Towards Understanding Generalization via Analytical Learning Theory

Kenji Kawaguchi KAWAGUCH@MIT.EDU
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139, USA

Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
University of Montreal
2900 Edouard Montpetit Blvd, Montreal, QC H3T 1J4, Canada

Vikas Verma VIKAS.VERMA@AALTO.FI
Aalto University
Konemiehentie 2, Espoo, Finland

Leslie Pack Kaelbling LPK@CSAIL.MIT.EDU
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139, USA

Abstract

This paper introduces a novel measure-theoretic theory for machine learning that does not require statistical assumptions. Based on this theory, a new regularization method in deep learning is derived and shown to outperform previous methods in CIFAR-10, CIFAR-100, and SVHN. Moreover, the proposed theory provides a theoretical basis for a family of practically successful regularization methods in deep learning. We discuss several consequences of our results on one-shot learning, representation learning, deep learning, and curriculum learning. Unlike statistical learning theory, the proposed learning theory analyzes each problem instance individually via measure theory, rather than a set of problem instances via statistics. As a result, it provides different types of results and insights when compared to statistical learning theory.

Keywords: Machine Learning, Measure Theory, Regularization method, Neural Network

1. Introduction

Statistical learning theory provides tight and illuminating results under its assumptions and for its objectives (e.g., Vapnik 1998; Mukherjee et al. 2006; Mohri et al. 2012). As the training datasets are considered as random variables, statistical learning theory was initially more concerned with the study of *data-independent* bounds based on the capacity of the hypothesis space (Vapnik, 1998), or the classical stability of learning algorithm (Bousquet and Elisseeff, 2002). Given the observations that these data-independent bounds could be overly

pessimistic for a "good" *training dataset*, *data-dependent* bounds have also been developed in statistical learning theory, such as the *luckiness framework* (Shawe-Taylor et al., 1998; Herbrich and Williamson, 2002), *empirical* Rademacher complexity of a hypothesis space (Koltchinskii and Panchenko, 2000; Bartlett et al., 2002), and the robustness of learning algorithm (Xu and Mannor, 2012).

Along this line of reasoning, we notice that the previous bounds, including data dependent ones, can be pessimistic for a "good" *problem instance*, which is defined by a tuple of a true (unknown) measure, a training dataset and a learned model (see Section 3 for further details). Accordingly, this paper proposes a learning theory designed to be strongly dependent on each individual problem instance. To achieve this goal, we directly analyse the generalization gap (difference between expected error and training error) and datasets as non-statistical objects via measure theory. This is in contrast to the setting of statistical learning theory wherein these objects are treated as random variables.

The non-statistical nature of our proposed theory can be of practical interest on its own merits. For example, the non-statistical nature captures well a situation wherein a training dataset is specified and fixed first (e.g., a UCL dataset, ImageNet, a medical image dataset, etc.), rather than remaining random with a certain distribution. Once a dataset is actually specified, there is no randomness remaining over the dataset (although one can artificially create randomness via an empirical distribution). For example, Zhang et al. (2017) empirically observed that *given a fixed (deterministic) dataset* (i.e., each of CIFAR10, ImageNet, and MNIST), test errors can be small despite the large capacity of the hypothesis space and possible instability of the learning algorithm. Understanding and explaining this empirical observation has become an active research area (Arpit et al., 2017; Krueger et al., 2017; Hoffer et al., 2017; Wu et al., 2017; Dziugaite and Roy, 2017; Dinh et al., 2017; Bartlett et al., 2017; Brutzkus et al., 2017).

For convenience within this paper, the proposed theory is called *analytical learning theory*, due to its non-statistical and analytical nature. While the scope of statistical learning theory covers both prior and posterior guarantees, analytical learning theory focuses on providing *prior* insights via *posterior* guarantees; i.e., the mathematical bounds are available before the learning is done, which provides insights a priori to understand the phenomenon and to design algorithms, but the numerical value of the bounds depend on the posterior quantities. A firm understanding of analytical learning theory requires a different style of thinking and a shift of technical basis from statistics (e.g., concentration inequalities) to measure theory. We present the foundation of analytical learning theory in Section 3 and several applications in Sections 4-5.

2. Preliminaries

In machine learning, a typical goal is to return a model $\hat{y}_{\mathcal{A}(S_m)}$ via a learning algorithm $\mathcal{A}$ given a dataset $S_m = \{s^{(1)}, \dots, s^{(m)}\}$ such that the expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] \triangleq \mathbb{E}_z[L\hat{y}_{\mathcal{A}(S_m)}(z)]$ with respect to a true (unknown) normalized measure μ is minimized. Here, $L\hat{y}$ is a function that combines a loss function ℓ and a model $\hat{y}$; e.g., in supervised learning, $L\hat{y}(z) = \ell(\hat{y}(x), y)$, where $z = (x, y)$ is a pair of an input x and a target y. Because the expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ is often not computable, we usually approximate the expected error by an empirical error $\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] \triangleq \frac{1}{m'}\sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z^{(i)})$

with a dataset $Z_{m'} = \{z^{(1)}, \ldots, z^{(m')}\}$. Accordingly, we define *the generalization gap* $\triangleq \mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$. One of the goals of learning theory is to explain and validate when and how minimizing $\hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$ is a sensible approach to minimizing $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ by analyzing the generalization gap, and to provide bounds on the performance of $\hat{y}_{\mathcal{A}(S_m)}$ on new data.

2.1 Discrepancy and variation

In the following, we define a quality of a dataset, called *discrepancy*, and a quality of a function, called *variation in the sense of Hardy and Krause*. These definitions have been used in harmonic analysis, number theory, and numerical analysis (Krause, 1903; Hardy, 1906; Hlawka, 1961; Niederreiter, 1978; Aistleitner et al., 2017). This study adopts these definitions in the context of machine learning. Intuitively, the *star-discrepancy* $D^*[T_m, \nu]$ evaluates how well a dataset $T_m = \{t^{(1)}, \ldots, t^{(m)}\}$ captures a normalized measure ν, and the *variation* $V[f]$ *in the sense of Hardy and Krause* computes how a function f varies in total w.r.t. each small perturbation of every cross combination of its variables.

2.1.1 Discrepancy of dataset with respect to a measure

For any $t = (t_1, \ldots, t_d) \in [0,1]^d$, let $B_t \triangleq [0, t_1] \times \cdots \times [0, t_d]$ be a closed axis-parallel box with one vertex at the origin. The *local discrepancy* $D[B_t; T_m, \nu]$ of a dataset $T_m = \{t^{(1)}, \ldots, t^{(m)}\}$ with respect to a normalized Borel measure ν on a set B_t is defined as

$$D[B_t; T_m, \nu] \triangleq \left(\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{B_t}(t^{(i)}) \right) - \nu\big(B_t\big)$$

where $\mathbb{1}_{B_t}$ is the indicator function of a set B_t. Figure 1 in Appendix A.1 shows an illustration of the *local discrepancy* $D[B_t; T_m, \nu]$ and related notation. The *star-discrepancy* $D^*[T_m, \nu]$ of a dataset $T_m = \{t^{(1)}, \ldots, t^{(m)}\}$ with respect to a normalized Borel measure ν is defined as

$$D^*[T_m, \nu] \triangleq \sup_{t \in [0,1]^d} \big| D[B_t; T_m, \nu] \big|.$$

2.1.2 Variations of a function

Let ∂_l be the partial derivative operator; that is, $\partial_l g(t_1, \ldots, t_k)$ is the partial derivative of a function g with respect to the l-th coordinate at a point $(t_1, \ldots, t_k)$. Let $\partial^k_{1,\ldots,k} \triangleq \partial_1, \ldots, \partial_k$. A partition P of $[0,1]^k$ with size $m_1^P, \ldots, m_k^P$ is a set of finite sequences $t_l^{(0)}, t_l^{(1)} \ldots, t_l^{(m_l^P)}$ $(l = 1, \ldots, k)$ such that $0 = t_l^{(0)} \leq t_l^{(1)} \leq \cdots \leq t_l^{(m_l^P)} = 1$ for $l = 1, \ldots, k$. We define a difference operator Δ_l^P with respect to a partition P as: given a function g and a point $(t_1, \ldots t_{l-1}, t_l^{(i)}, t_{l+1}, \ldots, t_k)$ in the partition P (for $i = 0, \ldots, m_l^P - 1$),

$$\Delta_l^P g(t_1, \ldots t_{l-1}, t_l^{(i)}, t_{l+1}, \ldots, t_k) = g(t_1, \ldots t_{l-1}, t_l^{(i+1)}, t_{l+1}, \ldots, t_k) - g(t_1, \ldots t_{l-1}, t_l^{(i)}, t_{l+1}, \ldots, t_k),$$

where $(t_1, \ldots t_{l-1}, t_l^{(i+1)}, t_{l+1}, \ldots, t_k)$ is the subsequent point in the partition P along the coordinate l. Let $\Delta^P_{1,\ldots,k} \triangleq \Delta_1^P \ldots \Delta_k^P$. Given a function f of d variables, let $f_{j_1 \ldots j_k}$ be the function restricted on $k \leq d$ variables such that $f_{j_1 \ldots j_k}(t_{j_1}, \ldots, t_{j_k}) = f(t_1, \ldots, t_d)$, where

$t_l \equiv 1$ for all $l \notin \{j_1, j_2, \dots j_k\}$. That is, $f_{j_1 \dots j_k}$ is a function of $(t_{j_1}, \dots, t_{j_k})$ with other original variables being fixed to be one.

The *variation of* $f_{j_1 \dots j_k}$ *on* $[0,1]^k$ *in the sense of Vitali* is defined as

$$V^{(k)}[f_{j_1 \dots j_k}] \triangleq \sup_{P \in \mathcal{P}_k} \sum_{i_1=1}^{m_1^P - 1} \cdots \sum_{i_k=1}^{m_k^P - 1} \left| \Delta^P_{1,\dots,k} f_{j_1 \dots j_k}(t_{j_1}^{(i_1)}, \dots, t_{j_k}^{(i_k)}) \right|,$$

where $\mathcal{P}_k$ is the set of all partitions of $[0,1]^k$. The *variation of* f *on* $[0,1]^d$ *in the sense of Hardy and Krause* is defined as

$$V[f] = \sum_{k=1}^{d} \sum_{1 \le j_1 < \cdots < j_k \le d} V^{(k)}[f_{j_1 \dots j_k}].$$

For example, if f is linear on its domain, $V[f] = \sum_{1 \le j_1 \le d} V^{(1)}[f_{j_1}]$ because $V^{(k)}[f_{j_1 \dots j_k}] = 0$ for all $k > 1$. The following proposition might be helpful in intuitively understanding the concept of the variation as well as in computing it when applicable. All the proofs in this paper are presented in Appendix B.

