P_\gamma^{\{t\}}}{W_f^{\{t\}}} (z_\gamma - P_\gamma^{\{t\}}) \right)^2 - \sum_{l \in \mathcal{L}(G;f)} \nu_{\gamma^l}^2 \left(\frac{P_{\gamma^l}^{\{t\}}}{W_l^{\{t\}}} \right)^2 (z_{\gamma^l} - P_{\gamma^l}^{\{t\}})^2 \right)
\end{aligned}
\tag{57}
$$

$$\geq C_f^{\{t\}} - 4\alpha^2 \Big(\sum_{l \in \mathcal{L}(G;f)} \nu_{\gamma^l}^2 \Big(\frac{P_{\gamma^l}^{\{t\}}}{W_l^{\{t\}}}\Big)^2 (z_{\gamma^l} - P_{\gamma^l}^{\{t\}})^2 \Big). \tag{58}$$

By Cauchy–Schwartz we also have

$$\Big(\sum_{\gamma \in \Gamma(G;f)} \nu_\gamma \frac{P_\gamma^{\{t\}}}{W_f^{\{t\}}} (z_\gamma - P_\gamma^{\{t\}}) \Big)^2 \leq \Big(\sum_{\gamma \in \Gamma(G;f)} \nu_\gamma \Big) \sum_{\gamma \in \Gamma(G;f)} \nu_\gamma \Big(\frac{P_\gamma^{\{t\}}}{W_l^{\{t\}}}\Big)^2 (z_\gamma - P_\gamma^{\{t\}})^2. \tag{59}$$

If we have $C_f^{\{t\}} > 0$, then $(W_f^{\{t\}})^2 > (W_{\gamma_L}^{\{t\}})^2$ for any $\gamma \in \Gamma(G;f)$. Thus, combining the estimate (57) with (59) we obtain

$$C_f^{\{t+1\}} \leq C_f^{\{t\}} + 4 \Big(\sum_{\gamma \in \Gamma(G;f)} \nu_\gamma \Big) \alpha^2 \Big(\sum_{l \in \mathcal{L}(G;f)} \nu_{\gamma^l} \Big(\frac{P_{\gamma^l}^{\{t\}}}{W_l^{\{t\}}}\Big)^2 (z_{\gamma^l} - P_{\gamma^l}^{\{t\}})^2 \Big). \tag{60}$$

Extending the sums in (60) from $\Gamma(G;f)$ to $\Gamma(G)$ and from $\mathcal{L}(G;f)$ to $\mathcal{L}(G)$, respectively, yields

$$C_f^{\{t+1\}} - C_f^{\{t\}} \leq 4 \|\nu\|_1 \alpha^2 \Big(\max_{e \in \mathcal{E} \setminus \mathcal{L}(G)} |W_e^{\{t\}}|^2 \Big)^{L-1} \mathcal{I}(W^{\{t\}}), \tag{61}$$

where we have used the bound $|W_f| \leq \max_{e \in \mathcal{E} \setminus \mathcal{L}(G)} |W_e|$ for all $f \in \mathcal{E} \setminus \mathcal{L}(G)$. Similarly, using (58) and the trivial bound $\nu_\gamma \leq \|\nu\|_1$ for any $\gamma \in \Gamma$, and by absorbing one ν_γ-term into $\mathcal{I}(W)$'s expression, we obtain

$$C_f^{\{t+1\}} \geq C_f^{\{t\}} - 4 \|\nu\|_1 \alpha^2 \Big(\max_{e \in \mathcal{E} \setminus \mathcal{L}(G)} |W_e^{\{t\}}|^2 \Big)^{L-1} \mathcal{I}(W^{\{t\}}) \tag{62}$$

for the lower bound.

Since $W^{\{t\}} \in \mathcal{S}$ by assumption, we have the bound $\max_{e \in \mathcal{E} \setminus \mathcal{L}(G)} |W_e^{\{t\}}|^2 \leq M^2$. This completes the proof. $\qquad\square$

B.6. Double induction

We now use Lemmas 6– 8 together in a double induction to finally prove Proposition 2. Let $\kappa > 0$ and denote the statements:

$$A(t) \equiv \{\mathcal{I}(W^{\{s\}}) \leq \mathcal{I}(W^{\{s-1\}}) e^{-2\nu_{\min}\kappa\alpha}, \forall s \in [t]\}, \tag{63}$$
$$B(t) \equiv \{W^{\{s\}} \in B(\epsilon, I) \cap \mathcal{S} \, \forall s \in [t]\}. \tag{64}$$

