Topologically Densified Distributions

Christoph D. Hofer [1] **Florian Graf** [1] **Marc Niethammer** [2] **Roland Kwitt** [1]

Abstract

We study regularization in the context of small sample-size learning with over-parameterized neural networks. Specifically, we shift focus from architectural properties, such as norms on the network weights, to properties of the internal representations before a linear classifier. Specifically, we impose a topological constraint on samples drawn from the probability measure induced in that space. This provably leads to mass concentration effects around the representations of training instances, i.e., a property beneficial for generalization. By leveraging previous work to impose topological constraints in a neural network setting, we provide empirical evidence (across various vision benchmarks) to support our claim for better generalization.

1. Introduction

Learning neural network predictors for complex tasks typically requires large amounts of data. Although such models are over-parameterized, they generalize well in practice. The mechanisms that govern generalization in such settings are still only partially understood (Zhang et al., 2017). Existing generalization bounds (Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2018; Arora et al., 2018) offer deeper insights, yet the vacuity of the bounds and their surprising behavior in terms of sample size (Nagarajan & Kolter, 2019) is a lasting concern.

In *small sample-size regimes*, achieving generalization is considerably more challenging and, in general, requires careful regularization, e.g., via various norms on the network weights, controlling the Lipschitz constants, or adaptation and adjustment of the training data. The latter does not exert regularization on parts of the function implemented by the network, but acts on the training data, e.g., by regularizing its internal representations. Prominent examples include modern augmentation techniques (Cubuk et al., 2019), or mixing strategies (Verma et al., 2019a) to control overconfident predictions. Not only do these approaches show remarkable practical success, but, to some extent, can also be legitimized formally, e.g., through *flattening* arguments in the representation space, or through variance reduction arguments, as in case of data augmentation (Dao et al., 2019).

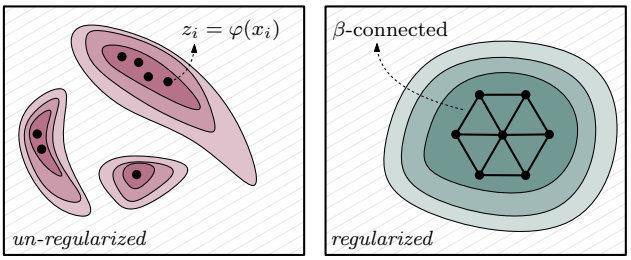


Figure 1. Illustration of how topological regularization affects probability mass concentration in the internal representation space of a neural network. Darker shading denotes regions of higher probability mass.

In this work, we contribute a regularization approach that hinges on *internal representations*. We consider neural networks as a functional composition of the form

$$\gamma \circ \varphi : \mathcal{X} \to \mathcal{Y} = \{1, \dots, K\} \ , \tag{1}$$

where $\varphi : \mathcal{X} \to \mathcal{Z}$ denotes a high-capacity feature extractor which maps into an internal representation space $\mathcal{Z}$. A linear classifier $\gamma : \mathcal{Z} \to \mathcal{Y}$ then predicts one of K classes. As customary, γ is typically of the form $Az + b$, followed by the argmax operator. In our setting, we focus on the representation space $\mathcal{Z}$ and, specifically, on the *push-forward probability measure* induced by φ on $\mathcal{Z}$. We then identify a *property of this measure that is beneficial to generalization*. This shifts attention away from properties of the network and instead focuses on its *utility* to implement these properties during learning, e.g., by means of regularization.

Our **contributions** are as follows: First, we formalize the intuition that in a neural network regime (where training samples are fitted perfectly), generalization is linked to probability mass concentration around the training samples in the internal representation space $\mathcal{Z}$. Second, we prove that a *topological constraint* on samples from the aforementioned

[1]Department of Computer Science, Univ. of Salzburg, Austria [2]Univ. of North Carolina, Chapel Hill, USA. Correspondence to: Christoph D. Hofer <chr.dav.hofer@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

push-forward measure (restricted to each class), leads to mass concentration. Third, relying on work by Hofer et al. (2019), we devise a regularizer to encourage the derived topological constraint during optimization and present empirical evidence across various vision benchmark datasets to support our claims.

1.1. Related work

Prior works, most related to ours, focus on regularizing statistics of internal representations in supervised learning settings. As opposed to *explicit* regularization which aims at restricting model capacity, e.g., by penalizing norms on network weights, regularizing representation statistics can be considered a less direct, data dependent, mechanism.

The intended objective of regularizing representation statistics vary across the literature. Glorot et al. (2011), for instance, encourage sparsity via an l_1 norm penalty. Cogswell et al. (2016) aim for redundancy minimization by penalizing the covariance matrix of representations. Choi & Rhee (2019) later extended this idea to perform class-wise regularization, i.e., a concept similar to (Behlarbi et al., 2017) where pairwise distances between representations of a class are minimized. In an effort to potentially replace batch normalization, Littwin & Wolf (2018) propose to control variations, across mini-batches, of each neuron's variance (before activation). This is shown to be beneficial in terms of reducing the number of modes in the output distribution of neurons. Liao et al. (2016) follow a different strategy and instead cluster representations to achieve parsimony, but at the cost of having to set the number of clusters (which can be difficult in small-sample size regimes). Motivated by the relation between the generalization error and the natural gradient (Roux et al., 2017), Joo et al. (2020) recently proposed to match the distribution of representations to a standard Gaussian, via the sliced Wasserstein distance. Yet, to the best of our knowledge, all mentioned regularization approaches in the realm of controlling internal representations, either only empirically demonstrate better generalization, or show a loose connection to the latter. *In contrast, we establish a direct (provable) connection between a property encouraged by our regularizer and the associated beneficial effects on generalization.*