Proposition 1. *Suppose that* $f_{j_1 \dots j_k}$ *is a function for which* $\partial^k_{1,\dots,k} f_{j_1 \dots j_k}$ *exists on* $[0,1]^k$. *Then,*

$$V^{(k)}[f_{j_1 \dots j_k}] \le \sup_{(t_{j_1}, \dots, t_{j_k}) \in [0,1]^k} \left| \partial^k_{1,\dots,k} f_{j_1 \dots j_k}(t_{j_1}, \dots, t_{j_k}) \right|.$$

If $\partial^k_{1,\dots,k} f_{j_1 \dots j_k}$ *is also continuous on* $[0,1]^k$,

$$V^{(k)}[f_{j_1 \dots j_k}] = \int_{[0,1]^k} \left| \partial^k_{1,\dots,k} f_{j_1 \dots j_k}(t_{j_1}, ..., t_{j_k}) \right| dt_{j_1} \cdots dt_{j_k}.$$

3. A basis of analytical learning theory

This study considers the problem of analyzing the generalization gap $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$ between the expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ and the training error $\hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$. For the purpose of general applicability, our base theory analyzes a more general quantity, which is the generalization gap $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$ between the expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ and any empirical error $\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$ with any dataset $Z_{m'}$ (of size m') including the training dataset with $Z_{m'} = S_m$. Whenever we write $Z_{m'}$, it is always including the case of $Z_{m'} = S_m$; i.e., the case where the model is evaluated on the training set.

With our notation, one can observe that the generalization gap is fully and deterministically specified by a *problem instance* $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$, where we identify an omitted measure space $(\mathcal{Z}, \Sigma, \mu)$ by the measure μ for brevity. Indeed, the expected error is defined by the Lebesgue integral of a function $L\hat{y}_{\mathcal{A}(S_m)}$ on a (unknown) normalized measure space $(\mathcal{Z}, \Sigma, \mu)$ as $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] = \int_{\mathcal{Z}} L\hat{y}_{\mathcal{A}(S_m)} d\mu$, which is a deterministic mathematical object. Accordingly, we introduce the following notion of *strong instance-dependence*: a mathematical object φ is said to be *strongly instance-dependent* in the theory of the generalization gap of the tuple $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$ if the object φ is invariant under any change of any mathematical object that contains or depends on any $\bar{\mu} \neq \mu$, any $\hat{y} \neq \hat{y}_{\mathcal{A}(S_m)}$, or any $\bar{S}_m$ such that $\bar{S}_m \neq S_m$ and $\bar{S}_m \neq Z_{m'}$. Analytical learning theory is designed to provide mathematical bounds and equations that are strongly instance-dependent.

3.1 Analytical decomposition of expected error

Let $(\mathcal{Z}, \Sigma, \mu)$ be any (unknown) normalized measure space that defines the expected error, $\mathbb{E}_\mu[L\hat{y}] = \int_{\mathcal{Z}} L\hat{y}\ d\mu$. Here, the measure space may correspond to an input-target pair as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ for supervised learning, the generative hidden space $\mathcal{Z}$ of $\mathcal{X} \times \mathcal{Y}$ for unsupervised / generative models, or anything else of interest (e.g., $\mathcal{Z} = \mathcal{X}$). Let $\mathcal{T}_*\mu$ be the pushforward measure of μ under a map $\mathcal{T}$. Let $\mathcal{T}(Z_{m'}) = \{\mathcal{T}(z^{(1)}), \ldots, \mathcal{T}(z^{(m')})\}$ be the image of the dataset $Z_{m'}$ under $\mathcal{T}$. Let $|\nu|(E)$ be the total variation of a measure ν on E. For vectors $a, b \in [0,1]^d$, let $[a, b] = \{t \in [0,1]^d : a \leq t \leq b\}$, where $\leq$ denotes the product order; that is, $a \leq t$ if and only if $a_j \leq t_j$ for $j = 1, \ldots, d$. This paper adopts the convention that the infimum of the empty set is positive infinity.

Theorem 1 is introduced below to exploit the various structures in machine learning through the decomposition $L\hat{y}_{\mathcal{A}(S_m)}(z) = (f \circ \mathcal{T})(z)$ where $\mathcal{T}(z)$ is the output of a representation function and f outputs the associated loss. Here, $\mathcal{T}(z)$ can be any intermediate representation on the path from the raw data (when $\mathcal{T}(z) = z$) to the output (when $\mathcal{T}(z) = L\hat{y}(z)$). The proposed theory holds true even if the representation $\mathcal{T}(z)$ is learned. The empirical error $\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$ can be the training error with $Z_{m'} = S_m$ or the test/validation error with $Z_{m'} \neq S_m$.

Theorem 1. *For any $L\hat{y}$, let $\mathcal{F}[L\hat{y}]$ be a set of all pairs $(\mathcal{T}, f)$ such that $\mathcal{T} : (\mathcal{Z}, \Sigma) \to ([0,1]^d, \mathcal{B}([0,1]^d))$ is a measurable function, $f : ([0,1]^d, \mathcal{B}([0,1]^d)) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is of bounded variation as $V[f] < \infty$, and*

$$L\hat{y}(z) = (f \circ \mathcal{T})(z) \quad \text{for all } z \in \mathcal{Z},$$

where $\mathcal{B}(A)$ indicates the Borel σ-algebra on A. Then, for any dataset pair $(S_m, Z_{m'})$ (including $Z_{m'} = S_m$) and any $L\hat{y}_{\mathcal{A}(S_m)}$,

(i) $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] \leq \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] + \inf_{(\mathcal{T},f)\in\hat{\mathcal{F}}} V[f]\ D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$,

where $\hat{\mathcal{F}} = \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$, and

(ii) for any $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$ such that f is right-continuous component-wise,

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] = \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] + \int_{[0,1]^d} \left((\mathcal{T}_*\mu)([\mathbf{0}, t]) - \frac{1}{m'} \sum_{i=1}^{m'} \mathbb{1}_{[\mathbf{0},t]}(\mathcal{T}(z_i)) \right) d\nu_f(t),$$

where $z_i \in Z_{m'}$, and ν_f is a signed measure corresponding to f as $f(t) = \nu_f([t, \mathbf{1}]) + f(\mathbf{1})$ and $|\nu_f|([0,1]^d) = V[f]$.

The statements in Theorem 1 hold for each individual instance $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$, for example, without taking a supremum over a set of other instances. In contrast, typically in previous bounds, when asserting that an upper bound holds on $\mathbb{E}_\mu[L\hat{y}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}]$ for any $\hat{y} \in \mathcal{H}$ (with high probability), what it means is that the upper bound holds on $\sup_{\hat{y}\in\mathcal{H}}(\mathbb{E}_\mu[L\hat{y}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}])$ (with high probability). Thus, in classical bounds including data-dependent ones, as $\mathcal{H}$ gets larger and more complex, the bounds tend to become more pessimistic for the actual instance $\hat{y}_{\mathcal{A}(S_m)}$ (learned with the actual instance S_m), which is avoided in Theorem 1.

Remark 1. The bound and the equation in Theorem 1 are strongly instance-dependent, and in particular, *invariant to hypothesis space $\mathcal{H}$ and the properties of learning algorithm $\mathcal{A}$ over datasets different from a given training dataset S_m* (and $Z_{m'}$).

Remark 2. Theorem 1 together with Remark 1 has an immediate practical consequence. For example, even if the true model is contained in some "small" hypothesis space $\mathcal{H}_1$, we might want to use a much more complex "larger" hypothesis space $\mathcal{H}_2$ in practice such that the optimization becomes easier and the training trajectory reaches a better model $\hat{y}_{\mathcal{A}(S_m)}$ at the end of the learning process (e.g., over-parameterization in deep learning potentially makes the non-convex optimization easier; see Dauphin et al. 2014; Choromanska et al. 2015; Soudry and Hoffer 2017). This is consistent with both Theorem 1 and practical observations in deep learning, although it can be puzzling from the viewpoint of previous results that explicitly or implicitly penalize the use of more complex "larger" hypothesis spaces (e.g., see Zhang et al. 2017).

Remark 3. Theorem 1 does not require statistical assumptions. Thus, it is applicable even when statistical assumptions required by statistical learning theory are violated in practice.

Theorem 1 produces bounds that can be zero even with $m = 1$ (and $m' = 1$) (as an examples are provided throughout the paper), supporting the concept of one-shot learning. This is true, even if the dataset is not drawn according to the measure μ. This is because although such a dataset may incur a lager value of D^* (than a usual i.i.d. drawn dataset), it can decrease $V[f]$ in the generalization bounds of $V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$. Furthermore, by being strongly instance-dependent on the learned model $\hat{y}_{\mathcal{A}(S_m)}$, Theorem 1 supports the concept of curriculum learning (Bengio et al., 2009a). This is because curriculum learning directly guides the learning to obtain a good model $\hat{y}_{\mathcal{A}(S_m)}$, which minimizes $V[f]$ by its definition.

3.2 Additionally using statistical assumption and general bounds on D^*

By additionally using the standard i.i.d. assumption, Proposition 2 provides a general bound on the star-discrepancy $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$ that appears in Theorem 1. It is a direct consequence of (Heinrich et al., 2001, Theorem 2).

Proposition 2. *Let $\mathcal{T}(Z_{m'}) = \{\mathcal{T}(z^{(1)}), \dots, \mathcal{T}(z^{(m')})\} = \{t^{(1)}, \dots, t^{(m')}\}$ be a set of i.i.d. random variables with values on $[0,1]^d$ and distribution $\mathcal{T}_*\mu$. Then, there exists a positive constant c_1 such that for all $m' \in \mathbb{N}^+$ and all $c_2 \geq c_1$, with probability at least $1 - \delta$,*

$$D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] \leq c_2 \sqrt{\frac{d}{m'}}$$

where $\delta = \frac{1}{c_2\sqrt{d}}(c_1 c_2^2 e^{-2c_2^2})^d$ with $c_1 c_2^2 e^{-2c_2^2} < 1$.

Remark 4. Proposition 2 is not probabilistically vacuous in the sense that we can increase c_2 to obtain $1 - \delta > 0$, at the cost of increasing the constant c_2 in the bound. Forcing $1 - \delta > 0$ still keeps c_2 constant without dependence on relevant variables such as d and m'. This is because $1 - \delta > 0$ if c_2 is large enough such that $c_1 c_2^2 < e^{2c_2^2}$, which depends only on the constants.

Using Proposition 2, one can immediately provide a statistical bound via Theorem 1 over random $Z_{m'}$. To see how such a result differs from that of statistical learning theory, consider the case of $Z_{m'} = S_m$. That is, we are looking at classic training error. Whereas statistical learning theory applies a statistical assumption to the whole object $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$, analytical learning theory first decomposes $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$ into $V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ and then applies the statistical assumption only to $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$. *This makes $V[f]$ strongly instance-dependent even with the statistical assumption.* For example, with $f(z) = L\hat{y}_{\mathcal{A}(S_m)}(z)$ and $\mathcal{T}(z) = z$, if the training dataset S_m satisfies the standard i.i.d. assumption, we have that with high probability,

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}] \leq c_2 V[L\hat{y}_{\mathcal{A}(S_m)}]\sqrt{\frac{d}{m}}, \tag{1}$$

where the term $V[L\hat{y}_{\mathcal{A}(S_m)}]$ is strongly instance-dependent.

In Equation (1), it is unnecessary for m to approach infinity in order for the generalization gap to go to zero. As an extreme example, if the variation of $\hat{y}_{\mathcal{A}(S_m)}$ aligns with that of the true y (i.e., $L\hat{y}_{\mathcal{A}(S_m)}$ is constant), we have that $V[L\hat{y}_{\mathcal{A}(S_m)}] = 0$ and the generalization gap becomes zero even with $m = 1$. This example illustrates the fact that Theorem 1 supports the concept of one-shot learning via the transfer of knowledge into the resulting model $\hat{y}_{\mathcal{A}(S_m)}$.

