Generalization in Machine Learning via Analytical Learning Theory

Kenji Kawaguchi
Massachusetts Institute of Technology

Yoshua Bengio
University of Montreal, CIFAR Fellow

Abstract

This paper introduces a novel measure-theoretic learning theory to analyze generalization behaviors of practical interest. The proposed learning theory has the following abilities: 1) to utilize the qualities of each *learned* representation on the path from raw inputs to outputs in representation learning, 2) to guarantee good generalization errors possibly with arbitrarily rich hypothesis spaces (e.g., arbitrarily large capacity and Rademacher complexity) and non-stable/non-robust learning algorithms, and 3) to clearly distinguish each individual problem instance from each other. Our generalization bounds are relative to a representation of the data, and hold true even if the representation is learned. We discuss several consequences of our results on deep learning, one-shot 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. Because of the differences in the assumptions and the objectives, the proposed learning theory is meant to be complementary to previous learning theory and is not designed to compete with it.

1. Introduction

Statistical learning theory provides tight and insightful 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 & 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 & Williamson, 2002), *empirical* Rademacher complexity of a hypothesis space (Koltchinskii & Panchenko, 2000; Bartlett et al., 2002), and the robustness of learning algorithm (Xu & 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, datasets 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 empirical 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 dataset to learn a model 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 hypothesis space and a 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 & 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 nature. A firm understanding of analytical learning theory might be conceptually challenging, as it 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, several applications in Sections 4-5, and additional discussions in Section 6.

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

$S_m = \{s^{(1)}, \ldots, s^{(m)}\}$ 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* $\triangleq \mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$, where the case of $Z_{m'} = S_m$ is of particular interest (e.g., in the empirical risk minimization). 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.

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.

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

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

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

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

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 $\partial_{1,\ldots,k} \triangleq \partial_1, \ldots, \partial_k$. 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 $i = 0, \ldots, m_l^P - 1$),

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

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 $\Delta_{1,\ldots,k}^P \triangleq \Delta_1^P \ldots \Delta_k^P$. 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 $l \notin \{j_1, j_2, \ldots j_k\}$. That is, $f_{j_1 \ldots j_k}$ is a function of $(t_{j_1}, \ldots, t_{j_k})$ with other original variables being fixed to be one. The *variation of* $f_{j_1 \ldots j_k}$ *on* $[0,1]^k$ *in the sense of Vitali* is defined as

$$\begin{aligned}&V^{(k)}[f_{j_1 \ldots j_k}]\\ &\triangleq \sup_{P \in \mathcal{P}_k} \sum_{i_1=1}^{m_1^P - 1} \cdots \sum_{i_k=1}^{m_k^P - 1} \left|\Delta_{1,\ldots,k}^P f_{j_1 \ldots j_k}(t_{j_1}^{(i_1)}, \ldots, t_{j_k}^{(i_k)})\right|,\end{aligned}$$

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

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

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* $f_{j_1 \ldots j_k}$ *is a function for which* $\partial_{1,\ldots,k} f_{j_1 \ldots j_k}$ *exists on* $[0,1]^k$. *Then,*

$$V^{(k)}[f_{j_1 \ldots j_k}] \leq \sup_{(t_{j_1}, \ldots, t_{j_k}) \in [0,1]^k} \left|\partial_{1,\ldots,k} f_{j_1 \ldots j_k}(t_{j_1}, \ldots, t_{j_k})\right|.$$

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

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

3. A basis of analytical learning theory

This study considers the problem of analyzing the generalization gap $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$, including the case of $Z_{m'} = S_m$. Whenever we write $Z_{m'}$, it is always including the case of $Z_{m'} = S_m$. With our notation, one can observe that the generalization gap is fully and deterministically specified by a tuple or a *problem instance* $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$, where we identify

Table 1. A simplified comparison, wherein GG denotes the generalization gap

	Statistical Learning Theory	Analytical Learning Theory
GG is characterized by	hypothesis space $\mathcal{H}$ or learning algorithm $\mathcal{A}$	a learned model $\hat{y}_{\mathcal{A}(S_m)}$
GG is decomposed via	statistics	measure theory
Statistical assumption	is required	can be additionally used
Relative advantage when	a (training) dataset S_m remains random	a (training) dataset S_m is specified
Relative advantage in	worst-case analysis	beyond worst-case analysis

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 in the theory of the generalization gap of the tuple $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$ is said to be *strongly instance-dependent* 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 $\bar{S}_m \neq Z_{m'}$.