Our technical contribution resides on the intersection of machine learning and algebraic topology (persistent homology in particular). Driven by various intents, several works have recently adopted concepts from algebraic topology. Rieck et al. (2019), for instance, study topological aspects of neural networks, represented as graphs, to guide architecture selection, Guss & Salakhutdinov (2018) aim to quantify dataset complexity. On the more theoretical side, Bianchini & Scarselli (2014) study the functional complexity of neural networks. Notably, these works *passively use* ideas from topology for post-hoc analysis of neural networks. More recently, various works have presented progress along the lines of *actively* controlling topological aspects. Chen et al. (2019) regularize decision boundaries of classifiers, Hofer et al. (2019) optimize connectivity of internal representations of autoencoders, and Rieck et al. (2019) match topological characteristics of input data to the topological characteristics of representations learned by an autoencoder. Technically, we rely on these advances to eventually implement a regularizer, *yet our primary objective is to study the connection between generalization and the topological properties of the probability measure induced by a neural network's feature extractor φ.*

1.2. Notation & Learning setup

In the context of Eq. (1), we refer to $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ as the sample, label and internal representation space. We assume that $\mathcal{Z}$ is equipped with a metric $\mathbb{d}$ and $\mathcal{Y} = [K] = \{1, \dots, K\}$. By P, we denote a probability measure on $\mathcal{X}$ and by Q the push-forward measure, induced by the measurable function $\varphi : \mathcal{X} \to \mathcal{Z}$, on the Borel σ-algebra Σ defined by $\mathbb{d}$ on $\mathcal{Z}$; Q^b denotes the product measure of Q. $B(z, r)$ refers to the *closed* (metric) ball of radius $r > 0$ around $z \in \mathcal{Z}$.

Our learning setup is to assume a deterministic relationship[1] between $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. This relationship is determined by $c : \operatorname{supp}(P) \to \mathcal{Y}$, where $\operatorname{supp}(P)$ refers to the support of P. We assume a training sample S, consisting of pairs $(x_1, y_1), \dots, (x_m, y_m)$ is the result of m i.i.d. draws of $X \sim P$, labeled via $y_i = c(x_i)$. By $S_{x|k}$ we denote the data instances, x_i, of class k. For a classifier $h : \mathcal{X} \to \mathcal{Y}$ and $X \sim P$, we define the *generalization error* as

$$\mathop{\mathbb{E}}_{X \sim P}[\mathbb{1}_{h,c}(X)] \ ,$$

where

$$\mathbb{1}_{h,c}(x) = \begin{cases} 0, & h(x) = c(x), \\ 1, & \text{else} \ . \end{cases}$$

For brevity, proofs are deferred to the suppl. material.

2. Topologically densified distributions

To build up intuition, consider $X \sim P$ and $\varphi(X) = Z$. As $Z \sim Q$ and the linear classifier γ operates on the internal representations yielded by φ, we can ask two questions: *(I) Which properties of Q are beneficial for generalization, and (II) how can we impose these properties?*

Increasing the probability that φ maps a sample of class k into the correct decision region (induced by γ) improves generalization. In Lemma 1 we will link this fact to a condition depending on Q.

[1] i.e., an arguably realistic setup (Kolchinsky et al., 2019) for many practical problems.

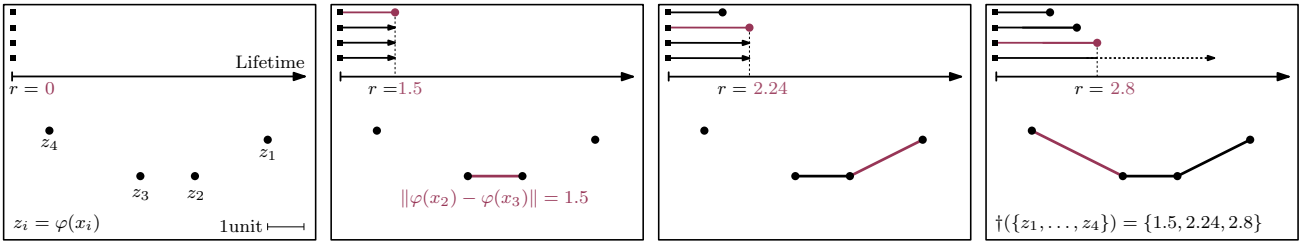


Figure 2. Illustration of *0-dimensional Vietoris-Rips persistent homology*. Starting from a point set $M = \{z_1, \ldots, z_4\}$, we iteratively construct a graph by sweeping $r \in [0, \infty)$. While increasing r, we add an edge between $z_i \neq z_j$ if $\|z_i - z_j\|_2 = r$ and track the *merging* of connected components as we successively add edges. The times (aka *death-times*) of those *merging events* are stored in the (multi-)set $\dagger(M)$, typically called a *persistence barcode* (as all connected components appear at $r = 0$, we omit the *birth* times).

At first, we introduce a way to measure *class-specific probability mass*. To this end, we define the restriction of Q (i.e., the push-forward of P via φ) to class k as

$$Q_k : \Sigma \to [0, 1] \quad \Sigma \ni \sigma \mapsto \frac{Q(\sigma \cap C_k)}{Q(C_k)} \ , \tag{2}$$

where $C_k = \varphi(c^{-1}(\{k\}))$ is the representation of class k in $\mathcal{Z}$. In the optimal case, the probability mass of the k-th decision region, measured via Q_k, tends towards one. The following lemma formalizes this notion by establishing a direct link between Q_k and the generalization error.

Lemma 1. *For any class $k \in [K]$, let $C_k = \varphi\left(c^{-1}(\{k\})\right)$ be its internal representation and $D_k = \gamma^{-1}(\{k\})$ be its decision region in $\mathcal{Z}$ w.r.t. γ. If, for $\varepsilon > 0$,*

$$\forall k : 1 - Q_k(D_k) \leq \varepsilon \ , \tag{3}$$

then

$$\mathop{\mathbb{E}}_{X \sim P}[\mathbb{1}_{\gamma \circ \varphi, c}(X)] \leq K\varepsilon \ .$$

While Lemma 1 partially answers *Question (I)* by yielding a property beneficial for generalization, it remains to find a mechanism to impose it.