For the purpose of the non-statistical decomposition of $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$, instead of Theorem 1, we might be tempted to conduct a simpler decomposition with the Hölder inequality or its variants. However, such a simpler decomposition is dominated by a difference between the true measure and the empirical measure on an arbitrary set in high-dimensional space, which suffers from the curse of dimensionality. Indeed, the proof of Theorem 1 is devoted to reformulating $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$ via the equivalence in the measure and the variation *before taking any inequality*, so that we can avoid such an issue. That is, the star-discrepancy evaluates the difference in the measures on high-dimensional boxes with one vertex at the origin, instead of on an arbitrary set.

The following proposition proves the existence of a dataset $Z_{m'}$ with a convergence rate of $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ that is asymptotically faster than $O(\sqrt{1/m})$ in terms of the dataset size m'. This is a direct consequence of (Aistleitner and Dick, 2014, Theorem 2).

Proposition 3. *Assume that $\mathcal{T}$ is a surjection. Let $\mathcal{T}_*\mu$ be any (non-negative) normalized Borel measure on $[0,1]^d$. Then, for any $m' \in \mathbb{N}^+$, there exists a dataset $Z_{m'}$ such that*

$$D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] \leq 63\sqrt{d}\frac{(2 + \log_2 m')^{(3d+1)/2}}{m'}.$$

This can be of interest when we can choose $\mathcal{T}$ to make d small without increasing $V[f]$ too much; i.e., it then provides a faster convergence rate than usual statistical guarantees. If $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ (which is true in many practical cases), we can have $d = 1$ by setting $\mathcal{T} : \mathcal{Z} \to [0,1]$, because there exists a bijection between $\mathcal{Z}$ and $(0,1)$. Then, although the variation of $\mathcal{T}$ is unbounded in general, $V[f]$ can be still small. For example, it is still zero if the variation of $\hat{y}_{\mathcal{A}(S_m)}$ aligns with that of the true y in this space of $[0,1]$.

3.3 General examples

The following example provides insights on the quality of *learned* representations:

Example 1. Let $\mathcal{T}(z) = (\phi(x), v)$ where ϕ is a map of any *learned* representation and v is a variable such that there exists a function f satisfying $L\hat{y}_{\mathcal{A}(S_m)}(z) = f(\phi(x), v)$ (for supervised learning, setting $v := y$ always satisfies this condition regardless of the information contained in $\phi(x)$). For example, $\phi(x)$ may represent the output of any intermediate hidden layer in deep learning (possibly the last hidden layer), and v may encode the noise left in the label y. Let f be a map such that $L\hat{y}_{\mathcal{A}(S_m)} = f(\mathcal{T}(z))$. Then, if $V[f] < \infty$, Theorem 1 implies that for any dataset pair $(S_m, Z_{m'})$ (including $Z_{m'} = S_m$), $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] \leq \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] + V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$.

Example 1 partially supports the concept of the *disentanglement* in deep learning (Bengio et al., 2009b) and proposes a new concrete method to measure the degree of disentanglement as follows. In the definition of $V[f] = \sum_{k=1}^{d} \sum_{1 \leq j_1 < \cdots < j_k \leq d} V^{(k)}[f_{j_1 \ldots j_k}]$, each term $V^{(k)}[f_{j_1 \ldots j_k}]$ can be viewed as measuring how *entangled* the $j_1, \ldots, j_k$-th variables are in a space of a learned (hidden) representation. We can observe this from the definition of $V^{(k)}[f_{j_1 \ldots j_k}]$ or from Proposition 1 as: $V^{(k)}[f_{j_1 \ldots j_k}] = \int_{[0,1]^k} \left| \partial^k_{1,\ldots,k} f_{j_1 \ldots j_k}(t_{j_1}, \ldots, t_{j_k}) \right| dt_{j_1} \cdots dt_{j_k}$, where $\partial^k_{1,\ldots,k} f_{j_1 \ldots j_k}(t_{j_1}, \ldots, t_{j_k})$ is the k-th order *cross* partial derivatives across the $j_1, \ldots, j_k$-th variables. If all the variables in a space of a learned (hidden) representation are completely disentangled in this sense, $V^{(k)}[f_{j_1 \ldots j_k}] = 0$ for all $k \geq 2$ and $V[f]$ is minimized to $V[f] = \sum_{j_1=1}^{d} V^{(1)}[f_{j_1}]$. Additionally, Appendices A.5 and A.6 provide discussion of the effect of flatness in measures and higher-order derivatives.

One of the reasons why analytical learning theory is complementary to statistical learning theory is the fact that we can naturally combine the both. For example, in Example 1, we cannot directly adopt the probabilistic bound on $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ from Section 3.2, if $\mathcal{T}(S_m)$ does not satisfy the i.i.d. assumption because $\mathcal{T}$ depends on the whole dataset S_m. In this case, to analyze $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$, we can use the approaches in statistical learning theory, such as Rademacher complexity or covering number. To see this, consider a set Φ such that $\mathcal{T} \in \Phi$ and Φ is independent of S_m. Then, by applying Proposition 2 with a union bound over a cover of Φ, we can obtain probabilistic bounds on D^* with the log of the *covering number* of Φ for all representations $\mathcal{T}' \in \Phi$. As in data-dependent approaches (e.g., Bartlett et al. 2017, Lemma A.9), one can also consider a sequence of sets $\{\Phi_j\}_j$ such that $\mathcal{T} \in \cup_j \Phi_j$, and one can obtain a data-dependent bound on $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ via a complexity of Φ_j.

The following example establishes the tightness of Theorem 1 (i) with the 0-1 loss in general, where $\iota : \{0, 1\} \to [0, 1]$ is an inclusion map:

Example 2. Theorem 1 (i) is tight in multi-class classification with 0-1 loss as follows. Let $\mathcal{T} = \iota \circ L\hat{y}_{\mathcal{A}(S_m)}$. Let f be an identity map. Then, $V[f] = 1$ and $L\hat{y}_{\mathcal{A}(S_m)}(z) = (f \circ \mathcal{T})(z)$ for all $z \in \mathcal{Z}$. Then, the pair of $\mathcal{T}$ and f satisfies the condition in Theorem 1 as $L\hat{y}_{\mathcal{A}(S_m)}$ and ι are measurable functions. Thus, from Theorem 1, $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] \leq V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] = |(\mathcal{T}_*\mu)(\{1\}) - \mathbb{E}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]|)$ (see Appendix B.5 for this derivation), which establishes a tightness of Theorem 1 (i) with the 0-1 loss as follows:

for any dataset pair $(S_m, Z_{m'})$ (including $Z_{m'} = S_m$), $\left|\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]\right| = V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$.

The following example applies Theorem 1 to a raw representation space $\mathcal{T}(z) = z$ and a loss space $\mathcal{T}(z) = (\iota \circ L\hat{y}_{\mathcal{A}(S_m)})(z)$:

Example 3. Consider a normalized domain $\mathcal{Z} = [0,1]^{d_z}$ and a Borel measure μ on $\mathcal{Z}$. For example, $\mathcal{Z}$ can be an unknown hidden generative space or an input-output space ($\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$). Let us apply Theorem 1 to this measure space with $\mathcal{T}(z) = z$ and $f = L\hat{y}_{\mathcal{A}(S_m)}$. Then, if $V[L\hat{y}_{\mathcal{A}(S_m)}] < \infty$, Theorem 1 implies that for any dataset pair $(S_m, Z_{m'})$ (including $Z_{m'} = S_m$) and any $L\hat{y}_{\mathcal{A}(S_m)}$, $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] \leq \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] + V[L\hat{y}_{\mathcal{A}(S_m)}]D^*[\mu, Z_{m'}]$.

Example 3 indicates that we can regularize $V[L\hat{y}_{\mathcal{A}(S_m)}]$ in some space $\mathcal{Z}$ to control the generalization gap. For example, letting the model $\hat{y}_{\mathcal{A}(S_m)}$ be invariant to a subspace that is not essential for prediction decreases the bound on $V[L\hat{y}_{\mathcal{A}(S_m)}]$. As an extreme example, if $x = g(y, \xi)$ with some generative function g and noise ξ (i.e., a setting considered in an information theoretic approach), $\hat{y}_{\mathcal{A}(S_m)}$ being invariant to ξ results in a smaller bound on $V[L\hat{y}_{\mathcal{A}(S_m)}]$. This is qualitatively related to an information theoretic observation such as in (Achille and Soatto, 2017).

4. Application to linear regression

Even in the classical setting of linear regression, recent papers (Zhang et al. 2017, Section 5; Kawaguchi et al. 2017, Section 3; Poggio et al. 2017, Section 5) suggest the need for further theoretical studies to better understand the question of precisely what makes a learned model generalize well, especially with an *arbitrarily* rich hypothesis space and algorithmic instability. Theorem 1 studies the question abstractly for machine learning in general. As a simple concrete example, this section considers linear regression. However, note that the theoretical results in this section can be directly applied to deep learning as described in Remark 7.

Let $S_m = \{s^{(i)}\}_{i=1}^m$ be a training dataset of the input-target pairs where $s^{(i)} = (x^{(i)}, y^{(i)})$. Let $\hat{y}_{\mathcal{A}(S_m)} = \hat{W}\phi(\cdot)$ be the learned model at the end of any training process. For example, in empirical risk minimization, the matrix $\hat{W}$ is an output of the training process, $\hat{W} := \operatorname{argmin}_W \hat{\mathbb{E}}_{S_m}[\frac{1}{2}\|W\phi(x) - y\|_2^2]$. Here, $\phi : (\mathcal{X}, \Sigma_x) \to ([0,1]^{d_\phi}, \mathcal{B}([0,1]^{d_\phi}))$ is any normalized measurable function, corresponding to fixed features. For any given variable v, let d_v be the dimensionality of the variable v. The goal is to minimize the expected error $\mathbb{E}_s[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2]$ of the learned model $\hat{W}\phi(\cdot)$.

4.1 Domains with linear Gaussian labels

In this subsection only, we assume that the target output y is structured such that $y = W^*\phi(x) + \xi$, where ξ is a zero-mean random variable independent of x. Many columns of W^* can be zeros (i.e., sparse) such that $W^*\phi(x)$ uses a small portion of the feature vector $\phi(x)$. Thus, this label assumption can be satisfied by including a sufficient number of elements from a basis with uniform approximation power (e.g., polynomial basis, Fourier basis, a set of step functions, etc.) to the feature vector $\phi(x)$ up to a desired approximation error. Note that we do not assume any knowledge of W^*.

Let μ_x be the (unknown) normalized measure for the input x (corresponding to the marginal distribution of (x, y)). Let $X_m = \{x^{(i)}\}_{i=1}^m$ and $\tilde{S}_m = \{(x^{(i)}, \xi^{(i)})\}_{i=1}^m$ be the input part and the (unknown) input-noise part of the same training dataset as S_m, respectively. We do not assume access to $\tilde{S}_m$. Let W_l be the l-th column of the matrix W.

