On Convergence and Generalization of Dropout Training

Poorya Mianjy
Department of Computer Science
Johns Hopkins University
mianjy@jhu.edu

Raman Arora
Department of Computer Science
Johns Hopkins University
arora@cs.jhu.edu

Abstract

We study dropout in two-layer neural networks with rectified linear unit (ReLU) activations. Under mild overparametrization and assuming that the limiting kernel can separate the data distribution with a positive margin, we show that dropout training with logistic loss achieves ϵ-suboptimality in test error in $O(1/\epsilon)$ iterations.

1 Introduction

Dropout is an algorithmic regularization approach that endows deep learning models with excellent generalization ability despite the non-convex nature of the underlying learning problem and the capacity of modern over-parameterized models to over-fit. Introduced by Hinton et al. [2012], Srivastava et al. [2014], dropout involves randomly pruning the network at every iteration of backpropagation by turning off a random subset of hidden nodes. Like many popular algorithmic approaches that emerged as heuristics from practitioners with deep insight into the learning problem, dropout, while extremely successful in practice, lacks a strong theoretical justification, especially from a computational learning theoretic perspective.

Dropout has been successful in several application areas including computer vision [Szegedy et al., 2015], natural language processing [Merity et al., 2017], and speech recognition [Dahl et al., 2013]. Motivated by explaining the empirical success of dropout, and inspired by simple, intuitive nature of the method, numerous works in recent years have focused on understanding its theoretical underpinnings [Baldi and Sadowski, 2013, Wager et al., 2013, Helmbold and Long, 2015, Gal and Ghahramani, 2016, Wei et al., 2020]. Most of these works, however, steer clear from the algorithmic and computational learning aspects of dropout. More precisely, none of the prior work, to the best of our knowledge, yields insights into the runtime of learning using dropout on non-linear neural networks. In this paper, we initiate a study into the iteration complexity of dropout training for achieving ϵ-suboptimality on true error – the misclassification error with respect to the underlying population – in two-layer neural networks with ReLU activations.

We leverage recent advances in the theory of deep learning in over-parameterized settings with extremely (or infinitely) wide networks [Jacot et al., 2018, Lee et al., 2019]. Analyzing two-layer ReLU networks in such a regime has led to a series of exciting results recently establishing that gradient descent (GD) or stochastic gradient descent (SGD) can successfully minimize the empirical error and the true error [Li and Liang, 2018, Du et al., 2019, Daniely, 2017, Zou et al., 2018, Allen-Zhu et al., 2019, Song and Yang, 2019, Arora et al., 2019, Cao and Gu, 2019, Oymak and Soltanolkotabi, 2020]. In a related line of research, several works attribute generalization in over-parametrized settings to the implicit inductive bias of optimization algorithms (through the geometry of local search methods) [Neyshabur et al., 2017]. However, many real-world state-of-the-art systems employ various explicit regularizers, and there is growing evidence that implicit bias may be unable to explain generalization even in a simpler setting of stochastic convex optimization [Dauber et al., 2020]. Here, we extend convergence guarantees and generalization bounds for GD-based methods with explicit

regularization due to dropout. We show that the key insights from analysis of GD-based methods in over-parameterized settings carry over to dropout training.

We summarize our key contributions as follows.

1. We give precise non-asymptotic convergence rates for achieving ϵ-subotimality in the test error via dropout training in two-layer ReLU networks.

2. We show that dropout training implicitly compresses the network. In particular, we show that there exists a sub-network, i.e., one of the iterates of dropout training, that can generalize as well as any complete network.

3. This work contributes to a growing body of work geared toward a theoretical understanding of GD-based methods for regularized risk minimization in over-parameterized settings.

The rest of the paper is organized as follows. In Section 2, we survey the related work. In Section 3, we formally introduce the problem setup and dropout training, state the key assumptions, and introduce the notation. In Section 4, we give the main results of the paper. In Section 5, we present a sketch of the proofs of our main results – the detailed proofs are deferred to the Appendix. We conclude the paper by providing empirical evidence for our theoretical results in Section 6.

2   Related Work

Empirical success of dropout has inspired a series of works aimed at understanding its theoretical underpinnings. Most of these works have focused on explaining the algorithmic regularization due to dropout in terms of *conventional* regularizers. Dropout training has been shown to be similar to *weight decay* in linear regression [Srivastava et al., 2014], in generalized linear models [Wager et al., 2013], and in a PAC-Bayes setting [McAllester, 2013]. These results have recently been extended to multi-variate regression models, where dropout induces a *nuclear norm* penalty in single hidden-layer linear networks [Mianjy et al., 2018] and deep linear networks [Mianjy and Arora, 2019], and a *weighted trace-norm* penalty in matrix completion [Arora et al., 2020]. In a recent work, Wei et al. [2020] characterize explicit regularization due to dropout in terms of the derivatives of the loss and the model, and argue for an implicit regularization effect that stems from the stochasticity in dropout updates.

A parallel strand of research has focused on bounding the generalization gap in dropout training, leveraging tools from uniform convergence. The early work of Wager et al. [2014] showed that under a certain topic model assumption on the data, dropout in linear classification can improve the decay of the excess risk of the empirical risk minimizer. Assuming certain norm-bounds on the weights of the network, the works of Wan et al. [2013], Zhai and Wang [2018], Gao and Zhou [2016] showed that the Rademacher complexity of networks trained by dropout decreases with the dropout rate. Finally, Arora et al. [2020] showed further that the Rademacher complexity can be bounded merely in terms of the explicit regularizer induced by dropout.

Despite the crucial insights provided by the previous art, there is not much known about the non-asymptotic convergence behaviour of dropout training in the literature. A very recent work by Senen-Cerda and Sanders [2020] shows for deep neural networks with polynomially bounded activations with continuous derivatives, under squared loss, that the network weights converge to a stationary set of system of ODEs. In contrast, our results leverages over-parameterization in two-layer networks with non-differentiable ReLU activations, works with logistic loss, and establishes ϵ-suboptimality in the true misclassification error.

Our results are inspired by the recent advances in over-parameterized settings. A large body of literature has focused on deriving optimization theoretic guarantees for (S)GD in this setting. In particular, Li and Liang [2018], Du et al. [2019] were among the first to provide convergence rates for empirical risk minimization using GD. Several subsequent works extended those results beyond two-layers, for smooth activation functions [Du et al., 2018], and general activation functions [Allen-Zhu et al., 2018, Zou et al., 2018].

Learning theoretic aspects of GD-based methods have been studied for several important target concept classes. Under linear-separability assumption, via a compression scheme, Brutzkus et al. [2017] showed that SGD can efficiently learn a two-layer ReLU network. Li and Liang [2018] further showed that SGD enjoys small generalization error on two-layer ReLU networks if the data follows a

well-separated mixture of distributions. Allen-Zhu et al. [2019] showed generalization error bounds for SGD in two- and three-layer networks with smooth activations where the concept class has fewer parameters. Arora et al. [2019] proved data-dependent generalization error bounds based on the neural tangent kernel by analyzing the Rademacher complexity of the class of networks reachable by GD.

When the data distribution can be well-classified in the *random feature space* induced by the gradient of the network at initialization, Cao and Gu [2019] provide generalization guarantees for SGD in networks with arbitrary depth. Nitanda and Suzuki [2019] studied convergence of GD in two-layer networks with smooth activations, when the data distribution is further *separable* in the infinite-width limit of the random feature space. Ji and Telgarsky [2019] adopted the same margin assumption and improved the convergence rate as well as the over-parameterization size for non-smooth ReLU activations. Here, we generalize the margin assumption in Nitanda and Suzuki [2019] to take into account the randomness injected by dropout into the gradient of the network at initialization, or equivalently, the scale of the corresponding random feature. Our work is most closely related to and inspired by Ji and Telgarsky [2019]; however, we analyze dropout training as opposed to plain SGD, give generalization bounds in expectation, and show the compression benefits of dropout training.

We emphasize that all of the results above focus on (S)GD in absence of any explicit regularization. We summarize a few papers that study regularization in the over-parameterized setting. The work of Wei et al. [2019] showed that even simple explicit regularizers such as weight decay can indeed provably improve the sample complexity of training using GD in the Neural Tangent Kernel (NTK) regime, appealing to a margin-maximization argument in homogeneous networks. We also note the recent works by Li et al. [2019] and Hu et al. [2020], which studied the robustness of GD to noisy labels, with explicit regularization in forms of early stopping; and squared norm of the distance from initialization, respectively.

3 Preliminaries and Notation

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{\pm 1\}$ denote the input and label spaces, respectively. We assume that the data is jointly distributed according to an unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. Given T i.i.d. examples $\mathcal{S}_T = \{(\mathrm{x}_t, \mathrm{y}_t)\}_{t=1}^T \sim \mathcal{D}^T$, the goal of learning is to find a hypothesis $f(\cdot; \Theta) : \mathcal{X} \to \mathbb{R}$, parameterized by Θ, that has a small *misclassification error* $\mathcal{R}(\Theta) := \mathbb{P}\{yf(\mathrm{x}; \Theta) < 0\}$. Given a convex surrogate loss function $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$, a common approach to the above learning problem is to solve the stochastic optimization problem $\min_\Theta L(\Theta) := \mathbb{E}_{\mathcal{D}}[\ell(yf(\mathrm{x}; \Theta))]$.

In this paper, we focus on logistic loss $\ell(z) = \log(1 + e^{-z})$, which is one of the most popular loss functions for classification tasks. We consider two-layer ReLU networks of width m, parameterized by the "weights" $\Theta = (\mathrm{W}, \mathrm{a}) \in \mathbb{R}^{m \times d} \times \mathbb{R}^m$, computing the function $f(\cdot; \Theta) : \mathrm{x} \mapsto \frac{1}{\sqrt{m}} \mathrm{a}^\top \sigma(\mathrm{Wx})$. We initialize the network with $a_r \sim \mathrm{Unif}(\{+1, -1\})$ and $\mathrm{w}_{r,1} \sim \mathcal{N}(0, \mathrm{I})$, for all hidden nodes $r \in [m]$. We then fix the top layer weights and train the hidden layer W using the dropout algorithm. We denote the weight matrix at time t by W_t, and $\mathrm{w}_{r,t}$ represents its r-th column. For the sake of simplicity of the presentation, we drop non-trainable arguments from all functions, e.g., we use $f(\cdot; \mathrm{W})$ in lieu of $f(\cdot; \Theta)$.