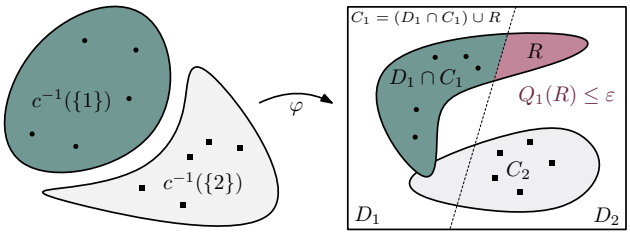


Figure 3. Eq. (3) of Lemma 1 controls how much probability mass of C_k is concentrated in D_k w.r.t. Q_k (only illustrated for $k = 1$ here). The smaller ε gets, the less mass is present in R, i.e., the region where errors on unseen data (of class $k = 1$) occur.

Inspired by recent work of Zhang et al. (2017), we continue by assuming $\gamma \circ \varphi$ to be powerful enough to fit any given training set S of size m. Specifically, we assume that for $z_i = \varphi(x_i)$ with $c(x_i) = k$, there exists $r > 0$ with $B(z_i, r) \subset D_k$. This is equivalent to a margin assumption on the training instances in $\mathcal{Z}$. With this in mind, increasing $Q_k(B(z_i, r))$ is beneficial for generalization, as it can only increase $Q_k(D_k)$. Our strategy hinges on this idea.

2.1. Topological densification

We show that a certain topological constraint on Q_k will *provably* lead to *probability mass concentration*. More precisely, given a reference set $M \subseteq \mathcal{Z}$ and its ε-extension

$$M_\varepsilon = \bigcup_{z \in M} B(z, \varepsilon), \quad \varepsilon > 0 \ , \tag{4}$$

the topological constraint provides a non-trivial lower bound on $Q_k(M_\varepsilon)$ in terms of $Q_k(M)$. Informally, we say that Q_k is *topologically densified* around M.

Our main arguments rely on the probability of an i.i.d. draw from Q_k^b (i.e., the product measure) to be *connected*. The latter is a topological property which can be computed using tools from algebraic topology. In particular, we quantify connectivity via 0*-dimensional (Vietoris Rips) persistent homology*, visually illustrated in Fig. 2. For a thorough technical introduction, we refer the reader to, e.g., (Edelsbrunner & Harer, 2010) or (Boissonnat et al., 2018).

Definition 1. Let $\beta > 0$. A set $M \subseteq \mathcal{Z}$ is *β-connected* iff all 0-dimensional death-times of its Vietoris-Rips persistent homology are in the *open* interval $(0, \beta)$.

As all information captured by 0-dimensional Vietoris-Rips persistent homology (Robins, 2000) is encoded in the minimum spanning tree (MST) on M (w.r.t. metric d), we can equivalently formulate Definition 1 in terms of the edges in the MST. In particular, we can say that each edge in the MST of M has edge length less than β. However, the topological perspective is preferable, as we can rely on previous work (Hofer et al., 2019) which shows how to backpropagate gradients through the persistent homology computation. This property is needed to implement a regularizer (see §2.4).

To capture β-connectedness of a sample, we define the indicator function $\mathfrak{c}_b^\beta : \mathcal{Z}^b \to \{0,1\}$ as

$$\mathfrak{c}_b^\beta(z_1,\ldots,z_b) = 1 \Leftrightarrow \{z_1,\ldots,z_b\} \text{ is } \beta\text{-connected} \ .$$

We now consider the probability of b-sized i.i.d. draws from Q_k, see Eq. (2), to be β-connected.

Definition 2. Let $\beta > 0, c_\beta \in [0,1]$, and $b \in \mathbb{N}$. We call Q_k (b, c_β)-*connected* if

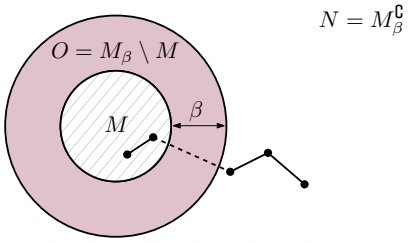
$$Q_k^b(\{\mathfrak{c}_b^\beta = 1\}) \geq c_\beta \ .$$

To underpin the relevance of this (probabilistic) connectivity property w.r.t. probability mass concentration, we sketch the key insights that lead to the main results of the paper.

Connectedness yields mass concentration. For sake of argument, assume Q_k to be (b, c_β)-connected. Now, consider a reference set $M \subseteq \mathcal{Z}$ together with two sets

$$N = M_\beta^\complement \quad \text{and} \quad O = M_\beta \setminus M \ ,$$

where $M_\beta^\complement$ denotes the set complement. These three sets are illustrated below on a toy example.



Apparently, M, N and O partition $\mathcal{Z}$. For an i.i.d. sample, $z_1,\ldots,z_b$, from Q_k, we can study how distributing the z_i among M, N, and O impacts β-connectedness. In particular, let $\sharp_M, \sharp_N, \sharp_O$ be the number of z_i's which fall within M, N, and O respectively, i.e.,

$$\sharp_M = |\{z_i\} \cap M|, \ \sharp_N = |\{z_i\} \cap N|, \ \sharp_O = |\{z_i\} \cap O| \ .$$

Observation 1: If the membership assignment is such that

$$\sharp_M \geq 1 \text{ and } \sharp_N \geq 1, \text{ but } \sharp_O = 0 \ ,$$

then $z_1,\ldots,z_b$ cannot be β-connected. This is easy to see, as for $z_M \in M$ and $z_N \in N$, we get $\mathrm{d}(z_M, z_N) \geq \beta$ and there are simply no z_i's in O (see illustration above).

Observation 2: The probability of

$$(\sharp_M, \sharp_N, \sharp_O) \in \{0,\ldots,b\}^3$$

is given by a Trinomial distribution with parameters $Q_k(M)$, $Q_k(O)$ and $Q_k(N) = 1 - Q_k(M) - Q_k(O)$.