Theorem 2. *Assume that the labels are structured as described above and $\|\hat{W} - W^*\| < \infty$. Then, Theorem 1 implies that*

$$\mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] - \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] \leq V[f]D^*[\phi_*\mu_x, \phi(X_m)] + A_1 + A_2, \quad (2)$$

where $f(t) = \frac{1}{2}\|\hat{W}t - W^*t\|_2^2$, $A_1 = \hat{\mathbb{E}}_{\tilde{S}_m}[\xi^\top(\hat{W} - W^*)\phi(x)]$, $A_2 = \mathbb{E}_\xi[\|\xi\|_2^2] - \hat{\mathbb{E}}_{\tilde{S}_m}[\|\xi\|_2^2]$, and

$$V[f] \leq \sum_{l=1}^{d_\phi} \|(\hat{W}_l - W_l^*)^\top(\hat{W} - W^*)\|_1 + \sum_{1 \leq l < l' \leq d_\phi} |(\hat{W}_l - W_l^*)^\top(\hat{W}_{l'} - W_{l'}^*)|.$$

Remark 5. Theorem 2 is tight in terms of both the minimizer and its value, which is explained below. The bound in Theorem 2 (i.e., the right-hand-side of Equation (2)) is minimized (to be the noise term A_2 only) if and only if $\hat{W} = W^*$ (see Appendix A.4 for pathological cases). Therefore, minimizing the bound in Theorem 2 is equivalent to minimizing the expected error $\mathbb{E}_s[\|\hat{W}\phi(x) - y\|_2^2]$ or generalization error (see Appendix A.4 for further details). Furthermore, the bound in Theorem 2 holds with equality if $\hat{W} = W^*$. Thus, the bound is tight in terms of the minimizer and its value.

Remark 6. For $D^*[\phi_*\mu_x, \phi(X_m)]$ and A_2, we can straightforwardly apply the probabilistic bounds under the standard i.i.d. statistical assumption. From Proposition 2, with high probability, $D^*[\phi_*\mu_x, \phi(X_m)] \leq O(\sqrt{d_\phi/m})$. From Hoeffding's inequality with $M \geq \|\xi\|_2^2$, with probability at least $1 - \delta$, $A_2 \leq M\sqrt{\ln(1/\delta)/2m}$.

It is not necessary for $D^*[\phi_*\mu_x, \phi(X_m)]$ to approach zero to minimize the expected error; irrespective of whether the training dataset satisfies a certain statistical assumption to bound $D^*[\phi_*\mu_x, \phi(X_m)]$, we can minimize the expected error via making $\hat{W}$ closer to W^* as shown in Theorem 2.

4.2 Domains with unstructured/random labels

In this subsection, we discard the linear Gaussian label assumption in the previous subsection and consider the worst case scenario where y is a variable independent of x. This corresponds to the random label experiment by Zhang et al. (2017), which posed another question: how to theoretically distinguish the generalization behaviors with structured labels from those with random labels. Generalization behaviors in practice are expected to be significantly different in problems with structured labels or random labels, even when the hypothesis space and learning algorithm remain unchanged.

As desired, Theorem 3 (unstructured labels) predicts a completely different generalization behavior from that in Theorem 2 (structured labels), even with an identical hypothesis space and learning algorithm. Here, we consider the normalization of y such that $y \in [0, 1]^{d_y}$. Let μ_s be the (unknown) normalized measure for the pair $s = (x, y)$.

Theorem 3. *Assume unstructured labels as described above. Let* $M = \sup_{t\in[0,1]} \|\hat{W}t - y\|_\infty$. *Assume that* $\|\hat{W}\| < \infty$ *and* $M < \infty$. *Then, Theorem 1 implies that*

$$\mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] - \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] \leq V[f]D^*[\mathcal{T}_*\mu_s, \mathcal{T}(S_m)], \tag{3}$$

where $\mathcal{T}(s) = (\phi(x), y)$, $f(t, y) = \frac{1}{2}\|\hat{W}t - y\|_2^2$, *and*

$$V[f] \leq (M+1)\sum_{l=1}^{d_\phi} \|\hat{W}_l\|_1 + \sum_{1\leq l<l'\leq d_\phi} |\hat{W}_l{}^\top \hat{W}_{l'}| + d_y M.$$

Unlike in the structured case (Theorem 2), minimizing the bound on the generalization gap in the unstructured case requires us to control the norm of $\hat{W}$, which corresponds to the traditional results from statistical learning theory. As in statistical learning theory, the generalization gap in Theorem 3 (unstructured labels) goes to zero as $D^*[\mathcal{T}_*\mu_s, \mathcal{T}(S_m)]$ approaches zero via certain statistical assumption: e.g., via Proposition 2, with high probability, $D^*[\mathcal{T}_*\mu_s, \mathcal{T}(S_m)] \leq O(\sqrt{(d_\phi + d_y)/m})$. This is in contrast to Theorem 2 (the structured case) where we require no statistical assumption for the generalization gap to approach zero within polynomial sample complexity.

Remark 7. (Theorems 2 and 3 on representation learning) *Theorems 2 and 3 hold true, even with learned representations* ϕ, *instead of fixed features.* Let $\phi(x)$ represent the last hidden layer in a neural network or the learned representation in representation learning in general. Consider the squared loss (square of output minus target). Then, the identical proofs of Theorems 2 and 3 work with the learned representation ϕ.

5. From analytical learning theory to methods in deep learning

This section further demonstrates the practical relevance of analytical learning theory by showing its application to derive empirical methods. The complete code of our method and experiments is publicly available at `https://github.com/Learning-and-Intelligent-Systems/DualCutout`.

5.1 Theory

We consider multi-class classification with a set Y of class labels. Then,

$$\begin{aligned}
&\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}] \\
&= \sum_{y\in Y} p(y)\mathbb{E}_{\mu_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{p}(y)\hat{\mathbb{E}}_{S_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}] \pm p(y)\hat{\mathbb{E}}_{S_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}] \\
&= \sum_{y\in Y} p(y)\left(\mathbb{E}_{\mu_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}]\right) + (p(y) - \hat{p}(y))\,\hat{\mathbb{E}}_{S_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}],
\end{aligned}$$

where $\hat{p}(y) \triangleq \frac{|S_{x|y}|}{m}$, $\hat{\mathbb{E}}_{S_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}] \triangleq \frac{1}{|S_{x|y}|}\sum_{x\in S_{x|y}} L\hat{y}_{\mathcal{A}(S_m)}(x, y)$, and $S_{x|y} \subseteq S_m$ is the set of the training input points x of the label y. Within the sum over y, by applying Theorem

1 (i) to each first term and Hoeffding's inequality to each second term, we have that with probability at least $1-\delta/2$,

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}]$$
$$\leq \sum_{y\in Y} p(y) \inf_{(\mathcal{T}_y, f_y)\in\hat{\mathcal{F}}_y} V[f_y] D^*[(\mathcal{T}_y)_*\mu_{x|y}, \mathcal{T}_y(S_{x|y})] + \hat{\mathbb{E}}_{S_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}]\sqrt{\frac{\log 2/\delta}{2|S_{x|y}|}}.$$

Assume that there exists a generative (unknown) hidden space $G : (y, z) \mapsto x$ where the true label of the input $x = G(y, z)$ is y for any z in the its normalized domain. We now set $\mathcal{T}_y : (x, y) \mapsto z$ and $f_y : z \mapsto L\hat{y}_{\mathcal{A}(S_m)}(x, y)$ where z is the *unknown* hidden space that does not affect the true label. This choice does not depend on the dataset although it is unknown. Thus, by applying Proposition 2 with these (T_y, f_y), we have that with probability at least $1-\delta$,

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{S_m}[L\hat{y}_{\mathcal{A}(S_m)}] \leq \sum_{y\in Y} c_2 p(y) V[f_y]\sqrt{\frac{d_z}{|S_{x|y}|}} + \hat{\mathbb{E}}_{S_{x|y}}[L\hat{y}_{\mathcal{A}(S_m)}]\sqrt{\frac{\log 2/\delta}{2|S_{x|y}|}}, \quad (4)$$

where d_z is the dimensionality of the generative hidden space of z and c_2 is a constant defined in Proposition 2.

Equation 4 tells us that if $V[f_y]$ is bounded by a constant, the generalization error *goes to zero in polynomial sample complexity even with an arbitrarily complex hypothesis space and non-stable learning algorithm.* If the loss is 0-1 loss, $V[f_y] = 0$ when $(y, z) \mapsto \hat{y}_{\mathcal{A}(S_m)}(x)$ is invariant over z. In other words, to control $V[f_y]$, we want to have a model that is more invariant over the space of z, which intuitively makes sense.

5.2 Methods

The above result provides a theoretical basis for a family of *consistency-based regularization* methods, including Π-Model (Laine and Aila, 2016), virtual adversarial training (Miyato et al., 2016) and regularization with stochastic transformations and perturbations (Sajjadi et al., 2016). These consistency-based regularization methods have been empirically successful heuristics. These methods are based on the intuition that perturbations of a data point $x \mapsto \tilde{x}$ should not change the output of a model as $\hat{y}(x) \approx \hat{y}(\tilde{x})$ if the true label is invariant under the perturbation; i.e., $y^*(x) = y^*(\tilde{x})$ where y^* outputs a correct label. This intuitive goal is achieved by minimizing $d(\hat{y}(x), \hat{y}(\tilde{x}))$ with respect to the trainable model $\hat{y}$, where $d(,)$ measures a distance between the two outputs. In Equation 4, these methods can be viewed to control $V[f_y]$ by making the model $\hat{y}$ more invariant over the space of z. Therefore, our theory formalizes the intuition of these regularization methods in terms of the generalization gap.

In order to more effectively minimize the bound on the generalization gap in Equation 4, we propose a new regularization method, called *dual-cutout*. For each training input x, our dual-cutout method minimizes the following regularization loss $\ell_{\text{reg}}(x, \theta)$ with respect to θ (in addition to the original classification loss):

$$\ell_{\text{reg}}(x, \theta) = \int_{(x_1,x_2)} \|h(x_1, \theta) - h(x_2, \theta)\|_2^2 dP(x_1, x_2|x),$$

where $h(x', \theta)$ is the post-softmax output of the last layer of a neural network with parameters θ (given an input x'), and $(x_1, x_2) \sim P(x_1, x_2|x)$ are two randomly sampled inputs of two random cutouts of a given natural input x. Here, we set $P(x_1, x_2|x) = P(x_1|x)P(x_2|x)$, and $P(x_1|x) = P(x_2|x)$ is the probability distribution over random cutout input x_1 given a original (non-cutout) input x; i.e., $P(x_1|x) = P(x_2|x)$ represents the same random cutout procedure as single-cutout method in the previous paper (DeVries and Taylor, 2017). As this additional regularization loss gets smaller, the model becomes more insensitive over the hidden space of z, implicitly minimizing $V[f_y]$ and the bound on the generalization gap in Equation 4.

Table 1 compares the test error of dual-cutout against single-cutout and the standard method for three benchmark datasets, namely CIFAR10, CIFAR100 and SVHN. Dual-cutout outperforms baseline methods by a significant margin.

Table 1: Test error (%) with different regularization methods.