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 & 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 & 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 & Williamson 2002 with contraposition). The data-dependent bounds with empirical Rademacher complexity (Koltchinskii & Panchenko, 2000; Bartlett et al., 2002) also depend on a non-singleton hypothesis space and its empirical Rademacher complexity. These bounds can be made more data-dependent by considering a sequence of hypothesis spaces instead, but such data-dependent bounds still contain the complexity of a non-singleton hypothesis space (and its ordering). Moreover, the definition of robustness itself depends on $\bar{S}_m$, which is not equal to S_m or $Z_{m'}$ (Xu & Mannor, 2012). Therefore, all of these data-dependent bounds are not strongly instance-dependent.

Analytical learning theory is designed to be strongly instance-dependent. To achieve this goal, unlike statistical learning theory, analytical learning theory directly analyzes the (deterministic) mathematical object $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$. Because of the difference in the settings and the aims, analytical learning theory is not designed to compete with previous learning theory results, which are based on a statistical viewpoint. Table 1 summarizes the major simplified differences between statistical learning theory and analytical learning theory, as the rest of this paper clarifies these differences. See Appendix A.2 for a graphical illustration of the difference.

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 study adopts the convention that the infimum of the empty set is positive infinity.

Theorem 1 is introduced 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, or even if the decomposition $(f \circ \mathcal{T})$ is unknown. 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 $Z_{m'} \neq S_m$. We can also set S_m to be the whole training-validation-test dataset

with $Z_{m'} = S_m$ or $Z_{m'}$ being any other dataset. While $Z_{m'} \subseteq \mathcal{Z}$ (to define the $\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$), a dataset S_m is arbitrary and $S_m \nsubseteq \mathcal{Z}$ is allowed when $Z_{m'} \neq S_m$.

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*

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

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 $|\nu_f|([0,1]^d) = V[f]$.

Theorem 1 holds 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 a $\mathcal{H}$ 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).

Remark 1. The bound and the equation in Theorem 1 are strongly instance-dependent, and in particular, *invariant to hypothesis spaces $\mathcal{H}$ and the properties of learning algorithm $\mathcal{A}$ over datasets different from a given dataset S_m* (and $Z_{m'}$) (e.g., stability and robustness). They are fully determined by each given instance $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$ without dependence on other instances.

Remark 2. Theorem 1 together with Remark 1 has a significant consequence in practice. 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 $\hat{y}_{\mathcal{A}(S_m)}$ in 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 & 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 penalizes the use of more complex "larger" hypothesis spaces (e.g., see Zhang et al. 2017).

Remark 3. The reason why both analytical learning theory and certain practical observations can possibly obtain results that might appear to contradict statistical learning theory is the difference in the settings or the sets of assumptions. The larger complexities of a hypothesis space and instability/nonrobustness (or other properties of the learning algorithm) indeed degrade the statistical guarantee over *different random problem instances* particularly with a worst-case distribution $\bar{\mu} \neq \mu$; however, it can improve the analytical guarantee and practical performances for *"good" problem instances* $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$.

For example, no free lunch theorems in machine learning are well-known results, typically proven with the worst-case distribution $\bar{\mu} \neq \mu$ for a fixed $\mathcal{A}$ or $\mathcal{H}$ over random different instances, rather than each individual instance $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$.

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. Proposition 2 is a direct consequence of (Heinrich et al., 2001, Theorem 2).

Proposition 2. *Let $\mathcal{T}(Z_{m'}) = \{\mathcal{T}(z^{(1)}), \ldots, \mathcal{T}(z^{(m')})\} = \{t^{(1)}, \ldots, t^{(m')}\}$ 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 $Z_{m'} = S_m$. 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,

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

where the term $V[L\hat{y}_{\mathcal{A}(S_m)}]$ is strongly instance-dependent. Indeed, in this case, the whole bound in Equation (1) is strongly instance-dependent. See Appendix A.3 for a conceptual discussion of using a statistical assumption when $Z_{m'} = S_m$.

In Equation (1), it is unnecessary for m to approach infinity in order for the generalization gap to go to zero. This is because it is multiplied by $V[L\hat{y}_{\mathcal{A}(S_m)}]$, which is a quality of a learned model $\hat{y}_{\mathcal{A}(S_m)}$. This strongly supports the concept of *one-shot learning*, in contrast to traditional results.