As it holds that $Q_k(O) = Q_k(M_\beta) - Q_k(M)$ and $Q_k(N) = 1 - Q_k(M_\beta)$, the probability of a b-sample *not* being β-connected can be expressed in terms of

$$\mathfrak{p} = Q_k(M) \quad \text{and} \quad \mathfrak{q} = Q_k(M_\beta) \ .$$

The key aspect of this construction is that we describe *events* where $z_1,\ldots,z_b$ cannot be β-connected, i.e.,

$$E = \{(z_i) \in \mathcal{Z}^b : \sharp_M \geq 1, \sharp_N \geq 1, \sharp_O = 0\}$$

and $\mathfrak{c}_b^\beta(E) = \{0\}$. As we will see, based on the Trinomial distribution, one can derive a polynomial Ψ expressing the probability of E, i.e.,

$$Q_k(E) = \Psi(\mathfrak{p}, \mathfrak{q}) \ .$$

Consequently, as c_β is defined to be the probability of a b-sample to be β-connected and E describes events where b-samples are *not* β-connected, it holds that

$$1 - c_\beta \geq Q_k(E) = \Psi(\mathfrak{p}, \mathfrak{q}) \ .$$

We will see, by properties of Ψ, that this relationship allows us to lower bound $\mathfrak{q} = Q_k(M_\beta)$, if $\mathfrak{p} = Q_k(M)$ is known. In other words, if M covers a certain mass, then we can infer the minimal mass which has to be covered by M_β. Our main result – presented in Theorem 1 – is slightly more general, as it not only considers the β-extension of M, but the $l \cdot \beta$-extension for $l \in \mathbb{N}$. In this more general case, the polynomial Ψ takes the form as in Definition 3 below.

Definition 3. Let $b, l \in \mathbb{N}$ and $p, q \in [0,1]$. For $p \leq q$, we define the polynomial

$$\Psi(p, q\,;\, b, l) = \sum_{\substack{(u,v,w) \\ \in I(b,l)}} \frac{b!}{u!\, v!\, w!} p^u (1-q)^v (q-p)^w \ ,$$

where the index set $I(b,l)$ is given by

$$I(b,l) = \{\, (u,v,w) \in \mathbb{N}_0^3 : \\ u + v + w = b \ \wedge \ u, v \geq 1 \ \wedge \ w \leq l - 1 \,\} \ .$$

The most important properties of Ψ are: (1) Ψ is *monotonically increasing* in p (and l); (2) Ψ is *monotonically decreasing* in q and (3) Ψ vanishes for $q = 1$.

Theorem 1. *Let $b, l \in \mathbb{N}$ and let Q_k be (b, c_β)-connected. Then, for all reference sets $M \in \Sigma$ and*

$$\mathfrak{p} = Q_k(M), \ \mathfrak{q} = Q_k(M_{l\cdot\beta})$$

it holds that

$$1 - c_\beta \geq \Psi\big(\mathfrak{p}, \mathfrak{q}\,;\, b, l\big) \ . \tag{5}$$

2.2. Ramifications of Theorem 1

By properties (1) – (3) of Ψ, Theorem 1 allows us to lower-bound the mass increase caused by extending (see Eq. (4)) a reference set M by $l \cdot \beta$. Recall that this is beneficial for generalization, if M is constructed from representations (in $\mathcal{Z}$) of correctly classified training instances.

In detail, assume that the mass of the reference set M, $\mathfrak{p} = Q_k(M)$, is fixed and let $\mathfrak{q} = Q_k(M_{l\cdot\beta})$ be the mass of the $l \cdot \beta$ extension. Then, by Theorem 1,

$$\mathfrak{q} \in \Big\{ q \in [\mathfrak{p}, 1] : 1 - c_\beta \geq \Psi(\mathfrak{p}, q\,;\, b, l) \Big\} = A \ , \qquad (6)$$

and thus A is non-empty. Now let $\mathcal{R}_{b,c_\beta}(\mathfrak{p}, l) = \min A$ identify the *smallest mass* in the $l \cdot \beta$-extension for which the inequality in Eq. (5) holds. As Ψ is monotonically decreasing, $\mathcal{R}_{b,c_\beta}(\mathfrak{p}, l)$ is monotonically increasing in c_β. This will motivate our regularization goal of *increasing* c_β.

Sufficient condition for mass concentration. Note that $\mathfrak{q} \geq \mathcal{R}_{b,c_\beta}(\mathfrak{p}, l) \geq \mathfrak{p}$ and thus, mass concentration is guaranteed as long as $\mathcal{R}_{b,c_\beta}(\mathfrak{p}, l) > \mathfrak{p}$. Otherwise, the mass in the $l \cdot \beta$-extension of M may not be greater than the mass in M. In fact, $\mathcal{R}_{b,c_\beta}(\mathfrak{p}, l) > \mathfrak{p}$ only holds if

$$1 - c_\beta < \Psi(\mathfrak{p}, \mathfrak{p}\,;\, b, l) = 1 - \mathfrak{p}^b - (1 - \mathfrak{p})^b \ . \qquad (7)$$

The behavior of Eq. (7) is specifically relevant in the region where $\mathfrak{p}$ is close to 0, as requiring a large mass in M would be detrimental. Notably, as we can see in Fig. 4, which shows Eq. (7) as a function of $\mathfrak{p} = Q_k(M)$, already small values of $\mathfrak{p}$ reach the *critical* threshold of $1 - c_\beta$.

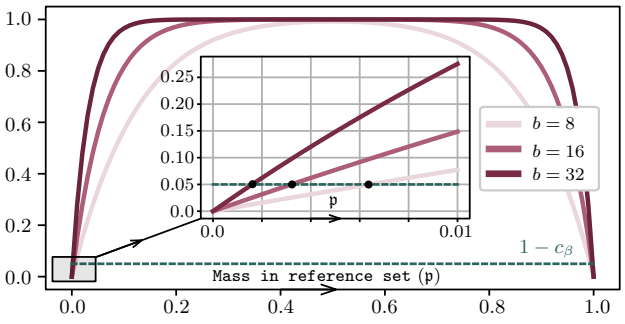


Figure 4. Illustration of when $1 - \mathfrak{p}^b - (1 - \mathfrak{p})^b > 1 - c_\beta$ holds, i.e., when mass concentration effects start to occur. The zoomed-in view shows the relevant region near 0.

Quantification of mass concentration. To understand how the minimal mass in $M_{l\cdot\beta}$ is boosted by the mass of the reference set M, we visualize (in Fig. 5) the minimial values for $\mathfrak{q} = M_{l\cdot\beta}$ as a function of $\mathfrak{p} = Q_k(M)$, i.e., the mass of the reference set M. Similar to Fig. 4, as $\mathfrak{p}$ approaches 0 (or 1), the mass concentration effect is rendered negligible.