Method	CIFAR-10	CIFAR-100	SVHN
Standard	3.79 ± 0.07	19.85 ± 0.14	2.47 ± 0.04
Single-cutout	3.19 ± 0.09	18.13 ± 0.28	2.23 ± 0.03
Dual-cutout	2.61 ± 0.04	17.54 ± 0.09	2.06 ± 0.06

We conducted all the experiments with the WideResNet28_10 (Zagoruyko and Komodakis, 2016) architecture and report the test errors at the end of 300 training epochs. We used SGD with the learning rate 0.1 and the momentum 0.9. At each step of SGD, to minimize the regularization loss $\ell_{\text{reg}}(x, \theta)$ of dual-cutout, we used the sampled gradient $\nabla_\theta \|h(x_1, \theta) - h(x_2, \theta)\|_2^2$ where (x_1, x_2) is sampled as $(x_1, x_2) \sim P(x_1, x_2|x)$. The learning rate was annealed at epochs 150 and 225 by a factor of 0.1. We used standard data-augmentation and preprocessing for all the datasets. For each dataset, we choose the cutout size as reported in (DeVries and Taylor, 2017). We performed five trials of each experiment and report the standard deviation and mean of test error in Table 1.

6. Discussion

Table 2 summarizes the major simplified differences between statistical learning theory and analytical learning theory. Because of the differences in the assumptions and the objectives, the proposed learning theory is not directly comparable in terms of sample complexity against previous learning theory. Instead of focusing on comparable sample-complexity, an-

Table 2: A simplified comparison, wherein GG denotes the generalization gap

	Statistical Learning Theory	Analytical Learning Theory
GG is characterized by	hypothesis spaces $\mathcal{H}$ or algorithms $\mathcal{A}$	a learned model $\hat{y}_{\mathcal{A}(S_m)}$
GG is decomposed via	statistics	measure theory
Statistical assumption	is required	can be additionally used
Main focus is when	a (training) dataset S_m remains random	a (training) dataset S_m is given
Bounds on GG are	not strongly instance-dependent	strongly instance-dependent

alytical learning theory focuses on complementing previous learning theory by providing additional practical insights. Indeed, the real-world phenomena that are analyzed are different in statistical learning theory and analytical learning theory. Typically in statistical learning theory, some upper bound holds over a fixed $\mathcal{H}$ or a fixed $\mathcal{A}$ *with high probability over different random datasets*. In contrast, in analytical learning theory, some upper bound holds *individually for each problem instance*.

An another difference between statistical learning theory and analytical learning theory lies in the property of strong instance-dependence. Any generalization bound that depends on a non-singleton hypothesis space $\mathcal{H} \neq \{\hat{y}_{\mathcal{A}(S_m)}\}$, such as ones with Rademacher complexity and VC dimension, is *not* strongly instance-dependent because the non-singleton hypothesis space contains $\hat{y} \neq \hat{y}_{\mathcal{A}(S_m)}$, and the bound is not invariant under an arbitrary change of $\mathcal{H}$. The definition of stability itself depends on $\bar{S}_m$ that is not equal to S_m and $Z_{m'}$ (Bousquet and Elisseeff, 2002), making the corresponding bounds be *not* strongly instance-dependent. Moreover, a generalization bound that depends on a concept of random datasets $\bar{S}_m$ different from S_m and $Z_{m'}$ (e.g., an additive term $O(\sqrt{1/m})$ that measures a deviation from an expectation over $\bar{S}_m \neq S_m, Z_{m'}$) is *not* strongly instance-dependent, because the bound is not invariant under an arbitrary change of $\bar{S}_m$.

Data dependence does not imply strong instance-dependence. For example, in the data-dependent bounds of the luckiness framework (Shawe-Taylor et al., 1998; Herbrich and Williamson, 2002), the definition of ω-smallness of the luckiness function contains a non-singleton hypothesis space $\mathcal{H}$, a sequence of non-singleton hypothesis spaces (ordered in a data-dependent way by a luckiness function), and a supremum over $\mathcal{H}$ with the probability over datasets $\bar{S}_m \neq S_m$ (with $Z_{m'} = S_m$) (e.g., see Definition 4 in Herbrich and Williamson 2002 with contraposition). As exemplified in the luckiness framework, one can usually turn both data-dependent and data-independent bounds into more data-dependent ones by considering a sequence of hypothesis spaces or sets of learning algorithms. However, such data-dependent bounds still contain the complexity of a non-singleton hypothesis space (and dependence on the definition of the sequence). The data-dependent bounds with empirical Rademacher complexity (Koltchinskii and Panchenko, 2000; Bartlett et al., 2002) also depend on a non-singleton hypothesis space and its empirical Rademacher complexity. Moreover, the definition of robustness itself depends on $\bar{S}_m$, which is not equal to S_m or $Z_{m'}$ (Xu and Mannor, 2012). Therefore, all of these data-dependent bounds are not strongly instance-dependent.

The fact that Theorem 1 is invariant to the complexity of hypothesis space $\mathcal{H}$ and certain details of a learning algorithm $\mathcal{A}$ can be both advantageous and disadvantageous, depending on the objective of the analysis. As we move towards the goal of artificial intelligence, $\mathcal{H}$ and $\mathcal{A}$ would become extremely complex, which can pose a challenge in theory. From this viewpoint, analytical learning theory can also be considered as a methodology to avoid such a challenge, producing theoretical insights for intelligent systems with arbitrarily complex $\mathcal{H}$ and $\mathcal{A}$, so long as other conditions are imposed on the actual functions being computed by them.

Acknowledgments

We gratefully acknowledge support from NSF grants 1420316, 1523767 and 1723381, from AFOSR FA9550-17-1-0165, from ONR grant N00014-14-1-0486, and from ARO grant W911 NF1410433, as well as support from NSERC, CIFAR and Canada Research Chairs. Vikas Verma was supported by Academy of Finland project 13312683 / Raiko Tapani AT kulut.

References

Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*, 2017.

Christoph Aistleitner. Covering numbers, dyadic chaining and discrepancy. *Journal of Complexity*, 27(6):531–540, 2011.

Christoph Aistleitner and Josef Dick. Low-discrepancy point sets for non-uniform measures. *Acta Arithmetica*, 163(4):345–369, 2014.

Christoph Aistleitner and Josef Dick. Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality. *Acta Arithmetica*, 2(167):143–171, 2015.

Christoph Aistleitner, Florian Pausinger, Anne Marie Svane, and Robert F Tichy. On functions of bounded variation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 162, pages 405–418. Cambridge University Press, 2017.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017.

Robert B Ash and Catherine Doleans-Dade. *Probability and measure theory*. Academic Press, 2000.

Peter L Bartlett, Stéphane Boucheron, and Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6241–6250, 2017.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009a.

Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *International Conference on Machine Learning*, pages 552–560, 2013.

Yoshua Bengio et al. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009b.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

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.

Anna Choromanska, MIkael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 192–204, 2015.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 2933–2941, 2014.

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

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, 2017.

Richard M Dudley. A course on empirical processes. In *Ecole d'été de Probabilités de Saint-Flour XII-1982*, pages 1–142. Springer, 1984.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.

Godfrey H Hardy. On double Fourier series and especially those which represent the double zeta-function with real and incommensurable parameters. *Quart. J. Math*, 37(1):53–79, 1906.

Stefan Heinrich, Henryk Woźniakowski, Grzegorz Wasilkowski, and Erich Novak. The inverse of the star-discrepancy depends linearly on the dimension. *Acta Arithmetica*, 3(96): 279–302, 2001.

Ralf Herbrich and Robert C Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2002.

Edmund Hlawka. Funktionen von beschränkter variatiou in der theorie der gleichverteilung. *Annali di Matematica Pura ed Applicata*, 54(1):325–333, 1961.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *arXiv preprint arXiv:1705.08741*, 2017.

Kenji Kawaguchi. Bounded optimal exploration in MDP. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Bayesian optimization with exponential convergence. In *Advances in Neural Information Processing (NIPS)*, 2015.

Kenji Kawaguchi, Yu Maruyama, and Xiaoyu Zheng. Global continuous optimization with error bound and fast convergence. *Journal of Artificial Intelligence Research*, 56:153–195, 2016.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

M Krause. Über fouriersche reihen mit zwei veränderlichen grössen. *Leipziger Ber*, 55: 164–197, 1903.

David Krueger, Nicolas Ballas, Stanislaw Jastrzebski, Devansh Arpit, Maxinder S Kanwal, Tegan Maharaj, Emmanuel Bengio, Asja Fischer, and Aaron Courville. Deep nets don't learn via memorization. In *Workshop Track of International Conference on Learning Representations*, 2017.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference for Learning Representations*, 2016.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. In *International Conference for Learning Representations*, 2016.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

Harald Niederreiter. Quasi-monte carlo methods and pseudo-random numbers. *Bulletin of the American Mathematical Society*, 84(6):957–1041, 1978.

Tomaso Poggio, Kenji Kawaguchi, Qianli Liao, Brando Miranda, Lorenzo Rosasco, Xavier Boix, Jack Hidary, and Hrushikesh Mhaskar. Theory of deep learning iii: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 1163–1171, 2016.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

Vladimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Appendix

Appendix A contains additional discussions to facilitate understanding this paper. Appendix B includes all the proofs of the theoretical results.

Appendix A. Additional discussions

A.1 An illustration of discrepancy

Figure 1 shows an illustration of the *local discrepancy* $D[B_t; T_m, \nu]$ and related notation in two dimensional space.

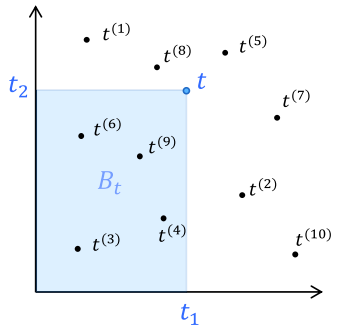


Figure 1: The local discrepancy $D[B_t; T_m, \nu]$ evaluates the difference between the empirical measure of the box B_t (the normalized number of data points in the box B_t, which is $4/10$) and the measure ν of the box B_t (the measure of the blue region)

A.2 An illustration of a difference in the scopes of statistical and analytical learning theories

Figure 2 shows a graphical illustration of a difference in the scopes of statistical learning theory and analytical learning theory. Here, μ^m is the product measure.

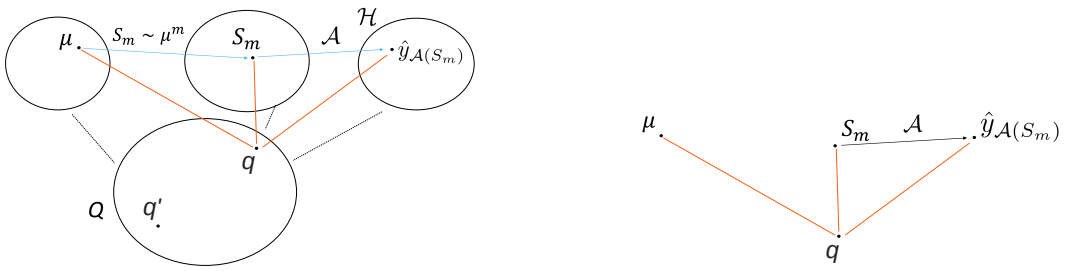
In the setting of statistical learning theory (Figure 2 (a)), our typical goal is to analyze the random expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ over the random datasets S_m by fixing a hypothesis

space and/or learning algorithm over random datasets. Due to the randomness over S_m, we do not know where q exactly lands in Q. The lower bound and necessary condition in the setting of statistical learning theory is typically obtained via a worst-case instance q' in Q. For example, classical no free lunch theorems and lower bounds on the generalization gap via VC dimension (e.g., Mohri et al. 2012, Section 3.4) have been derived with the worst-case distribution characterizing q' in Q. Such a necessary condition is only proven to be necessary for the worst-case $q' \in Q$, but is *not* proven to be necessary for other "good" cases $q \neq q'$. Intuitively, we are typically analyzing the quality of the set Q, instead of each individual $q \in Q$.