Let $\mathrm{B}_t \in \mathbb{R}^{m \times m}$, $t \in [T]$, be a random diagonal matrix with diagonal entries drawn independently and identically from a Bernoulli distribution with parameter q, i.e., $b_{r,t} \sim \mathrm{Bern}(q)$, where $b_{r,t}$ is the r-th diagonal entry of B_t. At the t-th iterate, dropout entails a SGD step on (the parameters of) the sub-network $g(\mathrm{W}; \mathrm{x}, \mathrm{B}_t) = \frac{1}{\sqrt{m}} \mathrm{a}^\top \mathrm{B}_t \sigma(\mathrm{Wx})$, yielding updates of the form $\mathrm{W}_{t+\frac{1}{2}} \leftarrow \mathrm{W}_t - \eta \nabla \ell(y_t g(\mathrm{W}_t; \mathrm{x}_t, \mathrm{B}_t))$. The iteration concludes with projecting the incoming weights – i.e. rows of $\mathrm{W}_{t+\frac{1}{2}}$ – onto a pre-specified Euclidean norm ball. We note that such max-norm constraints are standard in the practice of deep learning, and has been a staple to dropout training since it was proposed in Srivastava et al. [2014][1]. Finally, at test time, the weights are multiplied by q so as

---

[1]Quote from Srivastava et al. [2014]: "One particular form of regularization was found to be especially useful for dropout— constraining the norm of the incoming weight vector at each hidden unit to be upper bounded by a fixed constant c"

Algorithm 1 Dropout in Two-Layer Networks

Input: data $\mathcal{S}_T = \{(x_t, y_t)\}_{t=1}^T \sim \mathcal{D}^T$; Bernoulli masks $\mathcal{B}_T = \{B_t\}_{t=1}^T$;
  dropout rate $1 - q$; max-norm constraint parameter c; learning rate η

1: *initialize:* $w_{r,1} \sim \mathcal{N}(0, I)$ and $a_r \sim \mathrm{Unif}(\{+1, -1\}),\ r \in [m]$
2: **for** $t = 1, \ldots, T - 1$ **do**
3:   *forward:* $g(W_t; x_t, B_t) = \frac{1}{\sqrt{m}} a^\top B_t \sigma(W_t x_t)$
4:   *backward:* $\nabla L_t(W_t) = \nabla \ell(y_t g(W_t; x_t, B_t)) = \ell'(y_t g(W_t; x_t, B_t)) \cdot y_t \nabla g(W_t; x_t, B_t)$
5:   *update:* $W_{t+\frac{1}{2}} \leftarrow W_t - \eta \nabla L_t(W_t)$
6:   *max-norm:* $W_{t+1} \leftarrow \Pi_c(W_{t+\frac{1}{2}})$
7: **end for**

Test Time: re-scale the weights as $W_t \leftarrow q W_t$

---

to make sure that the output at test time is on par with the expected output at training time. The pseudo-code for dropout training is given in Algorithm 1[2].

Our analysis is motivated by recent developments in understanding the dynamics of (S)GD in the so-called *lazy regime*. Under certain initialization, learning rate, and network width requirements, these results show that the iterates of (S)GD tend to stay close to initialization; therefore, a first-order Taylor expansion of the t-th iterate around initialization, i.e. $f(x; W_t) \approx f(x; W_1) + \langle \nabla f(x; W_1), W_t - W_1 \rangle$, can be used as a proxy to track the evolution of the network predictions [Li and Liang, 2018, Chizat et al., 2018, Du et al., 2019, Lee et al., 2019]. In other words, training in lazy regime reduces to finding a linear predictor in the reproducing kernel Hilbert space (RKHS) associated with the gradient of the network at initialization, $\nabla f(\cdot; W_1)$. In this work, following Nitanda and Suzuki [2019], Ji and Telgarsky [2019], we assume that the data distribution is separable by a positive margin in the limiting RKHS:

Assumption 1 $((q, \gamma)$-Margin). Let $z \sim \mathcal{N}(0, I_d)$ and $b \sim \mathrm{Bern}(q)$ be a d-dimensional standard normal random vector, and a Bernoulli random variable with parameter q, respectively. There exists a *margin parameter* $\gamma > 0$, and a linear transformation $\psi : \mathbb{R}^d \to \mathbb{R}^d$ satisfying A) $\mathbb{E}_z[\|\psi(z)\|^2] < \infty$; B) $\|\psi(z)\|_2 \leq 1$ for all $z \in \mathbb{R}^d$; and C) $\mathbb{E}_{z,b}[y\langle \psi(z), bx\mathbb{I}[z^\top x \geq 0]\rangle] \geq \gamma$ for almost all $(x, y) \sim \mathcal{D}$.

The above assumption provides an infinite-width extension to the separability of data in the RKHS induced by $\nabla g(W_1; \cdot, B_1)$. To see that, define $V := [v_1, \ldots, v_m]^\top \in \mathbb{R}^{m \times d}$, where $v_r = \frac{1}{\sqrt{m}} a_r \psi(w_{r,1})$ for all $r \in [m]$, satisfying $\|V\|_F \leq 1$. For any given point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the margin attained by V is at least $y\langle \nabla g(W_1; x, B_1), V \rangle = \frac{1}{m} \sum_{r=1}^m y\langle \psi(w_{r,1}), b_{r,1} x\mathbb{I}\{w_{r,1}^\top x > 0\}\rangle$, which is a finite-width approximation of the quantity $\mathbb{E}[y\langle \psi(z), bx\mathbb{I}\{z^\top x > 0\}\rangle]$ in Assumption 1.

We remark that when $q = 1$ (pure SGD – no dropout), with probability one it holds that $b = 1$, so that Assumption 1 boils down to that of Nitanda and Suzuki [2019] and Ji and Telgarsky [2019]. When $q < 1$, this assumption translates to a margin of γ/q on the *full* features $\nabla f(\cdot; W_1)$, which is the appropriate scaling given that $\nabla f(\cdot; W_1) = \frac{1}{q}\mathbb{E}_B[\nabla g(W_1; \cdot, B)]$. Alternatively, dropout training eventually outputs a network with weights scaled down as $q W_t$, which (in expectation) corresponds to the shrinkage caused by the Bernoulli mask in $bx\mathbb{I}\{z^\top x > 0\}$. Regardless, we note that *our analysis can be carried over even without this scaling*; however, new polynomial factors of $1/q$ will be introduced in the required width in our results in Section 4.

3.1 Notation

We denote matrices, vectors, scalar variables and sets by Roman capital letters, Roman small letters, small letters, and script letters, respectively (e.g. Y, y, y, and $\mathcal{Y}$). The r-th entry of vector y, and the r-th row of matrix Y, are denoted by y_i and y_i, respectively. Furthermore, for a sequence of matrices $W_t, t \in \mathbb{N}$, the r-th row of the t-th matrix is denoted by $w_{r,t}$. Let $\mathbb{I}$ denote the indicator of an event, i.e., $\mathbb{I}\{y \in \mathcal{Y}\}$ is one if $y \in \mathcal{Y}$, and zero otherwise. For any integer d, we represent the set $\{1, \ldots, d\}$ by $[d]$. Let $\|x\|$ represent the ℓ_2-norm of vector x; and $\|X\|$, and $\|X\|_F$ represent the

---

[2]In a popular variant that is used in machine learning frameworks such as PyTorch, known as inverted dropout, (inverse) scaling is applied at the training time instead of the test time. The inverted dropout is equivalent to the method we study here, and can be analyzed in a similar manner.

operator norm, and the Frobenius norm of matrix X, respectively. $\langle \cdot, \cdot \rangle$ represents the standard inner product, for vectors or matrices, where $\langle X, X' \rangle = \text{Tr}(X^\top X')$. For a matrix $W \in \mathbb{R}^{m \times d}$, and a scalar $c > 0$, $\Pi_c(W)$ projects the rows of W onto the Euclidean ball of radius c with respect to the ℓ_2-norm.

For any $t \in [T]$ and any W, let $f_t(W) := f(x_t; W)$ denote the network output given input x_t, and let $g_t(W) := g(W; x_t, B_t)$ denote the corresponding output of the sub-network associated with the dropout pattern B_t. Let $L_t(W) = \ell(y_t g_t(W))$ and $Q_t(W) = -\ell'(y_t g_t(W))$ be the associated instantaneous loss and its negative derivative. The partial derivative of g_t with respect to the r-th hidden weight vector is given by $\frac{\partial g_t(W)}{\partial w_r} = \frac{1}{\sqrt{m}} a_r b_{r,t} \mathbb{I}\{w_r^\top x_t \geq 0\} x_t$. We denote the linearization of $g_t(W)$ based on features at time t by $g_t^{(k)}(W) := \langle \nabla g_t(W_k), W \rangle$; and its corresponding instantaneous loss and its negative derivative by $L_t^{(k)}(W) := \ell(y_t g_t^{(k)}(W))$ and $Q_t^{(k)}(W) := -\ell'(y_t g_t^{(k)}(W))$, respectively. Q plays an important role in deriving generalization bounds for dropout sub-networks $g(W_t; x, B_t)$; it has been recently used in [Cao and Gu, 2019, Ji and Telgarsky, 2019] for analyzing the convergence of SGD and bounding its generalization error.

We conclude this section by listing a few useful identities that are used throughout the paper. First, due to homogeneity of the ReLU, it holds that $g_t^{(t)}(W_t) = \langle \nabla g_t(W_t), W_t \rangle = g_t(W_t)$. Moreover, the norm of the network gradient, and the norm of the the gradient of the instantaneous loss can be upper-bounded as $\|\nabla g_t(W)\|_F^2 = \sum_{r=1}^m \|\frac{\partial g_t(W)}{\partial w_r}\|^2 \leq \frac{\|B_t\|_F^2}{m} \leq 1$, and $\|\nabla L_t(W)\|_F = -\ell'(y_t g_t(W))\|y_t \nabla g_t(W)\|_F \leq Q_t(W)$, respectively. Finally, the logistic loss satisfies $|\ell'(z)| \leq \ell(z)$, so that $Q_t(W) \leq L_t(W)$.