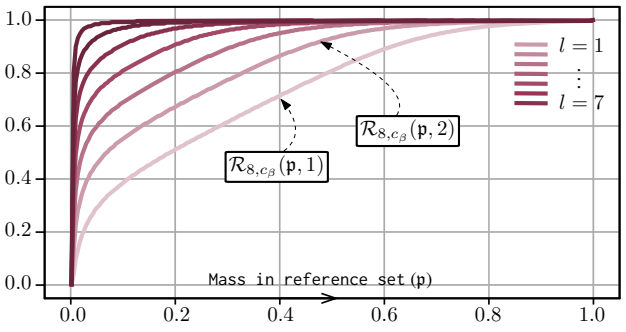


Figure 5. Illustration of $\mathcal{R}_{b,c_\beta}(\mathfrak{p}, l)$, i.e., the lower bound on $\mathfrak{q} = Q_k(M_{l\cdot\beta})$, plotted as a function of the mass $\mathfrak{p} = Q_k(M)$ of the reference set M (for $b = 8$ and different l).

However, already a small mass in M is sufficient for strong mass concentration in $M_{l\cdot\beta}$.

Next, we discuss the role of c_β, i.e., the probability of a b-sized sample from Q_k to be β-connected. Fig. 6 illustrates, for different choices of l, where $1 - c_\beta \geq \Psi(\mathfrak{p}, \mathfrak{q}\,;\, b, l)$ holds, as a function of $\mathfrak{q}$ with $\mathfrak{p} = 0.1$ fixed. Most importantly, as c_β is increased, the minimal mass in a particular $l \cdot \beta$ extension of M, characterized by $\mathcal{R}_{b,c_\beta}(\mathfrak{p}, l)$, shifts towards larger values.

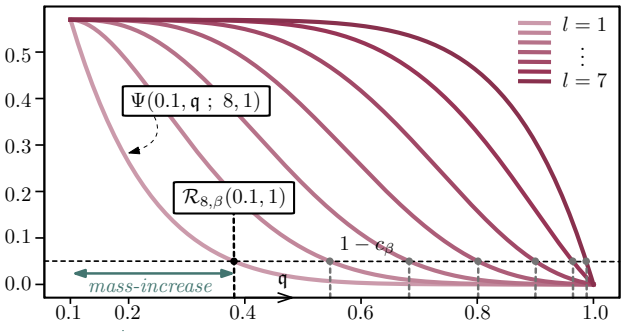


Figure 6. Illustration of Ψ for $\mathfrak{p} = 0.1, b = 8$ and different $l \in \mathbb{N}$. Points at which $1 - c_\beta = \Psi(\mathfrak{p}, q\,;\, b, l)$ holds are marked by dots.

Overall, Theorem 1 and the analysis presented above provide a possible answer to *Question (II)* stated at the beginning of §2. In particular, a mechanism to increase $Q_k(D_k)$ in Eq. (3) is to encourage Q_k to be (b, c_β)-connected. We conclude with the following summary:

If a measure is (b, c_β)-connected, then mass attracts mass; the higher c_β, the stronger the effect.

2.3. Limitations

As the mass, $\mathfrak{p}$, of the reference set M is typically unknown, all results are *relative* (w.r.t. $\mathfrak{p}$), not absolute. This warrants a discussion of potential limitations in a learning context.

Specifically, when learning from samples, an arguably natural choice for a class-specific reference set is to consider the union of balls around the representations of the training samples, yielding (for $r > 0$)

$$M^{(k)} = \bigcup_{z \in \varphi(S_{x|k})} B(z, r) \ . \tag{8}$$

Now, lets assume that Q_k is (b, c_β)-connected and the linear classifier γ attains zero error on $\varphi(S_{x|k})$. Two issues suppress good generalization:

First, the derived mass concentration is only beneficial if, for a given $l \in \mathbb{N}$, the reference set $M^{(k)}$ is located sufficiently far away from the decision boundary of class k, i.e.,

$$M^{(k)}_{l \cdot \beta} \subset D_k = \gamma^{-1}\big(\{k\}\big) \ . \tag{9}$$

In practice, we can, to some extent, induce such a configuration by selecting a loss function which yields a large margin in $\mathcal{Z}$. In that case, at least a $\mathcal{R}_{b,c_\beta}(Q_k(M^{(k)}), l)$ proportion of class k is correctly classified by γ. A violation of Eq. (9) would mean that mass is still concentrated, but the $l \cdot \beta$-extension might reach across the k-th decision region.

Second, the sample $S_{x|k}$ has to be *good* in the sense that $\mathfrak{p} = Q_k(M^{(k)})$ is sufficiently large (as noted earlier, see Fig. 4). This is somewhat related to the notion of *representativeness* of a training set, i.e., a topic well studied in works on learnability.

Overall, given that Q_k is (b, c_β)-connected, mass concentration effects provably occur; yet, advantages only come to light under the conditions outlined above. It is thus worth designing a *regularization* strategy to encourage (b, c_β)-connectedness during optimization. We describe such a strategy next.

2.4. Regularization

To encourage (b, c_β)-connectedness of Q_k, it is obvious that we have to consider multiple training instances of each class *jointly*. To be practical, we integrate this requirement into the prevalent setting of learning with mini-batches.

Our integration strategy is simple and, in fact, similar approaches (in a different context) have been investigated in prior work (see Hoffer et al., 2019). In detail, we construct each mini-batch, $\mathtt{B}$, as a collection of $\mathtt{n}$ sub-batches, i.e., $\mathtt{B} = (\mathtt{B}_1, \ldots, \mathtt{B}_\mathtt{n})$. Each sub-batch consists of b samples from the *same* class, thus the resulting mini-batch $\mathtt{B}$ is built from $\mathtt{n} \cdot b$ samples. Our regularizer is formulated as a loss term that penalizes deviations from a β-connected arrangement of the z_i in each sub-batch $\mathtt{B}_j$. To realize this, we leverage a recent approach from Hofer et al. (2019) which introduces a *differentiable* penalty on lifetimes of connected components (w.r.t. Vietoris-Rips persistent homology).