In this view, it becomes clear what is going on in some empirical observations such as in (Zhang et al., 2017). Intuitively, whereas statistical learning theory focuses more on analyzing the set Q, each element such as q (e.g., a "good" case or structured label case) and q' (e.g., the worst-case or random label case) can significantly differ from each other. Data-dependent analyses in statistical learning theory can be viewed as the ways to decrease the size of Q around each q.

In contrast, analytical learning theory (Figure 2 (b)) focuses on each q only, allowing tighter results for each "good" $q \in Q$ beyond the possibly "bad" quality of the set Q overall.

It is important to note that analyzing the set Q is of great interest on its own merits, and statistical learning theory has advantages over our proposed learning theory in this sense. Indeed, analyzing a set Q is a natural task along the way of thinking in theoretical computer science (e.g., categorizing a set Q of problem instances into polynomial solvable set or not). This situation where theory focuses more on Q and practical studies care about each $q \in Q$ is prevalent in computer science even outside the learning theory. For example, the size of Q analyzed in theory for optimal exploration in Markov decision processes (MDPs) has been shown to be often too loose for each practical problem instance $q \in Q$, and a way to partially mitigate this issue was recently proposed (Kawaguchi, 2016). Similarly, global optimization methods including Bayesian optimization approaches may suffer from a large



(a) Statistical learning theory

(b) Analytical learning theory

Figure 2: An illustration of a difference in the scopes with $Z_{m'} = S_m$: q represents a query about the generalization gap of a learned model $y_{\mathcal{A}(S_m)}$, which is a deterministic quantity of the tuple $(\mu, S_m, L\hat{y}_{\mathcal{A}(S_m)})$. Intuitively, whereas analytical learning theory analyzes q directly, statistical learning theory focuses more on analyzing the set Q that contains q. The set Q is defined by the sets of possible measures μ and randomly-drawn different datasets S_m and the hypothesis space $\mathcal{H}$ or learning algorithm $\mathcal{A}$.

complex Q for each practical problem instance $q \in Q$, which was partially mitigated in recent studies (Kawaguchi et al., 2015, 2016).

Furthermore, the issues of characterizing a set Q only via a worst-case instance q' (i.e., worst-case analysis) are well-recognized in theoretical computer science, and so-called *beyond worst-case analysis* (e.g., smoothed analysis) is an active research area to mitigate the issues. Moreover, a certain qualitative property of the set Q might tightly capture that of each instance $q \in Q$. However, to prove such an assertion, proving that a upper bound on $\forall q \in Q$ matches a lower bound on $\exists q' \in Q$ is insufficient.

A.3 On usage of statistical assumption with $Z_{m'} = S_m$

Using a statistical assumption on a dataset $Z_{m'}$ with $Z_{m'} \neq S_m$ is consistent with a practical situation where a dataset S_m is given first instead of remaining random. For $Z_{m'} = S_m$, we can view this formulation as a mathematical modeling of the following situation. Consider S_m as a random variable when collecting a dataset S_m, and then condition on the event of getting the collected dataset S_m once S_m is specified, focusing on minimization of the (future) expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ of the model $\hat{y}_{\mathcal{A}(S_m)}$ *learned with this particular specified dataset* S_m.

In this view, we can observe that if we draw an i.i.d. dataset S_m, a dataset S_m is guaranteed to be statistically "good" with high probability in terms of $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ (e.g., $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)] \leq c_2\sqrt{\frac{d}{m}}$ via Proposition 2). Thus, collecting a training dataset in a manner that satisfies the i.i.d. condition is an effective method. However, once a dataset S_m is actually specified, there is no longer randomness over S_m, and the specified dataset S_m is "good" (high probability event) or "bad" (low probability event). We get a "good" dataset with high probability, and we obtain probabilistic guarantees such as Equation (1).

In many practical studies, a dataset to learn a model is specified first as, for example, in studies with CIFAR-10, ImageNet, or UCI datasets. Thus, we might have a statistically "bad" dataset S_m with no randomness over S_m when these practical studies begin. Even then, we can minimize the expected error in Theorem 1 by minimizing $V[f]$ (and/or $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ as deterministic quantity) such that $V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ becomes marginal without the randomness over S_m.

A.4 Supplementary explanation in Remark 5

The bound is always minimized if $\hat{W} = W^*$, but it is not a necessary condition in a pathological case where the star-discrepancy D^* is zero and A_1 can be zero with $\hat{W} \neq W^*$.

In Section 4.1, the optimal solution to minimize the expected error $\mathbb{E}_s[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2]$ is attained at $\hat{W} = W^*$. To see this, we can expand the expected error as

$$\begin{aligned}
&\mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] \\
&= \mathbb{E}_x\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right] + \mathbb{E}_{x,\xi}\left[\frac{1}{2}\|\xi\|_2^2 + \xi^\top\left(W^*\phi(x) - \hat{W}\phi(x)\right)\right] \\
&= \mathbb{E}_x\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right] + \mathbb{E}_\xi\left[\frac{1}{2}\|\xi\|_2^2\right],
\end{aligned}$$

where the last line follows that ξ is a zero-mean random variable independent of x. From the last line of the above equation, we can conclude the above statement about the minimizer.

A.5 Flatness in measures

It has been empirically observed that deep networks (particularly in the unsupervised setting) tend to transform the data distribution into a flatter one closer to a uniform distribution in a space of a learned representation (e.g., see Bengio et al. 2013). If the distribution $\mathcal{T}_*\mu$ with the learned representation $\mathcal{T}$ is uniform, then there exist better bounds on $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$ such as $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] \leq 10\sqrt{d/m'}$ (Aistleitner, 2011). Intuitively, if the measure $T_*\mu$ is non-flat and concentrated near a highly curved manifold, then there are more opportunities for a greater mismatch between $\mathcal{T}_*\mu$ and $T(Z_{m'})$ to increase $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$ (see below for pathological cases). This intuitively suggests the benefit of the flattening property that is sometimes observed with deep representation learning: it is often illustrated with generative models or auto-encoders by showing how interpolating between the representations of two images (in representation space) corresponds (when projected in image space) to other images that are plausible (are on or near the manifold of natural images), rather than to the simple addition of two natural images (Bengio et al., 2009b).

If $\mathcal{T}_*\mu$ is concentrated in a single point, then $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] = 0$, but it implies that there is only a single value of $L\hat{y}_{\mathcal{A}(S_m)}(z) = f(\phi(x), v)$ because $(\phi(x), v)$ takes only one value. Hence, this is tight and consistent. On the other hand, to minimize the empirical error $\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$ with diverse label values, $T_*\mu$ should not concentrate on the small number of finite points.

If $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$ is small, it means that the learned representation is effective at minimizing the generalization gap. This insight can be practically exploited by aiming to make $\mathcal{T}_*\mu$ flatter and spread out the data points $\mathcal{T}(Z_{m'})$ in a limited volume. It would also be beneficial to directly regularize an approximated $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$ with the unknown μ replaced by some known measures (e.g., a finite-support measure corresponding to a validation dataset).

A.6 Effect of higher-order derivatives

Example 1 suggests a method of regularization or model selection to control higher-order derivatives of a learned model w.r.t. a learned representation. Let $f(t) = \ell(\hat{Y}(t), Y(t))$; here, $\hat{Y}$ and Y represent the learned model $\hat{y}_{\mathcal{A}(S_m)}$ and the target output y as a function of $t = \mathcal{T}(z)$, respectively. Then, for example, if ℓ is the square loss, and if $\hat{Y}$ and Y are smooth functions, $V[f]$ goes to zero as $\nabla^k\hat{Y} - \nabla^k Y \to 0$ for $k = 1, 2, ...$, which can be upper bounded by $\|\nabla^k\hat{Y}\| + \|\nabla^k Y\|$.

Appendix B. Proofs

We use the following fact in our proof.

Lemma 1. (theorem 3.1 in Aistleitner et al. 2017) *Every real-valued function f on $[0,1]^d$ such that $V[f] < \infty$ is Borel measurable.*

B.1 Proof of Proposition 1

Proof. By the definition, we have that

$$\Delta^P_{j_1,\dots,j_k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k}) = \Delta^P_{j_1,\dots,j_{k-1}}\left(\Delta^P_{j_k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k})\right)$$

By the mean value theorem on the single variable t_{j_k},

$$\Delta^P_{j_k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k}) = \left(\partial_k f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,c^{(i_k)}_{j_k})\right)(t^{(i_k+1)}_{j_k} - t^{(i_k)}_{j_k}),$$

where $c^{(i_k)}_{j_k} \in (t^{(i_k+1)}_{j_k}, t^{(i_k)}_{j_k})$. Thus, by repeatedly applying the mean value theorem,

$$\Delta^P_{j_k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k}) = \left(\partial^k_{1,\dots,k} f_{j_1\dots j_k}(c^{(i_1)}_{j_1},\dots,c^{(i_k)}_{j_k})\right)\prod_{l=1}^{k}(t^{(i_k+1)}_{j_k} - t^{(i_k)}_{j_k}),$$

where $c^{(i_l)}_{j_l} \in (t^{(i_l+1)}_{j_l}, t^{(i_l)}_{j_l})$ for all $l \in \{1,\dots,k\}$. Thus,

$$V^{(k)}[f_{j_1\dots j_k}] = \sup_{P\in\mathcal{P}_k}\sum_{i_1=1}^{m^P_1-1}\cdots\sum_{i_k=1}^{m^P_k-1}\left|\partial^k_{1,\dots,k} f_{j_1\dots j_k}(c^{(i_1)}_{j_1},\dots,c^{(i_k)}_{j_k})\right|\prod_{l=1}^{k}(t^{(i_k+1)}_{j_k} - t^{(i_k)}_{j_k}).$$

By taking supremum for $\left|\partial^k_{1,\dots,k} f_{j_1\dots j_k}(c^{(i_1)}_{j_1},\dots,c^{(i_k)}_{j_k})\right|$ and taking it out from the sum, we obtain the first statement. The second statement follows the fact that if $\partial^k_{1,\dots,k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k})$ is continuous, then $|\partial^k_{1,\dots,k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k})|$ is continuous and Riemann integrable. Thus, the right hand side on the above equation coincides with the definition of the Riemann integral of $|\partial^k_{1,\dots,k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k})|$ over $[0,1]^k$. □

B.2 Proof of Theorem 1

The proof of Theorem 1 relies on several existing proofs from different fields. Accordingly, along the proof, we also track the extra dependencies and structures that appear only in machine learning, to confirm the applicability of the previous proofs in the problem of machine learning. Let $\mathbb{1}_A$ be an indicator function of a set A. Let $\Omega = [0,1]^d$. Let $\mathbf{1} = (1,1,\dots,1) \in \Omega$ and $\mathbf{0} = (0,0,\dots,0) \in \Omega$ as in a standard convention. The following lemma follows theorem 1.6.12 in (Ash and Doleans-Dade, 2000).