4 Main Results

We begin with a simple observation that given the random initialization scheme in Algorithm 1, the ℓ_2-norm of the rows of W_1 are expected to be concentrated around $\sqrt{d}$. In fact, using Gaussian concentration inequality (Theorem A.1 in the appendix), it holds with probability at least $1 - 1/m$, uniformly for all $r \in [m]$, that $\|w_{r,1}\| \leq \sqrt{d} + 2\sqrt{\ln(m)}$. For the sake of the simplicity of the presentation, we assume that the event $\max_{r \in [m]} \|w_{r,1}\| \leq 2\sqrt{\ln(m)}$ holds through the run of dropout training. Alternatively, we can re-initialize the weights until this condition is satisfied, or multiply the probability of success in our theorems by a factor of $1 - 1/m$.

Our first result establishes that the true misclassification error of dropout training vanishes as $\tilde{\mathcal{O}}(1/T)$.

Theorem 4.1 (Learning with Dropout)**.** Let $c = \sqrt{d} + \max\{\frac{1}{14\gamma^2}, 2\sqrt{\ln(m)}\} + 1$ and $\lambda := 5\gamma^{-1} \ln(2\eta T) + \sqrt{44\gamma^{-2} \ln(24\eta c \sqrt{m} T^2)}$. Under Assumption 1, for any learning rate $\eta \in (0, \ln(2)]$ and any network of width satisfying $m \geq 2401\gamma^{-6}\lambda^2$, with probability one over the randomization due to dropout, we have that

$$\min_{t \in [T]} \mathbb{E}[\mathcal{R}(qW_t)] \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\mathcal{R}(qW_t)] \leq \frac{4\lambda^2}{\eta T} = \mathcal{O}\left(\frac{\ln(T)^2 + \ln(mdT)}{T}\right),$$

where the expectation is with respect to the initialization and the training samples.

Theorem 4.1 shows that dropout successfully trains the complete network $f(\cdot; W_t)$. Perhaps more interestingly, our next result shows that dropout successfully trains a potentially significantly narrower sub-network $g(W_t; \cdot, B_t)$. For this purpose, denote the misclassification error due to a network with weights W given a Bernoulli mask B as follows

$$\mathcal{R}(W; B) := \mathbb{P}\{yg(W; x, B) < 0\}.$$

Then the following result holds for the misclassification error of the iterates of dropout training.

Theorem 4.2 (Compression with Dropout)**.** Under the setting of Theorem 4.1, with probability at least $1 - \delta$ over initialization, the training data, and the randomization due to dropout, we have that

$$\min_{t \in [T]} \mathcal{R}(W_t; B_t) \leq \frac{1}{T} \sum_{t=1}^T \mathcal{R}(W_t; B_t) \leq \frac{12\lambda^2}{\eta T} + \frac{6\ln(1/\delta)}{T} = \mathcal{O}\left(\frac{\ln(mT)}{T}\right).$$

A few remarks are in order.

Theorem 4.1 gives a generalization error bound in expectation. A technical challenge here stems from the unboundedness of the logistic loss. In our analysis, the max-norm constraint in Algorithm 1 is essential to guarantee that the logistic loss remains bounded through the run of the algorithm, thereby controlling the loss of the iterates in expectation. However, akin to analysis of SGD in the lazy regime, the iterates of dropout training are not likely to leave a small proximity of the initialization whatsoever. Therefore, for the particular choice of c in the above theorems, the max-norm projections in Algorithm 1 will be virtually inactive for a typical run.

The expected width of the sub-networks in Theorem 4.2 is only qm. Using Hoeffding's inequality and a union bound argument, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds for all $t \in [T]$ that $g(W_t; x, B_t)$ has at most $qm + \sqrt{2m \ln (T/\delta)}$ active hidden neurons. That is, in a typical run of the dropout training, with high probability, there exists a sub-network of width $\approx qm + \tilde{\mathcal{O}}(\sqrt{m})$ whose generalization error is no larger than $\tilde{\mathcal{O}}(1/T)$. In Section 6, we further provide empirical evidence to verify this compression result. We note that dropout has long been considered as a means of network compression, improving post-hoc pruning [Gomez et al., 2019], in Bayesian settings [Molchanov et al., 2017], and in connection with the Lottery Ticket Hypothesis [Frankle and Carbin, 2019]. However, we are not aware of any theoretical result supporting that claim prior to our work.

Finally, the sub-optimality results in both Theorem 4.1 and Theorem 4.2 are agnostic to the dropout rate $1 - q$. This is precisely due to the (q, γ)-Margin assumption: if it holds, then so does (q', γ)-Margin for any $q' \in [q, 1]$. That is, these theorems hold for *any* dropout rate not exceeding $1 - q$. Therefore, in light of the remark above, larger admissible dropout rates are preferable since they result in higher compression rates, while enjoying the same generalization error guarantees.

5 Proofs

We begin by bounding $\mathbb{E}_{S_t}[\mathcal{R}(qW_t)]$, the expected population error of the iterates, in terms of $\mathbb{E}_{S_t, B_t}[L_t(W_t)]$, the expected instantaneous loss of the random sub-networks. In particular, using simple arguments including the smoothing property, the fact that W_t is independent from (x_t, y_t) given S_{t-1}, and that logistic loss upper-bounds the zero-one loss, it is easy to bound the expected population risk as $\mathbb{E}_{S_t}[\mathcal{R}(qW_t)] \leq \mathbb{E}_{S_t}[\ell(y_t f(x_t; qW_t))]$. Furthermore, using Jensen's inequality, we have that $\ell(y_t f_t(qW_t)) \leq \mathbb{E}_{B_t}[L_t(W_t)]$. The upper bound then follows from these two inequalities. In the following, we present the main ideas in bounding the average instantaneous loss of the iterates.

Under Algorithm 1, dropout iterates are guaranteed to remain in the set $\mathcal{W}_c := \{W \in \mathbb{R}^{m \times d} : \|w_r\| \leq c\}$. Using this property and the dropout update rule, we track the distance of consecutive iterates (W_{t+1}, W_t) from any competitor $U \in \mathcal{W}_c$, which leads to the following upper bound on the average instantaneous loss of iterates.

Lemma 5.1. Let $W_1, \ldots, W_T$ be the sequence of dropout iterates with a learning rate satisfying $\eta \leq \ln (2)$. Then, it holds for any $U \in \mathcal{W}_c$ that

$$\frac{1}{T} \sum_{t=1}^{T} L_t(W_t) \leq \frac{\|W_1 - U\|_F^2}{\eta T} + \frac{2}{T} \sum_{t=1}^{T} L_t^{(t)}(U). \tag{1}$$

Note that the upper bound in Equation (1) holds for *any* competitor $U \in \mathcal{W}_c$; however, we seek to minimize the upper-bound on the right hand side of Equation (1) by finding a *sweet spot* that maintains a trade-off between 1) the distance from initialization, and 2) the average instantaneous loss for the linearized models. Following Ji and Telgarsky [2019], we represent such a competitor as an interpolation between the initial weights W_1 and the *max-margin* competitor V, i.e. $U := W_1 + \lambda V$, where λ is the trade-off parameter. Recall that $V := [v_1, \cdots, v_m] \in \mathbb{R}^{d \times m}$, where $v_r = \frac{1}{\sqrt{m}} a_r \psi(w_{r,1})$ for any $r \in [m]$, and ψ is given by assumption 1 . Thus, the first term on the right hand side above can be conveniently bounded as $\frac{\lambda^2}{\eta T}$; Lemma 5.2 bounds the second term as follows.

Lemma 5.2. Under the setting of Theorem 4.1, it holds with probability at least $1 - \delta$ simultaneously for all iterates $t \in [T]$ that 1) $\|w_{r,t} - w_{r,1}\| \leq \frac{7\lambda}{2\gamma\sqrt{m}}$, for all $r \in [m]$; and 2) $L_t^{(t)}(U) \leq \frac{\lambda^2}{2\eta T}$.

Therefore, left hand side of Equation (1), i.e., the average instantaneous loss of the iterates, can be bounded with high probability as $\frac{2\lambda^2}{\eta T}$. To get the bound in expectation, as presented in Theorem 4.1,
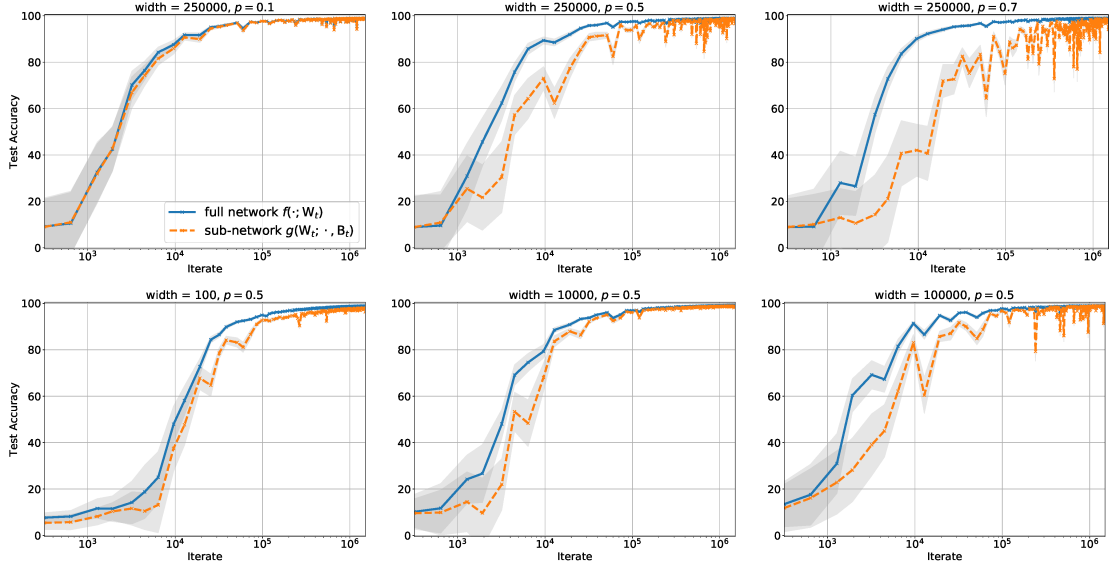
Figure 1: Test accuracy of the full network $f(\cdot; W_t)$ as well as the sub-networks $g(W_t; \cdot, B_t)$ drawn by dropout iterates, as a function of number of iterations t, for (**top**) fixed width $m = 250K$ and several dropout rates $1-p \in \{0.1, 0.5, 0.7\}$; and (**bottom**) fixed dropout rate $1-p = 0.5$ and several widths $m \in \{100, 10K, 100K\}$.