We will prove that there exists a $\kappa > 0$ such that when choosing α appropriately, firstly

$$A(t) \cap B(t) \Rightarrow B(t+1), \tag{65}$$

and secondly,

$$A(t) \cap B(t+1) \Rightarrow A(t+1). \tag{66}$$

Step 1: $A(t) \cap B(t) \Rightarrow B(t+1)$. We need to prove that $W^{\{t+1\}} \in B(\epsilon, I) \cap \mathcal{S}$ assuming (63) and (64). Using (61) from the proof of Lemma 8 repeatedly with the bound $\max_{e \in \mathcal{E}} |W_e^{\{t\}}| \leq M$, we obtain

$$C_f^{\{t+1\}} \leq C_f^{\{0\}} + 4 \|\nu\|_1 M^{2(L-1)} \alpha^2 \sum_{s=0}^{t} \mathcal{I}(W^{\{s\}}). \tag{67}$$

By (63), we can upper bound

$$\sum_{s=0}^{t} \mathcal{I}(W^{\{s\}}) \overset{(63)}{\leq} \sum_{s=0}^{t} \mathcal{I}(W^{\{0\}}) \exp(-2\nu_{\min}\kappa\alpha s) \leq \mathcal{I}(W^{\{0\}}) \frac{1}{1 - e^{-2\nu_{\min}\kappa\alpha}}. \tag{68}$$

If furthermore (C1) $0 < 2\nu_{\min}\kappa\alpha < 1$, then (i) the inequality $1/(1 - \exp(-2\nu_{\min}\kappa\alpha)) < 1/(\nu_{\min}\kappa\alpha)$ holds, so that

$$C_{\min}^{\{t+1\}} \overset{(67)}{\leq} C_{\min}^{\{0\}} + 4\left\|\nu\right\|_1 M^{2(L-1)}\alpha^2 \sum_{s=0}^{t} \mathcal{I}(W^{\{s\}}) \overset{(i)}{\leq} C_{\min}^{\{0\}} + 4\frac{\left\|\nu\right\|_1}{\nu_{\min}}M^{L-1}\alpha\kappa^{-1}\mathcal{I}(W^{\{0\}}). \tag{69}$$

In the same manner, we can also prove (69) for $C_f^{\{0\}}$ instead of $C_{\min}^{\{0\}}$. This yields

$$C_f^{\{t+1\}} \leq C_f^{\{0\}} + 4\frac{\left\|\nu\right\|_1}{\nu_{\min}\kappa}M^{2(L-1)}\alpha\mathcal{I}(W^{\{0\}}) \tag{70}$$

for any $f \in \mathcal{E}\backslash\mathcal{L}(G)$. Similarly, for a lower bound, we can use (62) repeatedly together with the bound (68) and condition (C1) yielding

$$C_f^{\{t+1\}} \geq C_f^{\{0\}} - 4\frac{\left\|\nu\right\|_1}{\nu_{\min}\kappa}M^{2(L-1)}\alpha\mathcal{I}(W^{\{0\}}). \tag{71}$$

for any $f \in \mathcal{E}\backslash\mathcal{L}(G)$. Now, suppose (D1) $C_{\min}^{\{0\}} - \kappa^{1/(L-1)} > 0$ and let (C2) the step size satisfy

$$\alpha \leq \nu_{\min}\kappa\frac{C_{\min}^{\{0\}} - \kappa^{1/(L-1)}}{8\left\|\nu\right\|_1 M^{2(L-1)}\mathcal{I}(W^{\{0\}})}. \tag{72}$$

We have (i) by (70) and (71) that

$$\begin{aligned}
C_f^{\{t+1\}} &\overset{(i)}{\in} [C_f^{\{0\}} - 4\frac{\left\|\nu\right\|_1}{\nu_{\min}}M^{2(L-1)}\alpha\kappa^{-1}\mathcal{I}(W^{\{0\}}), C_f^{\{0\}} + 4\frac{\left\|\nu\right\|_1}{\nu_{\min}}M^{2(L-1)}\alpha\kappa^{-1}\mathcal{I}(W^{\{0\}})] \\
&\overset{(72)}{\subseteq} [C_f^{\{0\}} - \frac{1}{2}(C_{\min}^{\{0\}} - \kappa^{1/(L-1)}), C_f^{\{0\}} + \frac{1}{2}(C_{\min}^{\{0\}} - \kappa^{1/(L-1)})] \\
&\overset{(D1)}{\subseteq} [C_f^{\{0\}} - C_f^{\{0\}}/2, C_f^{\{0\}} + C_f^{\{0\}}/2] \subseteq [C_f^{\{0\}}/2, 3C_f^{\{0\}}/2] = I_f.
\end{aligned} \tag{73}$$