Lemma 2. *For any* $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$,

$$\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) = \int_{\Omega} f(\omega)d(\mathcal{T}_*\mu)(\omega).$$

Proof of Lemma 2. By Lemma 1, f is a Borel measurable function. The rest of the proof of this lemma directly follows the proof of theorem 1.6.12 in (Ash and Doleans-Dade, 2000); we proceed from simpler cases to more general cases as follows. In the case of f

being an indicator function of some set A as $f = \mathbb{1}_A$, we have that

$$\begin{aligned}\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) &= \mu(\mathcal{Z} \cap \mathcal{T}^{-1}A) \\ &= (\mathcal{T}_*\mu)(\Omega \cap A) \\ &= \int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega).\end{aligned}$$

In the case of f being a non-negative simple function as $f = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$,

$$\begin{aligned}\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) &= \sum_{i=1}^n \alpha_i \int_{\mathcal{Z}} \mathbb{1}_{A_i}(\mathcal{T}(z))d\mu(z) \\ &= \sum_{i=1}^n \alpha_i \int_\Omega \mathbb{1}_{A_i}(\omega)d(\mathcal{T}_*\mu)(\omega) \\ &= \int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega),\end{aligned}$$

where the second line follows what we have proved for the case of f being an indicator function.

In the case of f being a non-negative Borel measurable function, let $(f_k)_{k\in\mathbb{N}}$ be an increasing sequence of simple functions such that $f(\omega) = \lim_{k\to\infty} f_k(\omega)$, $\omega \in \Omega$. Then, by what we have proved for simple functions, we have $\int_{\mathcal{Z}} f_k(\mathcal{T}(z))d\mu(z) = \int_\Omega f_k(\omega)d(\mathcal{T}_*\mu)(\omega)$. Then, by the monotone convergence theorem, we have $\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) = \int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega)$.

In the case of $f = f^+ - f^-$ being an arbitrary Borel measurable function, we have already proved the desired statement for each f^+ and f^-, and by the definition of Lebesgue integration, the statement for f holds.

□

Proof of Theorem 1. With Lemmas 1 and 2, the proof follows that of theorem 1 in (Aistleitner and Dick, 2015). For any $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$,

$$\begin{aligned}\int_{\mathcal{Z}} L\hat{y}_{\mathcal{A}(S_m)}(z)d\mu(z) - \frac{1}{m'}\sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z_i) &= \int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) - \frac{1}{m'}\sum_{i=1}^{m'} f(\mathcal{T}(z_i)) \\ &= \int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{m'}\sum_{i=1}^{m'} f(\mathcal{T}(z_i))\end{aligned}$$

where the second line follows the condition of $\mathcal{T}$ and f and the third line follows Lemma 2. In the following, we first consider the case where f is left-continuous, and then discard the left-continuity condition later.

Consider the case where f is left-continuous (for the second statement): Suppose that f is left-continuous coordinate-wise at every point in the domain. Given a pair of vectors (a, b), we write $a \leq b$ if the relation holds for every coordinate. Let $\tilde{f}(\omega) = f(\mathbf{1} - \omega) - f(\mathbf{1})$ for all $\omega \in \Omega$. Then, by theorem 3 and equation (20) in (Aistleitner and Dick, 2015), there exists signed Borel measure $\mu_{\tilde{f}}$ on Ω such that $\tilde{f}(\omega) = \mu_{\tilde{f}}([\mathbf{0}, \omega])$ for all $\omega \in \Omega$ and $|\mu_{\tilde{f}}|(\Omega) = V[f] + |\tilde{f}(0)| = V[f]$. Let ν_f be the reflected measure of $\mu_{\tilde{f}}$ as $\nu_f(A) = \mu_{\tilde{f}}(\mathbf{1} - A)$

for any Borel set $A \subset \Omega$ where $\mathbf{1} - A = \{\mathbf{1} - t : t \in A\}$. It follows that ν_f is a signed Borel measure and

$$|\nu_f|(\Omega) = |\mu_{\tilde{f}}|(\Omega) = V[f].$$

By using these, we can rewrite f as

$$\begin{aligned} f(\omega) &= f(\mathbf{1}) + \tilde{f}(\mathbf{1} - \omega) \\ &= f(\mathbf{1}) + \int_\Omega \mathbb{1}_{[\mathbf{0}, \mathbf{1}-\omega]}(t) d\mu_{\tilde{f}}(t) \\ &= f(\mathbf{1}) + \int_\Omega \mathbb{1}_{[\omega, \mathbf{1}]}(t) d\nu_f(t) \\ &= f(\mathbf{1}) + \int_\Omega \mathbb{1}_{[\mathbf{0}, t]}(\omega) d\nu_f(t), \end{aligned}$$

where the second line follows from $\{\mathbf{1} - t : t \in [\omega, \mathbf{1}]\} = [\mathbf{0}, \mathbf{1} - \omega]$. Then, by linearity,

$$\frac{1}{m'} \sum_{i=1}^{m'} f(\mathcal{T}(z_i)) - f(\mathbf{1}) = \int_\Omega \frac{1}{m'} \sum_{i=1}^{m'} \mathbb{1}_{[\mathbf{0}, t]}(\mathcal{T}(z_i)) d\nu_f(t),$$

and by the Fubini–Tonelli theorem and linearity,

$$\begin{aligned} \int_\Omega f(\omega) d(\mathcal{T}_* \mu)(\omega) - f(\mathbf{1}) &= \int_\Omega \int_\Omega \mathbb{1}_{[\mathbf{0}, t]}(\omega) d(\mathcal{T}_* \mu)(\omega) d\nu_f(t) \\ &= \int_\Omega (\mathcal{T}_* \mu)([\mathbf{0}, t]) d\nu_f(t). \end{aligned}$$

Therefore,

$$\int_\Omega f(\omega) d(\mathcal{T}_* \mu)(\omega) - \frac{1}{m'} \sum_{i=1}^{m'} f(\mathcal{T}(z_i)) = \int_\Omega \left((\mathcal{T}_* \mu)([\mathbf{0}, t]) - \frac{1}{m'} \sum_{i=1}^{m'} \mathbb{1}_{[\mathbf{0}, t]}(\mathcal{T}(z_i)) \right) d\nu_f(t),$$

which proves the second statement of this theorem by noticing that $f(t) = \nu_f([t, \mathbf{1}]) + f(\mathbf{1})$. Moreover, this implies that

$$\begin{aligned} \left| \int_\Omega f(\omega) d(\mathcal{T}_* \mu)(\omega) - \frac{1}{m'} \sum_{i=1}^{m'} f(\mathcal{T}(z_i)) \right| &\leq |d\nu_f(t)|(\Omega) D^*[\mathcal{T}_* \mu, \mathcal{T}(Z_{m'})] \\ &= V[f] D^*[\mathcal{T}_* \mu, \mathcal{T}(Z_{m'})]. \end{aligned}$$

Discard the left-continuity condition of f (for the first statement): Let f be given and fixed without left-continuity condition. For each fixed f, by the law of large numbers (strong law of large numbers and the multidimensional Glivenko–Cantelli theorem), for any $\epsilon > 0$, there exists a number n and a set $\bar{A}_n = \{\bar{\omega}_i\}_{i=1}^n$ such that both of the following two inequalities hold:

$$\left| \int_\Omega f(\omega) d(\mathcal{T}_* \mu)(\omega) - \frac{1}{n} \sum_{i=1}^{n} f(\bar{\omega}_i) \right| \leq \epsilon,$$

and

$$D^*[\mathcal{T}_*\mu, \bar{A}_n] \leq \epsilon.$$

Let $\bar{A}_n = \{\bar{\omega}_i\}_{i=1}$ be such a set. For each fixed f, let f_n be a left-continuous function such that $f_n(\omega) = f(\omega)$ for all $\omega \in \bar{A}_n \cup \mathcal{T}(Z_{m'})$ and $V[f_n] \leq V[f]$. This definition of f_n is non-vacuous and we can construct such a f_n as follows. Let $\mathcal{G}$ be the d-dimensional grid generated by the set $\{\mathbf{0}\} \cup \{\mathbf{1}\} \cup \bar{A}_n \cup \mathcal{T}(Z_{m'})$; $\mathcal{G}$ is the set of all points $\omega \in \Omega$ such that for $k \in \{1, \ldots, d\}$, the k-th coordinate value of ω is the k-th coordinate value of some element in the set $\{\mathbf{0}\} \cup \{\mathbf{1}\} \cup \bar{A}_n \cup \mathcal{T}(Z_{m'})$. We can construct a desired f_n by setting $f_n(\omega) = f(\mathrm{succ}_n(\omega))$, where $\mathrm{succ}_n(\omega)$ outputs an unique element $t \in \mathcal{G}$ satisfying the condition that $t \geq \omega$ and $t \leq t'$ for all $t' \in \{t' \in \mathcal{G} : t' \geq \omega\}$.

Then, by triangle inequality, we write

$$\begin{aligned}\left|\int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{m'}\sum_{i=1}^{m'}\underbrace{f(\mathcal{T}(z_i))}_{=f_n(\mathcal{T}(z_i))}\right| \leq& \left|\int_\Omega f_n(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{m'}\sum_{i=1}^{m'} f_n(\mathcal{T}(z_i))\right| \\ &+ \left|\frac{1}{n}\sum_{i=1}^{n}\underbrace{f_n(\bar{\omega}_i)}_{=f(\bar{\omega}_i)} - \int_\Omega f_n(\omega)d(\mathcal{T}_*\mu)(\omega)\right| \\ &+ \left|\int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{n}\sum_{i=1}^{n} f(\bar{\omega}_i)\right|.\end{aligned}$$

Because f_n is left-continuous, we can apply our previous result to the first and the second terms; the first term is at most $V[f_n]D^*[\mathcal{T}_*\mu, T(Z_{m'})] \leq V[f]D^*[\mathcal{T}_*\mu, T(Z_{m'})]$, and the second term is at most $V[f_n]D^*[\mathcal{T}_*\mu, \bar{A}_n] \leq \epsilon V[f]$. The third term is at most ϵ by the definition of $\bar{A}_n$. Since $\epsilon > 0$ can be arbitrarily small, we have that for each $(f, \mathcal{T}) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$, (deterministically,)

$$\left|\int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{m'}\sum_{i=1}^{m'} f(\mathcal{T}(z_i))\right| \leq V[f]D^*[\mathcal{T}_*\mu, T(Z_{m'})].$$

Putting together: for any $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$,

$$\left|\int_\mathcal{Z} L\hat{y}_{\mathcal{A}(S_m)}(z)d\mu(z) - \frac{1}{m'}\sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z_i)\right| \leq V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$$

Thus, $\left|\int_\mathcal{Z} L\hat{y}_{\mathcal{A}(S_m)}(z)d\mu(z) - \frac{1}{m'}\sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z_i)\right|$ is a lower bound of a set $Q = \{V[f]$ $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] : (\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]\}$. By the definition of infimum, $|\int_\mathcal{Z} L\hat{y}_{\mathcal{A}(S_m)}(z)d\mu(z)$ $- \frac{1}{m'}\sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z_i)| \leq \inf Q$, if $\inf Q$ exists. Because Q is a nonempty subset of real and lower bounded by 0, $\inf Q$ exists. Therefore,

$$\left|\int_\mathcal{Z} L\hat{y}_{\mathcal{A}(S_m)}(z)d\mu(z) - \frac{1}{m'}\sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z_i)\right| \leq \inf_{(\mathcal{T},f)\in\mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]} V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})],$$

which implies the first statement of this theorem.