Formally, let $\dagger(\mathtt{B}_i)$ contain the death-times (see Fig. 2) computed for sub-batch $\mathtt{B}_i$. Then, given the hyper-parameter $\beta > 0$, we set the *connectivity penalty* for mini-batch $\mathtt{B}$ as

$$\mathcal{L}(\mathtt{B}) = \sum_{i=1}^{\mathtt{n}} \sum_{d \in \dagger(\mathtt{B}_i)} |d - \beta| \ . \tag{10}$$

Notably, this is the same term as in (Hofer et al., 2019), however, motivated by a different objective.

Admittedly, to encourage (b, c_β)-connectivity of Q_k, it would suffice to use a less restrictive variant and only penalize lifetimes *greater* than β. However, we have empirically observed that Eq. (10) is more effective. This would imply that it is beneficial to prevent lifetimes from collapsing and, as a result, prevent Q_k to become *overly* dense. Currently, we can not formally explain this effect, but hypothesize that – to some extent – preventing lifetimes from collapsing preserves variance in the gradients, i.e., a property useful during SGD's *drift* phase (Shwartz-Ziv & Tishby, 2017).

3. Experiments

For our experiments[2], we draw on a setup common to many works in semi-supervised learning (Laine & Aila, 2017; Oliver et al., 2018; Verma et al., 2019b), both in terms of dataset selection and network architecture. As small sample-size experiments are typically presented as *baselines* in these works, we believe this to be an arguable choice. In particular, we present experiments on three (10 class) vision benchmark datasets: MNIST, SVHN and CIFAR10. For MNIST and SVHN, we limit training data to 250 instances, on CIFAR10 to 500 (and 1,000), respectively.

Architecture & Optimization. For CIFAR10 and SVHN we use the CNN-13 architecture of (Laine & Aila, 2017) which already includes dropout regularization (Srivastava et al., 2014). Only on MNIST we rely on a simpler CNN architecture with four convolutional blocks and max-pooling (w/o dropout). Both architectures have a final linear classifier $\gamma : \mathbb{R}^{128} \to \mathbb{R}^K$, use batch normalization (Ioffe & Szegedy, 2015), and fit our network decomposition of Eq. (1). Optimization is done by SGD with momentum (0.9) over 310 epochs with cross-entropy loss and cosine learning rate annealing (Loshchilov & Hutter, 2017) (without restarts). As all experiments use weight decay, it is important to note that batch normalization *combined* with weight decay only exerts regularization on the classifier γ. In fact, several works have shown that the combination of batch normalization and weight decay mainly affects the effective learning rate (van Laarhoven, 2017; Zhang et al., 2019).

The weighting of our regularization term is set such that

[2] `PyTorch` source code is available at `https://github.com/c-hofer/topologically_densified_distributions`

the range of the loss from Eq. (10) is comparable, in range, to the cross-entropy loss. We choose a sub-batch size of $b = 16$ and draw $\mathrm{n} = 8$ sub-batches (see §2.4); this amounts to a total batch size of 128. While, empirically, this setting facilitates stable optimization (w.r.t. batch norm statistics), we acknowledge that further theoretical insights could lead to a more informed choice. Additional parameter details are provided when relevant (and full details can be found in the supplementary material).

First, in §3.1, we investigate to which extent the (b, c_β)-connectivity property (imposed during optimization), translates to unseen data. In §3.2, we study the effect of β and whether this parameter can be reliably cross-validated on a *small* validation set. Finally, in §3.3, we compare to related work on regularizing statistics of internal representations.

3.1. Evaluating (b, c_β)-connectivity

For a fixed b, we study how well β-connectivity is achieved during optimization, as we vary β. In accordance with Definition 1, we measure β-connectivity via the lifetimes in the 0-dimensional persistence barcodes, computed over 500 random sub-batches (chosen from training/testing data).

Qualitatively, in Fig. 7 we see that increasing β not only translates to an increase of lifetimes in the internal representations of training instances, but equally translates to an increase in lifetimes on sub-batches of the testing data. While we observe a slight offset in the lifetime average (training *vs.* testing), and an increase in variance, these effects are largely constant across β. This suggests the *effect* of regularization is qualitatively invariant to the choice of β.

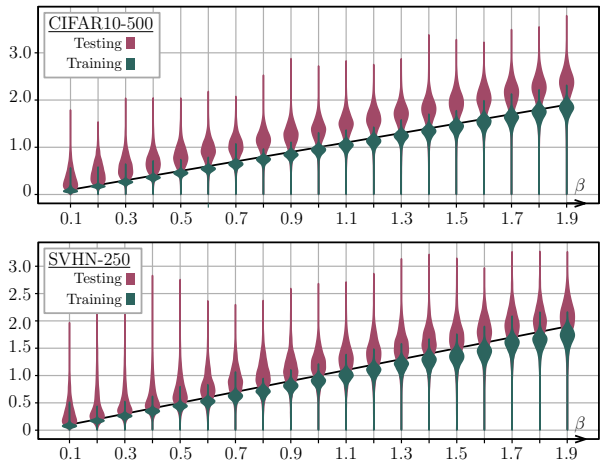


Figure 7. Lifetime distribution computed over 500 random sub-batches (of size 16) from (CIFAR10/SVHN) training and testing data, as a function of $\beta \in [0.1, 1.9]$ (set during optimization).

3.2. Selection of β

Using Eq. (10) as a loss term requires to set β a-priori. However, this choice can be crucial, as it assigns a notion of *scale* to $\mathcal{Z}$ and thus interacts with the linear classifier. In particular, β is interweaved with the Lipschitz constant of γ which is affected by weight decay.

While, at the moment, we do not have sufficient theoretical insights into the interaction between weight decay on γ and the choice of β, we argue that β can still be cross-validated (a common practice for most hyper-parameters). Yet, in small sample-size regimes, having a large labeled validation set is unrealistic. Thus, we study the behavior of cross-validating β, when the validation set is of size equal to the training corpus. To this end, Fig. 8 shows the testing error on CIFAR10 (using 500 training samples) and SVHN (using 250 training samples), over a range of β. Additionally, we overlay the variation in the error on the held-out validation sets. As we can see, the latter closely tracks the testing error as β is increased from 0.1 to 1.9. This indicates that choosing β through cross-validation can be done effectively. Fig. 8 additionally reveals that the testing error behaves smoothly around the optimal choice of β.