Then $W^{\{t+1\}} \in B(\epsilon, I)$ by (19). Hence, $M > W_f^{\{t+1\}} \overset{(18)}{>} \sqrt{1/2C_f^{\{0\}}} \geq \sqrt{1/2C_{\min}^{\{0\}}} > \delta$ for any $f \in \mathcal{E}\backslash\mathcal{L}(G)$. Moreover, since $C_e^{\{t+1\}} > 0$ for all $e \in \mathcal{E}\backslash\mathcal{L}(G)$, we have that if $f \in \mathcal{L}(G)$, then $M^2 > (W_j^{\{t+1\}})^2 > (W_f^{\{t+1\}})^2$ for some $j \in \mathcal{E}\backslash\mathcal{L}(G)$. Consequently we also obtain $M \geq |W_f^{\{t+1\}}|$ and $W^{\{t+1\}} \in \mathcal{S}$.

Step 2: $A(t) \cap B(t+1) \Rightarrow A(t+1)$. Suppose that $W^{\{s\}} \in B(\epsilon, I) \cap \mathcal{S}$ for $s = 0, 1, \ldots, t+1$. Using the bound in (70) which requires the induction hypothesis $A(t)$ and (C1) for $C_{\min}^{\{t\}}$, we obtain

$$C_{\min}^{\{t\}} \geq C_{\min}^{\{0\}} - 4\frac{\left\|\nu\right\|_1}{\nu_{\min}\kappa}M^{2(L-1)}\alpha\mathcal{I}(W^{\{0\}}). \tag{74}$$

Suppose now for a moment that (C2) the right-hand side of (74) is positive for some sufficiently small α. We could then use the PL inequality from Lemma 7 together with the bound $\min_{e\in\mathcal{E}\backslash\mathcal{L}(G)} |W_e^{\{t\}}|^{2(L-1)} \geq (C_{\min}^{\{t\}})^{L-1}$, that is,

$$\left\|\nabla\mathcal{I}(W^{\{t\}})\right\|_2^2 \geq 4\nu_{\min}(C_{\min}^{\{t\}})^{L-1}\mathcal{I}(W^{\{t\}}). \tag{75}$$

To see how, note that the argumentation around (55) together with (75) and (i) the induction hypothesis $B(t+1)$ we have $W^{\{t\}}, W^{\{t+1\}} \in B(\epsilon, I) \cap \mathcal{S}$ and (ii) the clause (L1) $\alpha \leq 1/(2\beta)$, implies

$$\begin{aligned}
\mathcal{I}(W^{\{t+1\}}) &\overset{(i,ii,75)}{\leq} \mathcal{I}(W^{\{t\}})\exp\big(-2\nu_{\min}\alpha(C_{\min}^{\{t\}})^{L-1}\big) \\
&\overset{(74)}{\leq} \mathcal{I}(W^{\{t\}})\exp\left(-2\nu_{\min}\alpha\big(C_{\min}^{\{0\}} - 4\frac{\left\|\nu\right\|_1}{\nu_{\min}\kappa}M^{2(L-1)}\alpha\mathcal{I}(W^{\{0\}})\big)\right) \\
&\overset{(iii)}{\leq} \mathcal{I}(W^{\{0\}})\exp\left(-2\nu_{\min}\alpha\big(C_{\min}^{\{0\}} - 4\frac{\left\|\nu\right\|_1}{\nu_{\min}\kappa}M^{2(L-1)}\alpha\mathcal{I}(W^{\{0\}})\big)^{L-1} - 2\nu_{\min}\alpha\kappa t\right)
\end{aligned} \tag{76}$$

where we have also used (iii) the induction hypothesis $A(t)$, i.e., that $\mathcal{I}(W^{\{t\}}) \leq \mathcal{I}(W^{\{0\}})\exp(-2\nu_{\min}\kappa\alpha t)$ holds.

We now investigate the exponent in (76) for a moment. Assuming (C2) and if (C3) the right-hand side of (76) is furthermore smaller than $\mathcal{I}(W^{\{0\}})\exp(-2\nu_{\min}\kappa\alpha(t+1))$, then the induction step would be complete. Note finally that both conditions (C2) and (C3) are satisfied when choosing

$$\kappa \leq \left(C_{\min}^{\{0\}} - 4\frac{\|\nu\|_1}{\nu_{\min}}M^{2(L-1)}\alpha\kappa^{-1}\mathcal{I}(W^{\{0\}})\right)^{L-1} \tag{77}$$

or equivalently

$$\alpha \leq \nu_{\min}\kappa\frac{C_{\min}^{\{0\}} - \kappa^{1/(L-1)}}{4\|\nu\|_1 M^{2(L-1)}\mathcal{I}(W^{\{0\}})}. \tag{78}$$