we need to control $L_t^{(t)}(U)$ in worst-case scenario. We take advantage of the max-norm constraints in Algorithm 1, and show in Lemma B.5 that with probability one all iterates satisfy $L_t^{(t)}(U) \leq \frac{c\sqrt{m}}{\ln(2)} + 1$. The proof of Theorem 4.1 then follows from carefully choosing δ, the confidence parameter. To prove the compression result in Theorem 4.2, we use the fact that the zero-one loss can be bounded in terms of the negative derivative of the logistic loss [Cao and Gu, 2019]. Therefore, we can bound $\mathcal{R}(W_t; B_t)$, the population risk of the sub-networks, in terms of $Q(W_t; B_t) = \mathbb{E}_{\mathcal{D}}[-\ell'(y_t g_t(W_t))]$. Following Ji and Telgarsky [2019], the proof of Theorem 4.2 then entails showing that $\sum_{t=1}^{T} Q(W_t; B_t)$ is close to $\sum_{t=1}^{T} Q_t(W_t)$, which itself is bounded by the average instantaneous loss of the iterates.

We now present the main ideas in proving Lemma 5.2, which closely follows Ji and Telgarsky [2019]. Since $L_t^{(t)}(U) \leq e^{-y_t \langle \nabla g_t(W_t), U \rangle}$, the proof entails lower bounding $y_t \langle \nabla g_t(W_t), U \rangle$, which can be decomposed as follows

$$y_t \langle \nabla g_t(W_t), U \rangle = y_t \langle \nabla g_t(W_1), W_1 \rangle + y_t \langle \nabla g_t(W_t) - \nabla g_t(W_1), W_1 \rangle$$
$$+ \lambda y_t \langle \nabla g_t(W_1), V \rangle + \lambda y_t \langle \nabla g_t(W_t) - \nabla g_i(B_t W_1), V \rangle. \quad (2)$$

By homogeneity of the ReLU activations, the first term in Equation (2) precisely computes $y_t g_t(W_1)$, which cannot be too negative under the initialization scheme used in Algorithm 1, as we show in Lemma B.3. On the other hand, we show in Lemma B.4 that under Assumption 1, V has a good margin with respect to the randomly initialized weights W_1, so that the third term in Equation (2) is concentrated around the margin parameter γ. The second and the fourth terms in Equation (2) can be bounded thanks to the lazy regime, where W_t remains close to W_1 at all times. In particular, provided $\|w_{r,t} - w_{r,1}\| \leq \frac{7\lambda}{2\gamma\sqrt{m}}$, we show in Lemma B.1 that at most only $O(1/\sqrt{m})$-fraction of neural activations change, and thus $\nabla g_t(W_t) - \nabla g_t(W_1)$ has a small norm. Lemma 5.2 then follows from carefully choosing λ and m such that the right hand side of Equation (2) is sufficiently positive.

6 Experimental Results

The goal of this section is to investigate if dropout indeed compresses the model, as predicted by Theorem 4.2. In particular, we seek to understand if the (sparse) dropout sub-networks $g(W; \cdot, B)$ – regardless of being visited by dropout during the training – are comparable to the full network $f(\cdot; W)$, in terms of the test accuracy.
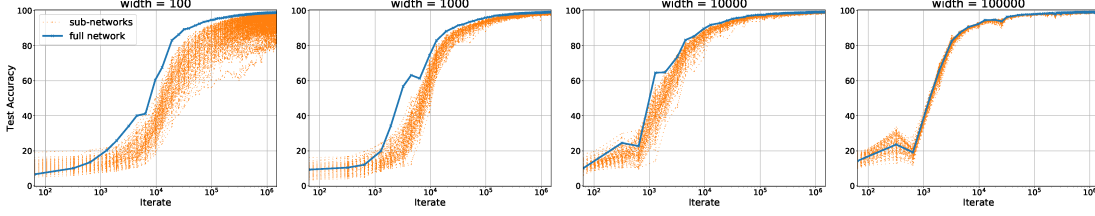
Figure 2: Test accuracy of the full network $f(\cdot; \mathrm{W}_t)$ as well as 100 random sub-networks $g(\mathrm{W}_t; \cdot, \mathrm{B}_1), \ldots, g(\mathrm{W}_t; \cdot, \mathrm{B}_{100})$ with dropout rate $1 - p = 0.8$, as a function of number of iterations t, for several width $m \in \{100, 1K, 10K, 100K\}$.

We train a convolutional neural network with a dropout layer on the top hidden layer, using cross-entropy loss, on the MNIST dataset. The network consists of two convolutional layers with max-pooling, followed by three fully-connected layers. All the activations are ReLU. We use a constant learning rate $\eta = 0.01$ and batch-size equal to 64 for all the experiments. We train several networks where except for the top layer widths ($m \in \{100, 500, 1K, 5K, 10K, 50K, 100K, 250K\}$), all other architectural parameters are fixed. We track the test accuracy over 25 epochs as a function of number of iterations, for the full network, the sub-networks visited by dropout, as well as random but fixed sub-networks that are drawn independently, using the same dropout rate. We run the experiments for several values of the dropout rate, $1 - p \in \{0.1, 0.2, 0.3, \ldots, 0.9\}$.

Figure 1 shows the test accuracy of the full network $f(\cdot; \mathrm{W}_t)$ (blue, solid curve) as well as the dropout iterates $g(\mathrm{W}_t; \cdot, \mathrm{B}_t)$ (orange, dashed curve), as a function of the number of iterations. Both curves are averaged over 50 independent runs of the experiment; the grey region captures the standard deviation. It can be seen that the (sparse) sub-networks drawn by dropout during the training, are indeed comparable to the full network in terms of the generalization error. As expected, the gap between the full network and the sparse sub-networks is higher for narrower networks, and for higher dropout rates. This figure verifies our compression result in Theorem 4.2.

Next, we show that dropout also generalizes to sub-networks that were not observed during the training. In other words, random sub-networks drawn from the same Bernoulli distribution, also performed well. We run the following experiment on the same convolutional network architecture described above with widths $m \in \{100, 1K, 10K, 100K\}$. We draw 100 sub-networks $g(\mathrm{W}; \cdot, \mathrm{B}_1), \ldots, g(\mathrm{W}; \cdot, \mathrm{B}_{100})$, corresponding to diagonal Bernoulli matrices $\mathrm{B}_1, \ldots, \mathrm{B}_{100}$, all generated by the same Bernoulli distribution used at training (Bernoulli parameter $p = 0.2$, i.e., dropout rate $1 - p = 0.8$). In Figure 2, we plot the generalization error of these sub-networks as well as the full network as a function of iteration number, as orange and blue curves, respectively. We observe that, as the width increases, the sub-networks become increasingly more competitive; it is remarkable that the effective width of these *typical* sub-networks are only $\approx 1/5$ of the full network.

7 Conclusion

Most of the results in the literature of over-parameterized neural networks focus on GD-based methods without any explicit regularization. On the other hand, recent theoretical investigations have challenged the view that *implicit bias* due to such GD-based methods can explain generalization in deep learning [Dauber et al., 2020]. Therefore, it seems crucial to explore algorithmic regularization techniques in over-parameterized neural networks. This paper takes a step towards understanding a popular algorithmic regularization technique in deep learning. In particular, assuming that the data distribution is separable in the RKHS induced by the neural tangent kernel, this paper presents precise iteration complexity results for dropout training in two-layer ReLU networks using the logistic loss.

Broader Impact

We investigate the convergence and generalization of a popular algorithmic regularization technique in deep learning. Although we can not think of any direct social impacts per se, we hope such theoretical studies serve the community in long-term, by improving our understanding of the foundations, which shall eventually lead to more powerful machine learning systems.

Acknowledgements

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pages 6155–6166, 2019.

Raman Arora, Peter Bartlett, Poorya Mianjy, and Nathan Srebro. Dropout: Explicit forms and capacity control. *arXiv preprint arXiv:2003.03397*, 2020.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer networks. *arXiv preprint arXiv:1901.08584*, 2019.

Pierre Baldi and Peter J Sadowski. Understanding dropout. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2814–2822, 2013.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10835–10845, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. arxiv e-prints, page. *arXiv preprint arXiv:1812.07956*, 2018.

George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for LVCSR using rectified linear units and dropout. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8609–8613. IEEE, 2013.

Amit Daniely. Sgd learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2017.

Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. *arXiv preprint arXiv:2003.06152*, 2020.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Int. Conf. Machine Learning (ICML)*, 2016.

Wei Gao and Zhi-Hua Zhou. Dropout rademacher complexity of deep neural networks. *Science China Information Sciences*, 59(7):072104, 2016.

Aidan N Gomez, Ivan Zhang, Kevin Swersky, Yarin Gal, and Geoffrey E Hinton. Learning sparse networks using targeted dropout. *arXiv preprint arXiv:1905.13678*, 2019.

David P Helmbold and Philip M Long. On the inductive bias of dropout. *Journal of Machine Learning Research (JMLR)*, 16:3403–3454, 2015.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Wei Hu, Zhiyuan Li, and Dingli Yu. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hke3gyHYwH.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow relu networks. *arXiv preprint arXiv:1909.12292*, 2019.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in neural information processing systems*, pages 8570–8581, 2019.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680*, 2019.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

David McAllester. A PAC-bayesian tutorial with a dropout bound. *arXiv preprint arXiv:1307.2118*, 2013.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing LSTM language models. *arXiv preprint arXiv:1708.02182*, 2017.

Poorya Mianjy and Raman Arora. On dropout and nuclear norm regularization. In *International Conference on Machine Learning*, pages 4575–4584, 2019.