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 that is asymptotically faster in terms of the dataset size m'. This is a direct consequence of (Aistleitner & 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*

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

4. Linear models possibly with rich hypothesis spaces, algorithmic instability and nonrobustness

Even in the classical setting of linear regression models, recent papers (Zhang et al. 2017, Section 5; Kawaguchi et al. 2017, Section 3; Poggio et al. 2017, Section 5) point out the need for further theoretical studies to better understand precisely what makes a learned model generalize well. Moreover, numerous related papers, even with their focus on deep learning, essentially ask the open question of understanding generalization with rich hypothesis spaces, algorithmic instability and nonrobustness (e.g., see Bartlett et al. 2017), which is applicable to linear models.

Theorem 1 with Remarks 1-3 answers the above open question abstractly for machine learning and deep learning in general. This section provides a more concrete answer for the case of linear models, which is a simple case that still captures the essence of the question.

Let $S_m = \{s^{(i)}\}_{i=1}^m$ be the training dataset of the input-target pairs as $s^{(i)} = (x^{(i)}, y^{(i)})$. 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, $\hat{W}$ is an output of the training process, $\hat{W} := \mathrm{minimize}_W \hat{\mathbb{E}}_{S_m}[\frac{1}{2}\|W\phi(x) - y\|_2^2]$. 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 structured 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 structured 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*

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

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. 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 $\hat{W} = W^*$. Therefore, the bound in Theorem 2 is tight in terms of both the minimizer and its value.

In contrast to the conventional wisdom,[1] Theorem 2 tightly concludes that all that matters is how close $\hat{W}$ is to W^* in the end of the learning process, producing a strongly instance-dependent bound. As in other important theories, although it becomes apparent in hindsight, Theorem 2 provides a crucial practical insight: we should make $\hat{W}$ closer to W^* at the end of the learning process, even if it increases the bound on the norm $\|W\|$ as well as the capacity and complexity of a hypothesis space, and even if it decreases the stability and robustness of the learning algorithm.

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 structured 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 open 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 Rademacher complexity) 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 the same Rademacher complexity and capacity) 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 the 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*

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

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 this case requires us to bound the norm of $\hat{W}$. This corresponds to the traditional wisdom from statistical learning theory, except that we do not require a pre-defined bound on $\|W\|$ over random problem instances; Theorem 3 is only sensitive to the actual value of $\|\hat{W}\|$ at the end of learning process with the given problem instance, producing a strongly instance-dependent bound. The generalization gap goes to zero as $D^*[\mathcal{T}_* \mu_s, \mathcal{T}(S_m)]$ approaches zero via certain statistical assumption as in statistical learning theory. We can guarantee it via Proposition 2 as follows: with high probability, $D^*[\mathcal{T}_* \mu_s, \mathcal{T}(S_m)] \leq O(\sqrt{(d_\phi + d_y)/m})$.

For linear models with over-parameterization, a recent impactful paper (Zhang et al. 2017, Section 5) poses open questions: "is it then possible to generalize with such a rich model class and no explicit regularization?" and "do all global minima generalize equally well?". Theorems 2 and 3 shed light on those questions with a mathematically tight reasoning for each problem instance (see Remark 5).

5. General results for machine/deep learning

In this section, we consider the applications of Theorem 1 to general models in machine learning including deep neural networks, in order to understand the qualitative nature of learning under the setting wherein the generalization gap is decomposed without statistical assumptions. Deriving more concrete statements for each specific model type (e.g., convolutional neural networks with ReLU units) is left to future work. We first state that the results from the previous section can be applied to deep learning.

Remark 7. *Theorems 2 and 3 hold true, 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

[1] A good description of the conventional wisdom is given in (Zhang et al., 2017; Bartlett et al., 2017).

the squared loss (square of output minus target). Then, the identical proofs of Theorems 2 and 3 work with the learned representation ϕ.

5.1. Theory with intermediate learned representation

An important question in representation learning such as deep learning is how to consider the quality of *learned* representations. The following example provides a general result to answer such a question with practical insights.

General Example 1. (Decomposition at a space of learned representation) 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 $Z_{m'} = S_m$),

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

General Example 1 partially supports the concept of the *disentanglement* in deep learning (Bengio et al., 2009b) and proposes a new concrete method to measure a degree of the 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: $V^{(k)}[f_{j_1 \ldots j_k}] = \int_{[0,1]^k} |\partial_{1,\ldots,k} f_{j_1 \ldots j_k}(t_{j_1}, \ldots, t_{j_k})| \, dt_{j_1} \cdots dt_{j_k}$, where $\partial_{1,\ldots,k} f_{j_1 \ldots j_k}(t_{j_1}, \ldots, t_{j_k})$ 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 $V[f] = \sum_{j_1=1}^{d} V^{(1)}[f_{j_1}]$.