□

B.3 Proof of Proposition 2

Proof. From theorem 2 in (Heinrich et al., 2001), there exists a positive constant c_1 such that for all $s \geq c_1\sqrt{d}$ and for all $m' \in \mathbb{N}^+$,

$$\mathbb{P}\left\{D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] \geq sm'^{-1/2}\right\} \leq \frac{1}{s}\left(\frac{c_1 s^2}{d}\right)^d e^{-2s^2},$$

where we used the fact that the VC dimension of the set of the axis-parallel boxes contained in $[0,1]^d$ with one vertex at the origin is d (e.g., see Dudley 1984). By setting $s = c_2\sqrt{d}$ for any $c_2 \geq c_1$, we obtain the desired result. □

B.4 Proof of Proposition 3

Proof. From theorem 1 in (Aistleitner and Dick, 2014), for any $m' \in \mathbb{N}^+$, there exists a set $T_{m'}$ of points $t_1, \ldots, t_{m'} \in [0,1]^d$ such that

$$D^*[\mathcal{T}_*\mu, T_{m'}] \leq 63\sqrt{d}\frac{(2+\log_2 m')^{(3d+1)/2}}{m'}.$$

Because $\mathcal{T}$ is a surjection, for such a $T_{m'}$, there exists $Z_{m'}$ such that $\mathcal{T}(Z_{m'}) = T_{m'}$. □

B.5 Proof of the inequality in Example 2

Let $\mu_{\mathcal{T}(Z_{m'})}$ be a (empirical) normalized measure with the finite support on $\mathcal{T}(Z_{m'})$. Then,

$$\begin{aligned}
\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] &\leq V[f]D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] \\
&= \max\{|(\mathcal{T}_*\mu)(\{0\}) - \mu_{\mathcal{T}(Z_{m'})}(\{0\})|, \\
&\qquad |(\mathcal{T}_*\mu(\{0,1\})) - \mu_{\mathcal{T}(Z_{m'})}(\{0,1\})|\} \\
&= |\mathcal{T}_*\mu(\{0\}) - \mu_{\mathcal{T}(Z_{m'})}(\{0\})| \\
&= |1 - \mathcal{T}_*\mu(\{1\}) - 1 + \mu_{\mathcal{T}(Z_{m'})}(\{1\})| \\
&= |\mathcal{T}_*\mu(\{1\}) - \mu_{\mathcal{T}(Z_{m'})}(\{1\})|.
\end{aligned}$$

Rewriting $\mu_{\mathcal{T}(Z_{m'})}(\{1\}) = \mathbb{E}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$ yields the desired inequality in Example 2.

B.6 Proof of Theorem 2

Proof. Let $L\hat{y}_{\mathcal{A}(S_m)}(x) = \frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2$ ($\mathcal{Z} = \mathcal{X}$). Since

$$\frac{1}{2}\|W\phi(x) - y\|_2^2 = \frac{1}{2}\|W\phi(x) - W^*\phi(x)\|_2^2 + \frac{1}{2}\|\xi\|_2^2 - \xi^\top\left(W\phi(x) - W^*\phi(x)\right),$$

we have

$$\begin{aligned}
&\mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] - \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] \\
&= \mathbb{E}_{\mu_x}[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{X_m}[L\hat{y}_{\mathcal{A}(S_m)}] + A_1 + A_2 \\
&\leq V[f]D^*[\phi_*\mu_x, \phi(X_m)] + A_1 + A_2,
\end{aligned}$$

where the last line is obtained by applying Theorem 1 to $\mathbb{E}_{\mu_x}[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{X_m}[L\hat{y}_{\mathcal{A}(S_m)}]$ as follows. Let $\mathcal{T}(x) = \phi(x)$ and $f(t) = \frac{1}{2}\|\hat{W}t - W^*t\|_2^2$, where $t \in \mathbb{R}^{d_\phi}$. Then, $L\hat{y}_{A(S_m)}(x) = (f \circ \mathcal{T})(x)$, and $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$ in Theorem 1 if $V[f] < \infty$. Therefore, by Theorem 1, if $V[f] < \infty$,

$$\mathbb{E}_{\mu_x}[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{X_m}[L\hat{y}_{\mathcal{A}(S_m)}] \leq V[f]D^*[\phi_*\mu_x, \phi(X_m)].$$

To upper bound $V[f]$ and to show $V[f] < \infty$, we invoke Proposition 1 as follows. We have that $\frac{\partial f}{\partial t_l} = (\hat{W}_l - W_l^*)^\top(\hat{W} - W^*)t$, and $\frac{\partial f}{\partial t_l \partial t_{l'}} = (\hat{W}_l - W_l^*)^\top(\hat{W}_{l'} - W_{l'}^*)$. Because the second derivatives are constant over t, the third and higher derivatives are zeros. Let $\tilde{t}_l = (t_1, \ldots, t_{d_\phi})^\top$ with $t_j \equiv 1$ for all $j \neq l$. Then, we have that

$$\begin{aligned}
\sum_{l=1}^{d} V^{(1)}[f_l] &= \sum_{l=1}^{d} \int_{[0,1]} |(\hat{W}_l - W_l^*)^\top(\hat{W} - W^*)\tilde{t}_l| dt_l \\
&\leq \sum_{l=1}^{d} \|(\hat{W}_l - W_l^*)^\top(\hat{W} - W^*)\|_1 \int_{[0,1]} \|\tilde{t}_l\|_\infty dt_l. \\
&= \sum_{l=1}^{d} \|(\hat{W}_l - W_l^*)^\top(\hat{W} - W^*)\|_1,
\end{aligned}$$

and

$$\sum_{1 \leq l < l' \leq d} V^{(2)}[f_{ll'}] \leq \sum_{1 \leq l < l' \leq d} |(\hat{W}_l - W_l^*)^\top(\hat{W}_{l'} - W_{l'}^*)|.$$

Since higher derivatives exist and are zeros, from Proposition 1, $V^{(k)}[f_{j_1 \ldots j_k}] = 0$ for $k = 3, \ldots, d$. By the definition of $V[f]$, we obtain the desired bound for $V[f]$, and we have $V[f] < \infty$ if $\|\hat{W} - W^*\| < \infty$ (where there is no need to specify the particular matrix norm because of the equivalence of the norm). □

B.7 Proof of Theorem 3

Proof. Let $W_{l'l}$ be the (l', l)-th entry of the matrix W. Let $L\hat{y}_{\mathcal{A}(S_m)}(s) = \frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2$ ($\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$). Let $\mathcal{T}(s) = (\phi(x), y)$ and $f(t, y) = \frac{1}{2}\|\hat{W}t - y\|_2^2$. Then, $\ell(s) = (f \circ \mathcal{T})(s)$, and $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$ in Theorem 1 if $V[f] < \infty$. Therefore, by Theorem 1, if $V[f] < \infty$,

$$\mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] - \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] \leq V[f]D^*[\mathcal{T}_*\mu_s, \mathcal{T}(S_m)].$$

To upper bound $V[f]$ and to show $V[f] < \infty$, we invoke Proposition 1 as follows. For the first derivatives, we have that $\frac{\partial f}{\partial t_l} = \hat{W}_l^\top(\hat{W}t - y)$ and $\frac{\partial f}{\partial y_l} = -(\hat{W}t - y)_l$. For the second derivatives, we have that $\frac{\partial^2 f}{\partial t_l \partial t_{l'}} = \hat{W}_l^\top \hat{W}_{l'}$,

$$\frac{\partial^2 f}{\partial y_l \partial y_{l'}} = \begin{cases} 1 & \text{if } l = l' \\ 0 & \text{if } l \neq l, \end{cases}$$

and $\frac{\partial^2 f}{\partial t_l \partial y_{l'}} = -\hat{W}_{l'l}$. Because the second derivatives are constant in t and y, the third and higher derivatives are zeros. Then, because $|\frac{\partial f}{\partial t_l}| \leq M \|\hat{W}_l\|_1$ and $|\frac{\partial f}{\partial y_l}| \leq M$, with $l = j_1$,

$$\sum_{j_1=1}^{d_\phi} V^{(1)}[f_{j_1}] \leq M \sum_{l=1}^{d_\phi} \|\hat{W}_l\|_1,$$

and

$$\sum_{j_1=d_\phi+1}^{d_\phi+d_y} V^{(1)}[f_{j_1}] \leq d_y M.$$

Furthermore, for $j_1, j_2 \in \{1, \ldots, d_\phi\}$, with $l = j_1$ and $l' = j_2$,

$$V^{(2)}[f_{j_1 j_2}] \leq |\hat{W}_l^\top \hat{W}_{l'}|.$$

For $j_1 \in \{1, \ldots, d_\phi\}$ and $j_2 \in \{d_\phi + 1, \ldots, d_\phi + d_y\}$, with $l = j_1$ and $l' = j_2 - d_\phi$,

$$V^{(2)}[f_{j_1 j_2}] \leq |\hat{W}_{l'l}|,$$

and for $j_1, j_2 \in \{d_\phi + 1, \ldots, d_\phi + d_y\}$,

$$V^{(2)}[f_{j_1 j_2}] \leq \begin{cases} 1 & \text{if } j_1 = j_2 \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned} \sum_{1 \leq j_1 < j_2 \leq d_\phi + d_y} V^{(2)}[f_{j_1 j_2}] &= \sum_{1 \leq l < l' \leq d_\phi} |\hat{W}_l^\top \hat{W}_{l'}| + \sum_{l=1}^{d_\phi} \sum_{l'=1}^{d_y} |\hat{W}_{l'l}| \\ &= \sum_{1 \leq l < l' \leq d_\phi} |\hat{W}_l^\top \hat{W}_{l'}| + \sum_{l=1}^{d_\phi} \|\hat{W}_l\|_1. \end{aligned}$$

Therefore,

$$\begin{aligned} V[f] &= \sum_{k=1}^{d_\phi + d_y} \sum_{1 \leq j_1 < \cdots < j_k \leq d_\phi + d_y} V^{(k)}[f_{j_1 \ldots j_k}] \\ &= \sum_{k=1}^{2} \sum_{1 \leq j_1 < \cdots < j_k \leq d_\phi + d_y} V^{(k)}[f_{j_1 \ldots j_k}] \\ &\leq (M+1) \sum_{l=1}^{d_\phi} \|\hat{W}_l\|_1 + \sum_{1 \leq l < l' \leq d_\phi} |\hat{W}_l^\top \hat{W}_{l'}| + d_y M. \end{aligned}$$

Here, we have $V[f] < \infty$ because $\|\hat{W}\| < \infty$ and $M < \infty$ (and the equivalence of the norm). □