Poorya Mianjy, Raman Arora, and Rene Vidal. On the implicit bias of dropout. In *International Conference on Machine Learning*, pages 3540–3548, 2018.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2498–2507. JMLR. org, 2017.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Atsushi Nitanda and Taiji Suzuki. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Albert Senen-Cerda and Jaron Sanders. Almost sure convergence of dropout algorithms for neural networks. *arXiv preprint arXiv:2002.02247*, 2020.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1), 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Stefan Wager, Sida Wang, and Percy S Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Stefan Wager, William Fithian, Sida Wang, and Percy S Liang. Altitude training: Strong bounds for single-layer dropout. In *Adv. Neural Information Processing Systems*, 2014.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066, 2013.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9709–9721, 2019.

Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. *arXiv preprint arXiv:2002.12915*, 2020.

Ke Zhai and Huan Wang. Adaptive dropout with rademacher complexity regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1uxsye0Z.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

"On Convergence and Generalization of Dropout Training"

A Auxiliary Theorems

Theorem A.1 (Gaussian Concentration Vershynin [2018])**.** Consider a random vector $z \sim \mathcal{N}(0, I_d)$ and a ρ-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ (with respect to the Euclidean metric). Then $f(z)$ is ρ-sub-Gaussian and it holds for all $t \geq 0$:

$$\mathbb{P}\{f(z) - \mathbb{E}[f(z)] \geq t\} \leq e^{\frac{-t^2}{2\rho^2}}$$

Theorem A.2 (Hoeffding's inequality Vershynin [2018])**.** Let $X_1, \ldots, X_n$ be independent, mean zero random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for every $t > 0$, we have

$$\mathbb{P}\{\sum_{i=1}^{n} X_i \geq t\} \leq e^{-\frac{2t^2}{\sum_{i=1}^{n}(m_i - M_i)^2}}$$

Theorem A.3 (Theorem 1 of Beygelzimer et al. [2011])**.** Let $X_1, \ldots, X_T$ be a sequence of real-valued random variables. Let $\mathbb{E}_t[Y] := \mathbb{E}[Y|X_1, \ldots, X_{t-1}]$. Assume, for all t, that $X_t \leq R$ and that $\mathbb{E}_t[X_t] = 0$. Define the random variable $S_t := \sum_{k=1}^{t} X_t$, and $V_t := \sum_{k=1}^{t} \mathbb{E}_k[X_k^2]$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have the following guarantee:

$$S_t \leq R \ln\left(\frac{1}{\delta}\right) + (e - 2)\frac{V_t}{R}$$

B Proofs

The following Lemma bounds $|R_t|$, where $R_t := \{r \in [m]| \ \mathbb{I}\{w_{r,t}^\top x_t > 0\} \neq \mathbb{I}\{w_{r,1}^\top x_t > 0\}\}$ is the set of hidden nodes at time t whose activation on sample x_t is different from the initialization.

Lemma B.1. Assume that $\|w_{r,1} - w_{r,t}\| \leq D$ holds for all $r \in [m]$, where D is a positive constant. Then, with probability at least $1 - \frac{\delta}{3}$, we have that

$$|R_t| \leq mD + \sqrt{\frac{m \ln(3T/\delta)}{2}}, \text{ for all } t \in [T].$$

Proof of Lemma B.1. Assume that $r \in R_t$. Then it holds that

$$
\begin{aligned}
|w_{r,1}^\top x_t| &\leq |w_{r,1}^\top x_t| + |w_{r,t}^\top x_t| \\
&= |(w_{r,1} - w_{r,t})^\top x_t| & (r \in R_t) \\
&\leq \|w_{r,1} - w_{r,t}\|\|x_t\| & \text{(Cauchy-Schwarz)} \\
&= \|w_{r,1} - w_{r,t}\| \leq D & (\|x_t\| = 1)
\end{aligned}
$$

Since $w_{r,1}^\top x_t$ is a standard Gaussian random variable, by anti-concentration property of the Gaussian distribution, $\mathbb{E}[\mathbb{I}\{|w_{r,1}^\top x_t| \leq D\}] = \Pr\{|w_{r,1}^\top x_t| \leq D\} \leq \frac{2D}{\sqrt{2\pi}}$. On the other hand, we have that

$$|R_t| = \left|\{r| \ \mathbb{I}\{w_{r,t}^\top x_t > 0\} \neq \mathbb{I}\{w_{r,1}^\top x_t > 0\}\}\right| \leq \left|\{r| \ |w_{r,1}^\top x_t| \leq D\}\right| = \sum_{r=1}^{m} \mathbb{I}\{|w_{r,1}^\top x_t| \leq D\}$$

By Hoeffding's inequality, we have the following with probability at least $1 - \frac{\delta}{3T}$:

$$\frac{1}{m}\sum_{r=1}^{m} \mathbb{I}\{|w_{r,1}^\top x_t| \leq D\} \leq \Pr\{|w_{r,1}^\top x_t| \leq D\} + \sqrt{\frac{\ln(3T/\delta)}{2m}} \leq \frac{2D}{\sqrt{2\pi}} + \sqrt{\frac{\ln(3T/\delta)}{2m}}.$$

Multiplying both sides by m and applying union bound on $t \in [T]$ completes the proof. □

Lemma B.2. For any $t \in [T]$, let $\mathcal{B}_t := \{\mathrm{B}_1, \ldots, \mathrm{B}_t\}$ denote the set of Bernoulli masks up to time t. Then it holds almost surely that:

$$\sum_{t=1}^{T} \ell(y_t f_t(q\mathrm{W}_t)) \leq \mathbb{E}_{\mathcal{B}_T}[\sum_{t=1}^{T} L_t(\mathrm{W}_t)]. \tag{3}$$

Proof of Lemma B.2. For any $a, b \in \mathbb{R}$, the function $\ell(z) = \log(1 + \exp(az + b))$ is convex in z. We have the following inequalities:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{B}_T}[\sum_{t=1}^{T} L_t(\mathrm{W}_t)] &= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{B}_t}[\ell(y_t \cdot \frac{1}{\sqrt{m}} \mathrm{a}^{\top} \mathrm{B}_t \sigma(\mathrm{W}_t \mathrm{x}_t))] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\mathcal{B}_{t-1}}[\mathbb{E}_{\mathrm{B}_t} \ell(y_t \cdot \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r b_{r,t} \sigma(\mathrm{w}_{r,t}^{\top} \mathrm{x}_t))|\mathcal{B}_{t-1}] \quad \text{(smoothing property)} \\
&\geq \sum_{t=1}^{T} \mathbb{E}_{\mathcal{B}_{t-1}}[\ell(y_t \cdot \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \mathbb{E}_{\mathrm{B}_t}[b_{r,t}] \sigma(\mathrm{w}_{r,t}^{\top} \mathrm{x}_t))|\mathcal{B}_{t-1}] \quad \text{(Jensen's inequality)} \\
&= \sum_{t=1}^{T} \ell(y_t \cdot \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(q \mathrm{w}_{r,t}^{\top} \mathrm{x}_t)) \quad\quad (\mathbb{E}[b_{r,t}] = q, \text{homogeneity of ReLU}) \\
&= \sum_{t=1}^{T} \ell(y_t f_t(q\mathrm{W}_t))
\end{aligned}
$$

which completes the proof. $\square$

Proof of Lemma 5.1. Using the dropout update rule in Algorithm 1, we start by analyzing the distance of consecutive iterates from the reference point U, assuming that $\Pi_c(\mathrm{U}) = \mathrm{U}$:

$$
\begin{aligned}
\|\mathrm{W}_{t+1} - \mathrm{U}\|_F^2 &= \|\Pi_c(\mathrm{W}_{t+\frac{1}{2}}) - \mathrm{U}\|_F^2 \\
&\leq \|\mathrm{W}_{t+\frac{1}{2}} - \mathrm{U}\|_F^2 \quad\quad (\mathrm{U} \in \mathcal{W}_c) \\
&= \|\mathrm{W}_t - \eta \nabla L_t(\mathrm{W}_t) - \mathrm{U}\|_F^2 \\
&= \|\mathrm{W}_t - \mathrm{U}\|_F^2 - 2\eta \langle \nabla L_t(\mathrm{W}_t), \mathrm{W}_t - \mathrm{U} \rangle + \eta^2 \|\nabla L_t(\mathrm{W}_t)\|_F^2
\end{aligned}
$$

The last term on the right hand side above is bounded as follows:

$$
\begin{aligned}
\eta^2 \|\nabla L_t(\mathrm{W}_t)\|_F^2 &= \eta^2 \|\ell'(y_t g_t(\mathrm{W}_t)) y_t \nabla g_t(\mathrm{W}_t)\|_F^2 \\
&= \eta^2 \left(-\ell'(y_t g_t(\mathrm{W}_t)) \|\nabla g_t(\mathrm{W}_t)\|_F\right)^2 \\
&= \eta^2 Q_t(\mathrm{W}_t)^2 \sum_{r=1}^{m} \|\frac{\partial g_t(\mathrm{W}_t)}{\partial \mathrm{w}_{r,t}}\|^2 \\
&\leq \eta^2 Q_t(\mathrm{W}_t)^2 \quad\quad (\|\frac{\partial g_t(\mathrm{W}_t)}{\partial \mathrm{w}_{r,r}}\| \leq \frac{1}{\sqrt{m}}) \\
&\leq \frac{\eta^2}{\ln(2)} Q_t(\mathrm{W}_t) \quad\quad (Q_t(\cdot) \leq 1/\ln(2)) \\
&\leq \eta Q_t(\mathrm{W}_t) \quad\quad (\text{assumption } \eta \leq \ln(2)) \\
&\leq \eta L_t(\mathrm{W}_t) \quad\quad (Q_t(\cdot) \leq L_t(\cdot))
\end{aligned}
$$

The second term can be bounded as follows:

$$
\begin{aligned}
\langle \nabla L_t(\mathrm{W}_t), \mathrm{W}_t - \mathrm{U} \rangle &= \ell'(y_t g_t(\mathrm{W}_t)) \langle y_t \nabla g_t(\mathrm{W}_t), \mathrm{W}_t - \mathrm{U} \rangle \\
&= \ell'(y_t g_t(\mathrm{W}_t))(y_t g_t(\mathrm{W}_t) - y_t g_t^{(t)}(\mathrm{U})) \quad \text{(Homogeneity, definition of } g_t^{(t)}) \\
&\geq (\ell(y_t g_t(\mathrm{W}_t)) - \ell(y_t g_t^{(t)}(\mathrm{U}))) \quad\quad (\text{convexity of } \ell(\cdot)) \\
&= L_t(\mathrm{W}_t) - L_t^{(t)}(\mathrm{U})
\end{aligned}
$$