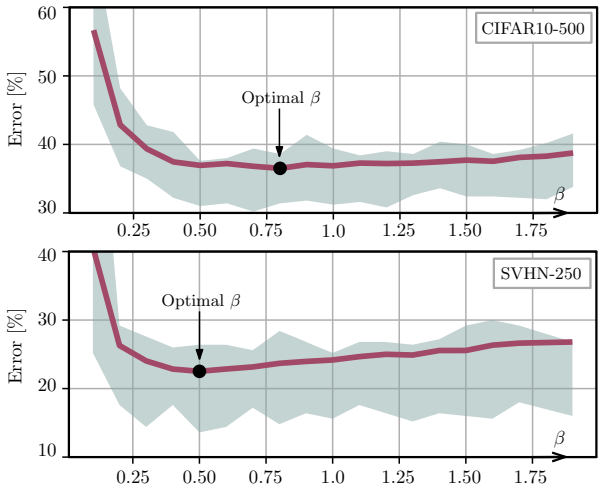


Figure 8. Testing error (purple; averaged over 10 runs) over various choices of $\beta \in [0.1, 1.9]$. The shaded region shows the variation in the testing error on small-validation sets. This indicates that the choice of β can be cross-validated effectively.

3.3. Comparison to the state-of-the-art

Finally, we present a comparison to different state-of-the-art regularizers. Specifically, we evaluate against works that regularize statistics of internal representations (right before the linear classifier). This includes the *DeCov* loss of Cogswell et al. (2016), as well as its class-wise extensions (*cw-CR* and *cw-VR*), proposed in Choi & Rhee (2019). As a representative of an alternative approach, we provide re-

Regularization	MNIST-250	SVHN-250	CIFAR10-500	CIFAR10-1k
Vanilla	7.1 ± 1.0	30.1 ± 2.9	39.4 ± 1.5	29.5 ± 0.8
+ Jac.-Reg. (Hoffman et al., 2019)	6.2 ± 0.8	33.1 ± 2.8	39.7 ± 2.0	29.8 ± 1.2
+ DeCov (Cogswell et al., 2016)	6.5 ± 1.1	28.9 ± 2.2	38.2 ± 1.5	29.0 ± 0.6
+ VR (Choi & Rhee, 2019)	6.1 ± 0.5	28.2 ± 2.4	38.6 ± 1.4	29.3 ± 0.7
+ cw-CR (Choi & Rhee, 2019)	7.0 ± 0.6	28.8 ± 2.9	39.0 ± 1.9	29.1 ± 0.7
+ cw-VR (Choi & Rhee, 2019)	6.2 ± 0.8	28.4 ± 2.5	38.5 ± 1.6	29.0 ± 0.7
+ Sub-batches	7.1 ± 0.5	27.5 ± 2.6	38.3 ± 3.0	28.9 ± 0.4
+ Sub-batches + Top.-Reg. (**Ours**)	$\mathbf{5.6 \pm 0.7}$	$\mathbf{22.5 \pm 2.0}$	$\mathbf{36.5 \pm 1.2}$	$\mathbf{28.5 \pm 0.6}$
+ Sub-batches + Top.-Reg. (**Ours**) ‡	$\mathbf{5.9 \pm 0.3}$	$\mathbf{23.3 \pm 1.1}$	$\mathbf{36.8 \pm 0.3}$	$\mathbf{28.8 \pm 0.3}$

Table 1. Comparison to state-of-the-art regularizers added to *Vanilla* training which includes batch normalization, dropout (0.5; except for MNIST) and weight decay. Reported is the lowest achievable test error [%] ($\pm$ std. deviation) over a hyper-parameter grid, averaged over 10 cross-validation runs. Numbers attached to the dataset names indicates the number of training instances used. The last row (‡) lists the results of our approach when β is cross-validated (and all other hyper-parameters are fixed) as described in §3.3.

sults when penalizing the network Jacobian, as proposed in (Sokolić et al., 2017; Hoffman et al., 2019). For these comparison experiments, we empirically found a batch size of 32 to produce the best results. To account for the difference in the update steps of SGD w.r.t. to our approach (caused by the sub-batch construction), we adjusted the number of epochs accordingly. All approaches are evaluated on the same training/testing splits and achieve zero training error.

To establish a *strong baseline*, we decided to conduct an extensive hyper-parameter search over a grid of (1) initial learning rate, (2) weight decay and (3) weighting of the regularization terms. For each grid point, we run 10 cross-validation runs, average, and then pick the *lowest achievable error* on the test set. This establishes an *lower bound* on the error if hyper-parameters were chosen via a validation set. Table 1 lists the corresponding results.

To test our regularizer against these lower bounds, we fix all hyper-parameters and cross-validate β, as discussed in §3.2. Notably, topological regularization *consistently* exhibits the lowest error, even when compared to the optimistic performance estimate of the other regularizers. This strongly supports our claim that mass concentration is beneficial.

4. Discussion

As emphasized earlier, our theoretical results are *relative* in nature, in particular, relative to a reference set M, naturally determined by the representations of the training samples.

In §2, we linked mass concentration to generalization and showed that mass in the β-extension of M increases, as the probability c_β, i.e., the probability of a b-sized sample from Q_k to be β-connected, increases. Results in Table 1 empirically support this. *However, can mass concentration be directly observed?* While it is challenging to measure this, we can perform a proxy experiment. In detail, we select two models trained with equal β and define the reference sets (per class) via balls of radius $r > 0$, see Eq. (8), around 500 randomly selected training samples. By successively increasing r and counting *test samples* that occur in $B(z_i, r)$ and $B(z_i, r + \beta)$, resp., we obtain estimates of $\mathfrak{p} = Q_k(M)$ and $\mathfrak{q} = Q_k(M_\beta)$. As β-connectivity is not *strictly* enforced, but used for regularization, we have to account for the lifetime shift seen in Fig. 9. Hence, to estimate $\mathfrak{p}$ and $\mathfrak{q}$, we use $\beta = 1.4$, which is higher than the sought-for $\beta = 0.8$ during training. Fig. 9 (left) shows the mass estimates for two CIFAR10 models, trained on 500 and 1,000 samples, respectively.