To also satisfy condition (C1), we thus require that

$$\alpha \leq \min\left(\frac{1}{2\nu_{\min}\kappa}, \nu_{\min}\kappa\frac{C_{\min}^{\{0\}} - \kappa^{1/(L-1)}}{4\|\nu\|_1 M^{2(L-1)}\mathcal{I}(W^{\{0\}})}\right). \tag{79}$$

Step 3. Let us summarize. Convergence occurs at rate at most $2\nu_{\min}\kappa\alpha$ if conditions (L1), (D1), (C1)–(C3) hold. Hence we have to choose $\kappa > 0$ such that $C_{\min}^{\{0\}} - \kappa^{L-1} > 0$ and

$$\alpha \leq \min\left(\nu_{\min}\kappa\frac{C_{\min}^{\{0\}} - \kappa^{1/(L-1)}}{8\|\nu\|_1 M^{2(L-1)}\mathcal{I}(W^{\{0\}})}, \frac{1}{2\beta}, \frac{1}{2\nu_{\min}\kappa}\right). \tag{80}$$

Note that we can maximize the convergence rate $2\nu_{\min}\alpha\kappa$ by maximizing $\kappa^2(C_{\min}^{\{0\}} - \kappa^{1/(L-1)})$, which occurs when $\kappa = (C_{\min}^{\{0\}})^{L-1}(1 + 1/(2(L-1)))^{-(L-1)} \geq \mathrm{e}^{-1/2}(C_{\min}^{\{0\}})^{L-1}$. Substituting this in (80) we require a step size

$$\alpha \leq \min\left(\nu_{\min}\frac{\mathrm{e}^{1/2}(C_{\min}^{\{0\}})^L}{8\|\nu\|_1 (2L-1)M^{2(L-1)}\mathcal{I}(W^{\{0\}})}, \frac{1}{2\beta}, \frac{1}{2\nu_{\min}(C_{\min}^{\{0\}})^{L-1}}\right). \tag{81}$$

Finally, we have the bound $\beta \leq 6\nu_{\max}|\Gamma(G)|M^{2(L-1)}$ from Lemma 6 in $\mathcal{S}$, so that

$$\alpha \leq \min\left(\nu_{\min}\frac{\mathrm{e}^{1/2}(C_{\min}^{\{0\}})^L}{8\|\nu\|_1 (2L-1)M^{2(L-1)}\mathcal{I}(W^{\{0\}})}, \frac{1}{12\nu_{\max}|\Gamma(G)|M^{2(L-1)}}, \frac{1}{2\nu_{\min}(C_{\min}^{\{0\}})^{L-1}}\right). \tag{82}$$

This completes our proof of Proposition 2. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

B.7. Convergence rate in the case of *Dropconnect* – Proof of Corollary 2

Suppose that the base graph G has no cycles and every path is of length L. Then by definition in Lemma 4, we have

$$\eta_\gamma = \sum_{\{g\in\mathcal{G}|\gamma\in\Gamma(g)\}} \mathbb{P}[G_F = g] = \sum_{g\in\mathcal{G}} \mathbb{1}[\gamma\in\Gamma(g)]\mathbb{P}[G_F = g]$$

$$= \sum_{g\in\mathcal{G}} \mathbb{P}[\gamma\in\Gamma(g)|G_F = g]\mathbb{P}[G_F = g] = \mathbb{P}[\gamma\in\Gamma(G_F)] \overset{\text{(i)}}{=} p^L \tag{83}$$

where (i) we have used *Dropconnect*'s distribution on F.

Now suppose that additionally we make the stronger assumption that G is an arborescence. Then by definition in Corollary 1 $\nu_\gamma = \mathbb{E}[X^2]\eta_\gamma$, and subsequently we can calculate $\|\nu\|_1 = \mathbb{E}[X^2]\sum_{\gamma\in\Gamma(G)}\nu_\gamma = \mathbb{E}[X^2]|\Gamma(G)|p^L = \mathbb{E}[X^2]d_L p^L$.

Now, since by assumption $\max_\gamma |z_\gamma| \leq M^L$ and $|W_f| \leq M$ for all $f \in \mathcal{E}$, then $\mathcal{I}(W^{\{0\}}) \leq O(|\Gamma(G)|M^{2L})$ so that substitution of in the definition of α in Proposition 2 yields $\alpha = O((C_{\min}^{\{0\}})^L/(LM^{4L}))$, where we have used that $C_{\min} \leq M^2$. Finally multiplying by τ gives the rate $\alpha\tau = O((p^L(C_{\min}^{\{0\}})^2 L)/(L(d_L)^2 M^{4L}))$.

Substituting these results in the rate $\tau\alpha$ in Proposition 2 yields the result. $\qquad\qquad\qquad\qquad\qquad\square$