13

Plugging back the above inequalities we get

$$\|W_{t+1} - U\|_F^2 \leq \|W_{t+\frac{1}{2}} - U\|_F^2 \leq \|W_t - U\|_F^2 - 2\eta(L_t(W_t) - L_t^{(t)}(U)) + \eta L_t(W_t)$$

$$= \|W_t - U\|_F^2 - \eta L_t(W_t) + 2\eta L_t^{(t)}(U) \tag{4}$$

Rearranging, dividing both sides by η, and averaging over iterates we arrive at

$$\frac{1}{T} \sum_{t=1}^{T} L_t(W_t) \leq \sum_{t=1}^{T} \frac{\|W_t - U\|_F^2 - \|W_{t+1} - U\|_F^2}{\eta T} + \frac{2}{T} \sum_{t=1}^{T} L_t^{(t)}(U)$$

$$\leq \frac{\|W_1 - U\|_F^2}{\eta T} + \frac{2}{T} \sum_{t=1}^{T} L_t^{(t)}(U) \qquad \text{(Telescopic sum)}$$

$\square$

Lemma B.3. With probability at least $1 - \delta/3$ it holds uniformly over all $t \in [T]$ that $|g_t(W_1)| \leq \sqrt{2 \ln(6T/\delta)}$, provided that $m \geq 25 \ln(6T/\delta)$.

Proof of Lemma B.3. The proof is similar to the proof of Lemma A.1 in Ji and Telgarsky [2019], except for that we have to take into account the randomness due to dropout as well. In particular, there are four different sources of randomness in $g_t(W_1) = g(W_1; x_t, B_t)$: 1) the randomly initialized hidden layer weights W_1, 2) the randomly initialized top layer weights a, 3) the input vector x_t, $t \in [T]$, and 4) the Bernoulli masks B_t, $t \in [T]$. Given input x_t and the dropout mask B_t, let $h_t(W) = \frac{1}{\sqrt{m}} B_t \sigma(Wx_t) \in \mathbb{R}^m$ denote the (scaled) output of the dropout layer with hidden weights W. It is easy to see that the function $g : W \mapsto \|h_t(W)\|$ is 1-Lipschitz:

$$|g(W) - g(W')| = |\|h_t(W)\| - \|h_t(W')\||$$

$$\leq \|h_t(W) - h_t(W')\| \qquad \text{(Reverse Triangle Inequality)}$$

$$= \sqrt{\sum_{r=1}^{m} (\frac{1}{\sqrt{m}} b_i^{(t)} \sigma(\langle w_{r,1}, x_t \rangle) - \frac{1}{\sqrt{m}} b_i^{(t)} \sigma(\langle w'_{r,1}, x_t \rangle))^2}$$

$$= \frac{\sqrt{\sum_{r=1}^{m} (\langle w_{r,1}, x_t \rangle - \langle w'_{r,1}, x_t \rangle)^2}}{\sqrt{m}} \qquad \text{(1-Lipschitzness of ReLU)}$$

$$\leq \frac{\sqrt{\sum_{r=1}^{m} \|w_{r,1} - w'_{r,1}\|^2 \|x_t\|^2}}{\sqrt{m}} \qquad \text{(Cauchy-Schwarz)}$$

$$= \frac{\|W - W'\|_F}{\sqrt{m}}$$

Using Gaussian concentration (Lemma A.1), we get that $\|h_t(W_1)\| - \mathbb{E}_{W_1}[\|h_t(W_1)\|] \leq \sqrt{\frac{2 \ln\left(\frac{6T}{\delta}\right)}{m}}$ with probability at least $1 - \frac{\delta}{6T}$. It also holds that:

$$\mathbb{E}_{W_1}[\|h_t(W_1)\|] \leq \sqrt{\mathbb{E}_{W_1}[\|h_t(W_1)\|^2]}$$

$$= \sqrt{\sum_{r=1}^{m} \mathbb{E}_{w_{r,1}} (\frac{1}{\sqrt{m}} b_{r,t} \sigma(w_{r,1}^\top x_t))^2}$$

$$\leq \sqrt{\frac{\sum_{r=1}^{m} \mathbb{E}_{w_{r,1}}[\sigma(w_{r,1}^\top x_t)^2]}{m}}$$

$$= \sqrt{\mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma(z)^2]} = \frac{1}{\sqrt{2}}$$

As a result, we have with probability at least $1 - \frac{\delta}{6T}$ that $\|h_t(W_1)\| \leq \sqrt{\frac{2 \ln(6T/\delta)}{m}} + \frac{\sqrt{2}}{2} \leq 1$ whenever $m \geq 25 \ln(6T/\delta)$. Now, taking a union bound over all $t \in [T]$, we get that $\|h_t(W_1)\| \leq 1$ holds

14

simultaneously for all iterates. Conditioned on this event, the random variable $g_t(W_1) = \langle a, h_t(W_1) \rangle$ is zero mean and 1-sub-Gaussian, so that by the general Hoeffding's inequality, for any t, with probability at least $1 - \frac{\delta}{6T}$, it holds that $|g_t(W_1)| \le \sqrt{2 \ln (6T/\delta)}$. Taking union bound over all $t \in [T]$, with probability $1 - \delta/6$ it holds that $|g_t(W_1)| \le \sqrt{2 \ln (6T/\delta)}$ simultaneously for all $t \in [T]$. Finally, the probability that both of these events hold is no less than $(1 - \delta/6)^2 \ge 1 - \delta/3$, which completes the proof. $\qquad\square$

Lemma B.4. Under Assumption 1, for any $\delta \in (0, 1)$, with probability at least $1 - \delta/3$ it holds uniformly for all $t \in [T]$ that:

$$y_t g_t^{(1)}(V) = y_t \langle \nabla g_t(W_1), V \rangle \ge \gamma - \sqrt{\frac{2 \ln (3T/\delta)}{m}}$$

Proof of Lemma B.4. By Assumption 1, it holds that $\mathbb{E}_{z,b}[y\langle \psi(z), bx\mathbb{I}\{z^\top x > 0\}\rangle] \ge \gamma$ for all (x, y) in the domain of $\mathcal{D}$. We observe that $y_t g_t^{(1)}(V)$ is an empirical estimate of this quantity:

$$
\begin{aligned}
y_t g_t^{(1)}(V) &= y_t \langle \nabla g_t(W_1), V \rangle \\
&= y_t \sum_{r=1}^{m} \langle \frac{1}{\sqrt{m}} a_r b_{r,t} \mathbb{I}\{x_t^\top w_{r,1} > 0\} x_t, \frac{1}{\sqrt{m}} a_r \psi(w_{r,1}) \rangle \\
&= \frac{1}{m} \sum_{r=1}^{m} y_t \langle \psi(w_{r,1}), b_{r,t} x_t \mathbb{I}\{w_{r,1}^\top x_t > 0\} \rangle
\end{aligned}
$$

For $t, r \in [T] \times [m]$, let $\gamma_{t,r} := y_t \langle \psi(w_{r,1}), b_{r,t} x_t \mathbb{I}\{w_{r,1}^\top x_t > 0\} \rangle$. Note that $\mathbb{E}_{W_1, B_t}[\gamma_{t,r}] = \mathbb{E}_{z,b}[y_t \langle \psi(z), bx_t \mathbb{I}\{z^\top x_t > 0\} \rangle]$. Also, for any t, the random variable $\gamma_{t,r}$ is bounded almost surely as follows:

$$|\gamma_{t,r}| \le |y_t| \, \|\psi(w_{r,1})\| \, |b_{r,t}| \, \|x_t\| \, \left|\mathbb{I}\{w_{r,1}^\top x_t > 0\}\right| \le 1.$$

Therefore by Hoeffding's inequality (Theorem A.2), with probability at least $1 - \frac{\delta}{3T}$, it holds that:

$$y_t g_t^{(1)}(V) - \gamma \ge y_t g_t^{(1)}(V) - \mathbb{E}[y_t g_t^{(1)}(V)] \ge -\sqrt{\frac{2 \ln (3T/\delta)}{m}}$$

Applying a union bound over t finishes the proof. $\qquad\square$

Proof of Lemma 5.2. We adopt the proof of Theorem 2.2 in Ji and Telgarsky [2019] for dropout training. Assume that $\|w_{r,t} - w_{r,1}\| \le \frac{7\lambda}{2\gamma\sqrt{m}}$ holds for the first T iterates of Algorithm 1. Then with probability at least $1 - (\frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3}) = 1 - \delta$, Lemma B.1, Lemma B.3, and Lemma B.4 hold simultaneously. We first prove that $L_t^{(t)}(U) \le \frac{\lambda^2}{2\eta T}$ for all $t \in [T]$. Using the inequality $\log(1 + z) \le z$, we get that

$$L_t^{(t)}(U) = \log(1 + e^{-y_t \langle \nabla g_t(W_t), U \rangle}) \le e^{-y_t \langle \nabla g_t(W_t), U \rangle}$$

To upper-bound the right hand side, we lower-bound $y_t \langle \nabla g_t(W_t), U \rangle$. By definition of U, we have

$$y_t \langle \nabla g_t(W_t), U \rangle = y_t \langle \nabla g_t(W_t), W_1 \rangle + \lambda y_t \langle \nabla g_t(W_t), V \rangle \qquad (5)$$

We bound each of the terms separately. The first term can be decomposed as follows:

$$
\begin{aligned}
y_t \langle \nabla g_t(W_t), W_1 \rangle &= y_t \langle \nabla g_t(W_1), W_1 \rangle + y_t \langle \nabla g_t(W_t) - \nabla g_t(W_1), W_1 \rangle \\
&\ge -|y_t g_t(W_1)| - |y_t \langle \nabla g_t(W_t) - \nabla g_t(W_1), W_1 \rangle| \qquad (6)
\end{aligned}
$$