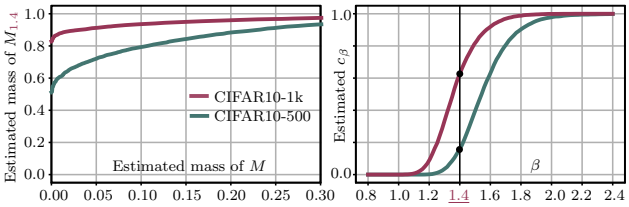


Figure 9. (*Left*) Estimated mass in the reference set M *vs.* the estimated mass in M_β, shown for $\beta = 1.4$; (*Right*) Estimated probability of a b-sized sample to be β-connected. All estimations are computed on testing data.

As we can see – especially for small (estimated) mass in the reference set – the mass concentration effect is much stronger for the model trained with 1,000 samples. The underlying reason is that using more samples improves how (b, c_β)-connectivity transfers to testing data. Estimates for c_β across both models confirm the latter, see Fig. 9 (right). This strongly indicates that mass concentration is not only a theoretical result, but is indeed observed on real data.

Overall, the presented analysis suggests that studying and *controlling* topological properties of representations is promising. Yet, we have only started to scratch at the surface

of how topological aspects influence generalization. We argue that further (formal) understanding of these connections could offer novel insights into the generalization puzzle.

Acknowledgements

This work was partially funded by the Austrian Science Fund (FWF): project FWF P31799-N38 and the Land Salzburg (WISS 2025) under project numbers 20102-F1901166-KZP and 20204-WISS/225/197-2019.

References

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. Stronger generalization bounds for deep nets via a compression approach. In *ICML*, 2018.

Bartlett, P., Foster, D., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. In *NIPS*, 2017.

Behlarbi, S., Chatelain, C., Herault, R., and Adam, S. Neural networks regularization through class-wise invariant representation learning. *arXiv*, 2017. https://arxiv.org/abs/1709.01867.

Bianchini, M. and Scarselli, F. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Trans. Neural Netw. Learn. Syst.*, 25(8):1533–1565, 2014.

Boissonnat, J.-D., Chazal, F., and Yvinec, M. *Geometric and Topological Inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2018.

Chen, C., Ni, X., Bai, Q., and Wang, Y. A topological regularizer for classifiers via persistent homology. In *AISTATS*, 2019.

Choi, D. and Rhee, W. Utilizing class information for deep network representation shaping. In *AAAI*, 2019.

Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., and Batra, D. Reducing overfitting in deep networks by decorrelating representations. In *ICLR*, 2016.

Cubuk, E., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019.

Dao, T., Gu, A., Rather, A., Smith, V., Sa, C. D., and Re, C. A kernel theory of modern data augmentation. In *ICML*, 2019.

Edelsbrunner, H. and Harer, J. L. *Computational Topology : An Introduction*. American Mathematical Society, 2010.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *AISTATS*, 2011.

Golowich, N., Rakhlin, A., and Shamir, O. Size-independent sample complexity of neural networks. In *COLT*, 2018.

Guss, W. and Salakhutdinov, R. On characterizing the capacity of neural networks using algebraic topology. *arXiv*, 2018. https://arxiv.org/abs/1802.04443.

Hofer, C., Kwitt, R., , Dixit, M., and Niethammer, M. Connectivity-optimized representation learning via persistent homology. In *ICML*, 2019.

Hoffer, E., Ben-Nun, T., Hubara, I., Giladi, N., Hoefler, T., and Soudry, D. Augment your batch: better training with larger batches. *arXiv*, 2019. https://arxiv.org/abs/1901.09335.

Hoffman, J., Roberts, D., and Yaida, S. Robust learning with jacobian regularization. *arXiv*, 2019. https://arxiv.org/abs/1908.02729.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Joo, T., Kang, D., and Kim, B. Regularizing activations in neural networks via distribution matching with the Wasserstein metric. In *ICLR*, 2020.

Kolchinsky, A., Tracey, B., and Kuyk, S. V. Caveats for information bottleneck in deterministic scenarios. In *ICLR*, 2019.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. In *ICLR*, 2017.

Liao, R., Schwing, A., Zemel, R., and Urtasun, R. Learning deep parsimonious representations. In *NIPS*, 2016.

Littwin, E. and Wolf, L. Regularizing by the variance of the activations' sample-variances. In *NIPS*, 2018.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Nagarajan, V. and Kolter, J. Uniform convergence may be unable to explain generalization in deep learning. In *NeurIPS*, 2019.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *NIPS*, 2017.

Oliver, A., Odena, A., Raffel, C., Cubuk, E., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In *NIPS*, 2018.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *ICLR*, 2019.

Robins, V. *Computational topology at multiple resolutions: foundations and applications to fractals and dynamics*. PhD thesis, University of Colorado, 6 2000.

Roux, N., Manzagol, P.-A., and Bengio, Y. Topmoumoute online natural gradient algorithm. In *NIPS*, 2017.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv*, 2017. https://arxiv.org/abs/1703.00810.

Sokolić, J., Giryes, R., Sapiro, G., and Rodrigues, M. Robust large margin deep neural networks. *IEEE Trans. Signal Process.*, 65(16):4265–4280, 2017.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15:1929–1958, 2014.

van Laarhoven, T. l_2 regularization versus batch and weight normalization. *arXiv*, 2017. https://arxiv.org/abs/1706.05350.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Y.Bengio. Manifold mixup: Better representations by interpolating hidden states. In *ICML*, 2019a.

Verma, V., Lamb, A., Kannala, J., Bengio, Y., and Lopez-Paz, D. Interpolation consistency training for semi-supervised learning. In *IJCAI*, 2019b.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Zhang, G., Wang, C., Xu, B., and Grosse, R. Three mechanisms of weight decay regularization. In *ICLR*, 2019.