By Lemma B.3, the first term on right hand side is lower-bounded by $-|g_t(W_1)| \geq -\sqrt{2\ln(6T/\delta)}$. We bound the second term as follows:

$$|y_t\langle\nabla g_t(W_t) - \nabla g_t(W_1), W_1\rangle| = \left|\frac{y_t}{\sqrt{m}}\sum_{r=1}^{m} a_r b_{r,t}(\mathbb{I}\{w_{r,t}^\top x_t > 0\} - \mathbb{I}\{w_{r,1}^\top x_t > 0\})w_{r,1}^\top x_t\right|$$

$$\leq \frac{1}{\sqrt{m}}\sum_{r\in R_t}|a_r b_{r,t}\langle w_{r,1}, x_t\rangle| \qquad \text{(Triangle inequality)}$$

$$\leq \frac{1}{\sqrt{m}}\sum_{r\in R_t}|\langle w_{r,t} - w_{r,1}, x_t\rangle| \qquad (r\in R_t)$$

$$\leq \frac{|R_t|\,\|w_{r,t} - w_{r,1}\|}{\sqrt{m}}$$

$$\leq \frac{49\lambda^2}{4\gamma^2\sqrt{m}} + \sqrt{\frac{49\lambda^2\ln(3T/\delta)}{8\gamma^2 m}} \qquad \text{(Lemma B.1)}$$

$$\leq \frac{\lambda\gamma}{2} \qquad (7)$$

where the last inequality holds when $m \geq \max\{98\gamma^{-4}\ln(3T/\delta), 2401\gamma^{-6}\lambda^2\} = 2401\gamma^{-6}\lambda^2$. The second term in Equation 5 is bounded as follows:

$$y_t\langle\nabla g_t(W_t), V\rangle = y_t\langle\nabla g_t(W_1), V\rangle + y_t\langle\nabla g_t(W_t) - \nabla g_t(W_1), V\rangle$$

$$\geq y_t\langle\nabla g_t(W_1), V\rangle - |y_t\langle\nabla g_t(W_t) - \nabla g_t(W_1), V\rangle|$$

$$= y_t g_t^{(1)}(V) - \left|\frac{y_t}{\sqrt{m}}\sum_{r=1}^{m}a_r b_{r,t}(\mathbb{I}\{w_{r,t}^\top x_t > 0\} - \mathbb{I}\{w_{r,1}^\top x_t > 0\})\langle\frac{1}{\sqrt{m}}a_r\psi(w_{r,1}), x_t\rangle\right|$$

$$\geq \gamma - \sqrt{\frac{2\ln(3T/\delta)}{m}} - \frac{1}{m}\sum_{r\in R_t}|a_r b_{r,t}\langle\psi(w_{r,1}), x_t\rangle| \qquad \text{(Lemma B.4)}$$

$$\geq \gamma - \sqrt{\frac{2\ln(3T/\delta)}{m}} - \frac{|R_t|}{m}$$

$$\geq \gamma - \sqrt{\frac{2\ln(3T/\delta)}{m}} - \frac{7\lambda}{2\gamma\sqrt{m}} - \sqrt{\frac{\ln(3T/\delta)}{2m}} \qquad \text{(Lemma B.1)}$$

$$\geq \gamma - \frac{\gamma^2}{7} - \frac{\gamma^2}{14} - \frac{\gamma^2}{14} = \gamma - \frac{2\gamma^2}{7} \geq \frac{5\gamma}{7} \qquad (8)$$

where the penultimate inequality holds when $m \geq \max\{98\gamma^{-4}\ln(3T/\delta), 2401\gamma^{-6}\lambda^2\} = 2401\gamma^{-6}\lambda^2$. Plugging Equations (7) and (8) in Equation 5, we get that

$$y_t\langle\nabla g_t(W_t), U\rangle \geq -\sqrt{2\ln(6T/\delta)} + \frac{3\lambda\gamma}{14} \geq \ln\left(\frac{2\eta T}{\lambda^2}\right), \qquad (9)$$

where the inequality hold true for $\lambda := 5\gamma^{-1}\ln(2\eta T) + \sqrt{44\gamma^{-2}\ln(6T/\delta)}$. Thus, we have that

$$L_t^{(t)}(U) = \log(1 + e^{-y_t\langle\nabla g_t(W_t), U\rangle}) \leq \frac{\lambda^2}{2\eta T}.$$

We now prove by induction that $\|w_{r,t} - w_{r,1}\| \leq \frac{7\lambda}{2\gamma\sqrt{m}}$ holds throughout dropout training. First, we show that the claim holds for $t = 2$:

$$\|w_{r,2} - w_{r,1}\| = \|\Pi_c(\eta\frac{\partial L_t(B_1 W_1)}{\partial w_{r,1}})\| \leq \|\eta\frac{\partial L_t(B_1 W_1)}{\partial w_{r,1}}\|$$

$$\leq \|\eta\ell'(y_t f_t(B_1 W_1))y_i\frac{\partial f_t(B_1 W_1)}{\partial w_{r,1}}\|$$

$$\leq \frac{\eta}{\ln(2)\sqrt{m}} \leq \frac{7\lambda}{2\gamma\sqrt{m}}, \qquad (\eta \leq \ln(2))$$

16

which proves the basic step. Now by induction hypothesis, we assume that the claim holds for all $k \in [t]$, i.e., it holds that $\|w_{r,k} - w_{r,1}\| \leq \frac{7\lambda}{2\gamma\sqrt{m}}$. Therefore, it holds that $\|w_{r,k}\| \leq \|w_{r,1}\| + \|w_{r,k} - w_{r,1}\| \leq c - 1 + 1 \leq c$, where we used the triangle inequality, the fact that $\|w_{r,1}\| \leq c - 1$, and that $m \geq 2401\gamma^{-6}\lambda^2$. This, in particular, means that all iterates $1 < k \leq t$ remain in $\mathcal{W}_c$:

$$W_k = \Pi_c(W_{k-\frac{1}{2}}) = W_{k-\frac{1}{2}} \text{ for all } 1 < k \leq t. \tag{10}$$

For the $t+1$-th iterate, we first upper-bound the distance from initialization in terms of the Q function:

$$\|w_{r,t+1} - w_{r,1}\| = \|\Pi_c(w_{r,t} - \eta\frac{\partial L_t(W_t)}{\partial w_{r,t}}) - w_{r,1}\|$$

$$\leq \|w_{r,t} - \eta\frac{\partial L_t(W_t)}{\partial w_{r,t}} - w_{r,1}\|$$

$$\leq \|\eta\frac{\partial L_t(W_t)}{\partial w_{r,t}}\| + \|w_{r,t} - w_{r,1}\|$$

$$\leq \sum_{k=1}^{t} \|\eta\frac{\partial L_k(W_k)}{\partial w_{r,k}}\|$$

$$\leq \eta\sum_{k=1}^{t} -\ell'(y_k g_k(W_k))\|y_k\frac{\partial g_k(W_k)}{\partial w_{r,k}}\|$$

$$\leq \frac{\eta}{\sqrt{m}}\sum_{k=1}^{t} -\ell'(y_k g_k(W_k))$$

The idea is to turn the right hand side above into a telescopic sum using the identity $W_{k+1} - W_k = W_{k+\frac{1}{2}} - W_k = \eta\ell'(y_k g_k(W_k))y_k\nabla g_k(W_k)$, $k \in [t-1]$. By induction hypothesis, for all $k \in [t]$, Equation (8) guarantees $y_k\langle\nabla g_k(W_k), V\rangle \geq \frac{5\gamma}{7}$. Thus, multiplying the right hand side of (11) by $\frac{7}{5\gamma}y_k\langle\nabla g_k(W_k), V\rangle$, we get that:

$$\|w_{r,t+1} - w_{r,1}\| \leq \frac{7\eta}{5\gamma\sqrt{m}}\sum_{k=1}^{t} -\ell'(y_k g_k(W_k))y_k\langle\nabla g_k(W_k), V\rangle$$

$$= \frac{7}{5\gamma\sqrt{m}}\sum_{k=1}^{t} \langle\eta\nabla L_k(W_k), V\rangle$$

$$= \frac{7}{5\gamma\sqrt{m}}\langle W_{t+\frac{1}{2}} - W_1, V\rangle \qquad \text{(Equation (10))}$$

$$= \frac{7\langle W_{t+\frac{1}{2}} - U, V\rangle + 7\langle U - W_1, V\rangle}{5\gamma\sqrt{m}}$$

$$\leq \frac{7\|W_{t+\frac{1}{2}} - U\|_F\|V\|_F + 7\langle\lambda V, V\rangle}{5\gamma\sqrt{m}} \qquad \text{(Cauchy-Schwarz)}$$

$$\leq \frac{7\|W_{t+\frac{1}{2}} - U\|_F + 7\lambda}{5\gamma\sqrt{m}} \tag{11}$$

17

Again by induction hypothesis, Equation ([4]) and Equation ([9]) hold for all $k \in [t]$, which are used to bound $\|W_{t+\frac{1}{2}} - U\|_F$ as follows:

$$\|W_{t+\frac{1}{2}} - U\|_F^2 \leq \|W_t - U\|_F^2 - 2\eta(L_t(W_t) - L_t^{(t)}(U)) + \eta L_t(W_t) \qquad \text{(Equation ([4]))}$$

$$\leq \|W_t - U\|_F^2 + 2\eta L_t^{(t)}(U))$$

$$\leq \|W_1 - U\|_F^2 + 2\eta \sum_{k=1}^{t} L_k^{(k)}(U)$$

$$\leq \|\lambda V\|_F^2 + 2\eta t \frac{\lambda^2}{2\eta T} \qquad \text{(Equation [9])}$$

$$\leq \lambda^2 + \frac{\lambda^2 t}{T} \qquad (\|V\|_F \leq 1)$$

$$\leq 2\lambda^2$$

$$\implies \|W_{t+\frac{1}{2}} - U\|_F \leq \sqrt{2}\lambda \qquad (12)$$

Plugging Equation ([12]) back in Equation ([11]), we arrive at:

$$\|w_{r,t+1} - w_{r,1}\| \leq \frac{7\sqrt{2}\lambda + 7\lambda}{5\gamma\sqrt{m}} \leq \frac{7\lambda}{2\gamma\sqrt{m}}$$
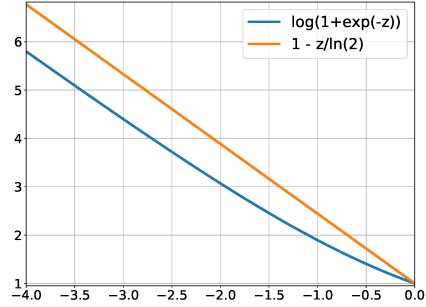
Which completes the induction step and the proof. $\qquad \square$

A crucial step in giving generalization bounds in expectation via upper-bounding the logistic loss is to control the maximum value the loss can take on any iterate of the algorithm. In particular, we need to upper-bound the instantaneous loss of $g_t^{(t)}(U)$, which appears in the right hand side of Lemma [5.1]. To that end, we note that the logistic loss only grows linearly for $z < 0$. More formally, it holds for all $z < 0$ that:

$$\log(1 + e^{-z}) \leq \frac{-z}{\ln(2)} + 1 \qquad (13)$$

as depicted in Figure [B].

Lemma B.5. Under Algorithm [1], it holds with probability one for all iterates that $L_t^{(t)}(U) \leq \frac{c\sqrt{m}}{\ln(2)} + 1$.

Proof of Lemma B.5. Recall that $L_t^{(t)}(U) = \ell(y_t g_t^{(t)}(U))$. First we bound the argument inside the loss function:

$$\left| y_t g_t^{(t)}(W_t) \right| = |y_t \langle \nabla g_t(W_t), U \rangle|$$

$$\leq \sum_{r=1}^{m} \left| \langle \frac{\partial g_t(W_t)}{\partial w_{r,t}}, u_r \rangle \right| \qquad \text{(triangle inequality)}$$

$$\leq \sum_{r=1}^{m} \|\frac{\partial g_t(W_t)}{\partial w_{r,t}}\| \|w_{r,1} + \lambda v_r\| \qquad \text{(Cauchy-Schwarz)}$$

$$\leq \sum_{r=1}^{m} \frac{c - 1 + \lambda/\sqrt{m}}{\sqrt{m}} \qquad (\|w_{r,1}\| \leq c - 1, \|v_r\| \leq 1/\sqrt{m})$$

$$\leq c\sqrt{m} \qquad (\lambda \leq \sqrt{m})$$

Now using Equation ([13]), we get that

$$L_t^{(t)}(U) = \log(1 + e^{-y_t \langle \nabla g_t(W_t), U \rangle}) \leq \log(1 + \exp(c\sqrt{m})) \leq \frac{c\sqrt{m}}{\ln(2)} + 1.$$

$\qquad \square$

Proof of Theorem 4.1. Note that W_t is conditionally independent from x_t given $x_1, \ldots, x_{t-1}$. Thus,

$$\mathbb{E}_{\mathcal{S}_T}[L_t(W_t)] = \mathbb{E}_{\mathcal{S}_{t-1}}[\mathbb{E}_{(x_t, y_t)}\ell(y_t f_t(W_t))|\mathcal{S}_{t-1}] = \mathbb{E}_{\mathcal{S}_{t-1}}[L(W_t)]$$

Using the fact that logistic loss upper-bounds the zero-one loss, taking expectation over initialization, taking average over iterates, and using Lemma B.2, we get that:

$$\mathbb{E}_{W_1, a, \mathcal{S}_T}[\frac{1}{T}\sum_{t=1}^{T}\mathcal{R}(qW_t)] \leq \mathbb{E}_{W_1, a, \mathcal{S}_T}[\frac{1}{T}\sum_{t=1}^{T}\ell(y_t f_t(qW_t))] \qquad (\mathbb{I}\{z < 0\} \leq \ell(z))$$

$$\leq \mathbb{E}_{W_1, a, \mathcal{S}_T, \mathcal{B}_T}[\frac{1}{T}\sum_{t=1}^{T}L_t(W_t)] \qquad (\text{Lemma B.2})$$

$$\leq \frac{\mathbb{E}_{W_1}[\|W_1 - U\|_F^2]}{2\eta T} + \frac{2}{T}\sum_{t=1}^{T}\mathbb{E}_{W_1, a, \mathcal{S}_T, \mathcal{B}_T}[L_t^{(t)}(U)] \qquad (\text{Lemma 5.1})$$

The first term is upper-bounded by $\frac{\lambda^2}{2\eta T}$ since $\|W_1 - U\|_F^2 = \|W_1 - W_1 - \lambda V\|_F^2 = \lambda^2\|V\|_F^2 \leq \lambda^2$. Bounding the second term is based on the following two facts:

1. By Lemma 5.2, with probability at least $1 - \delta$, it holds that $L_t^{(t)}(U) \leq \frac{\lambda^2}{2\eta T} =: u_1$.

2. By Lemma B.5, it holds with probability one that $L_t^{(t)}(U) \leq \frac{c\sqrt{m}}{\ln(2)} + 1 \leq 2c\sqrt{m} =: u_2$.

Therefore, the expected value of $L_t^{(t)}(U)$ can be upper-bounded as:

$$\mathbb{E}[L_t^{(t)}(U)] \leq (1 - \delta)u_1 + \delta u_2 \leq \frac{\lambda^2}{2\eta T} + 2\delta c\sqrt{m}$$

Choosing $\delta := \frac{1}{4\eta c\sqrt{m}T}$ guarantees that

$$\mathbb{E}[L_t^{(t)}(U)] \leq \frac{\lambda^2}{2\eta T} + \frac{1}{2\eta T} \leq \frac{\lambda^2}{\eta T},$$

where $\lambda := 5\gamma^{-1}\ln(2\eta T) + \sqrt{44\gamma^{-2}\ln(24\eta c\sqrt{m}T^2)}$. $\qquad\qquad\square$

Proof of Theorem 4.2. First, recall the following property of the logistic loss:

$$\mathbb{I}\{z < 0\} \leq -2\ln(2)\ell'(z) \leq 2\ln(2)\ell(z)$$

which implies that $\mathcal{R}(W_t; B_t) \leq 2\ln(2)Q(W_t; B_t)$, where $Q(W; B) := \mathbb{E}_{\mathcal{D}}[-\ell'(yg(W; x, B)]$ is the expected value of the negative derivative of the logistic loss. On the other hand, taking the empirical average over the training data, and using Lemma 5.1 and Lemma 5.2, we conclude that the following holds with probability at least $1 - \delta$:

$$\frac{1}{T}\sum_{t=1}^{T}Q_t(W_t) \leq \frac{1}{T}\sum_{t=1}^{T}L_t(W_t)$$

$$\leq \frac{\|W_1 - U\|_F^2}{\eta T} + \frac{2}{T}\sum_{t=1}^{T}L_t^{(t)}(U) \qquad (\text{Lemma 5.1})$$

$$\leq \frac{\lambda^2}{\eta T} + \frac{2}{T}\sum_{t=1}^{T}\frac{\lambda^2}{2\eta T} \qquad (\text{Lemma 5.2})$$

$$\leq \frac{2\lambda^2}{\eta T}.$$

Given the dropout masks $\mathcal{B}_T$, since $Q(W_t; B_t) = \mathbb{E}_{\mathcal{D}}[Q_t(W_t)]$, we know that $\sum_{t=1}^{T}Q(W_t; B_t) - \sum_{t=1}^{T}Q_t(W_t)$ is a martingale difference with respect to the past observations, $\mathcal{S}_{T-1}$. We next show that the average of $Q_t(W_t)$ on the right hand side above is close to the average of $Q(W_t; B_t)$, using

Theorem A.3, similar to Lemma 4.3. of Ji and Telgarsky [2019]. First, this martingale difference sequence is bounded almost surely as $R := 1/\ln(2) \geq Q(W_t; B_t) - Q_t(W_t)$, simply because $0 \leq -\ell'(z) \leq 1/\ln(2)$. The conditional variance can be bounded as:

$$
\begin{aligned}
V_t &:= \sum_{t=1}^{T} \mathbb{E}[(Q(W_t; B_t) - Q_t(W_t))^2 | \mathcal{S}_{t-1}] \\
&= \sum_{t=1}^{T} Q(W_t; B_t)^2 - 2Q(W_t; B_t)\mathbb{E}[Q_t(W_t)|\mathcal{S}_{t-1}] + \mathbb{E}[Q_t(W_t)^2|\mathcal{S}_{t-1}] \\
&\leq \sum_{t=1}^{T} \mathbb{E}[Q_t(W_t)^2|\mathcal{S}_{t-1}] && (\mathbb{E}[Q_t(W_t)|\mathcal{S}_{t-1}] = Q(W_t; B_t)) \\
&\leq \frac{1}{\ln(2)} \sum_{t=1}^{T} \mathbb{E}[Q_t(W_t)|\mathcal{S}_{t-1}] && (0 \leq Q_t(W_t) \leq 1/\ln(2)) \\
&= \frac{1}{\ln(2)} \sum_{t=1}^{T} Q(W_t; B_t)
\end{aligned}
$$

Thus, using Theorem A.3 with $R \leq 1/\ln(2)$ and $V_t \leq \sum_{t=1}^{T} Q(W_t; B_t)/\ln(2)$, we conclude that with probability at least $1 - \delta$ it holds that

$$
\sum_{t=1}^{T} Q(W_t; B_t) - \sum_{t=1}^{T} Q_t(W_t) \leq (e - 2) \sum_{t=1}^{T} Q(W_t; B_t) + \frac{\ln(1/\delta)}{\ln(2)}
$$

$$
\implies \frac{1}{T} \sum_{t=1}^{T} Q(W_t; B_t) \leq \frac{4}{T} \sum_{t=1}^{T} Q_t(W_t) + \frac{4\log(1/\delta)}{T}
$$

Plugging the above back in $\mathcal{R}(W_t; B_t) \leq 2\ln(2) Q(W_t; B_t)$, and averaging over iterates we have:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathcal{R}(W_t; B_t) \leq \frac{16\ln(2)\lambda^2}{\eta T} + \frac{8\ln(2)\ln(1/\delta)}{T}
$$

which completes the proof. $\qquad\square$