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 Appendix A.5 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 $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$ is small, it means that the learned representation ϕ 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).

Furthermore, General Example 1 suggests a method of regularization or model selection to control higher 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 $y_{\mathcal{A}(S_m)}$ 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 $\|\nabla^k \hat{Y}\| + \|\nabla^k Y\|$.

5.2. Theory with raw space and loss space

This subsection applies Theorem 1 to a raw representation space $\mathcal{T}(z) = z$ and a loss space $\mathcal{T}(z) = (\iota \circ L\hat{y}_{\mathcal{A}(S_m)})(z)$, where $\iota : \{0, 1\} \to [0, 1]$ is an inclusion map.

General Example 2. (Decomposition in the space of $\mathcal{Z}$) 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 $L\hat{y}_{\mathcal{A}(S_m)}$,

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] \leq \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] + V[L\hat{y}_{\mathcal{A}(S_m)}]D^*[\mu, Z_{m'}].$$

This 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 $y_{\mathcal{A}(S_m)}$ 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 & Soatto, 2017).

General Example 3. (Decomposition in the loss space) Consider multi-class classification with 0-1 loss. 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.7 for this derivation), which establishes a tightness of Theorem 1 with the 0-1 loss as follows: for any dataset pair $(S_m, Z_{m'})$ (including $Z_{m'} = S_m$),

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

6. Discussion

By generating strongly instance-dependent bounds, throughout this paper, Theorem 1 has been shown to answer the following open questions: 1) how to mathematically analyze generalization behaviors of machine learning models possibly with arbitrarily rich hypothesis spaces (large capacity, Rademacher complexity, etc.) and non-stable/non-robust learning algorithms, and 2) how to theoretically distinguish generalization behaviors with structured labels from unstructured random labels. Perhaps more importantly, Theorem 1 presents a way to measure the quality of the learned model $y_{\mathcal{A}(S_m)}$ along with each instance $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$, and has produced several new theoretical insights on the generalization gap.

Theorem 1 also provides a different theoretical insight that has another immediate practical consequence on how to handle a dataset. In Theorem 1, the quality of a dataset S_m is determined by the *product* of its similarity to μ (the star-discrepancy) and the quality of the learned function $L\hat{y}_{\mathcal{A}(S_m)}$ through the dataset (the variation). That is, while conventional wisdom based on statistical learning theory tells us to collect a statistically "good" (e.g., i.i.d.) dataset from a distribution closer to the true μ, Theorem 1 tells us that there may be better choices. According to Theorem 1, collecting a dataset that can induce a better model $\hat{y}_{\mathcal{A}(S_m)}$ with a lower empirical error and lower variation of $L\hat{y}_{\mathcal{A}(S_m)}$ also results in better generalization gap and expected error, even if the similarity to μ is reduced and the statistical property (e.g., independence) becomes invalid (e.g., in Example 2, even if $D^*[\mu, S_m]$ approaches infinity, if $V[L\hat{y}_{\mathcal{A}(S_m)}]$ goes to zero faster, the generalization gap approaches zero). This is consistent with emerging practical heuristics in deep learning; it is becoming a common practice to add new data points to explicitly improve a currently learned model (e.g., by probing the model), rather than considering the statistical property in the dataset.

As we discussed in Section 3.2, Theorem 1 produces generalization bounds that can be zero even with $m = 1$ (and $m' = 1$), supporting the concept of one-shot learning. This is true in general, even if the dataset *completely* differs from the measure μ (e.g., learning with different distributions). This is because although it increases D^*, 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, particularly on the learned model $L\hat{y}_{\mathcal{A}(S_m)}$ in the end, Theorem 1 supports the concept of curriculum learning (Bengio et al., 2009a), which directly guides the learning to obtain a good model $y_{\mathcal{A}(S_m)}$ in the end. Moreover, consider certain types of curriculum learning that can violate statistical assumptions and degrade statistical guarantees. Theorem 1 supports even such types of curriculum learning, because learning a better model $y_{\mathcal{A}(S_m)}$ in the end decreases $V[f]$.

As we discussed in Section 3.2, $V[f]$ is always strongly instance-dependent, but a probabilistic bound on $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ may not always remain strongly instance-dependent, depending on the choice of $\mathcal{T}$. For example, if $\mathcal{T}$ is learned with S_m, we cannot directly adopt the probabilistic bound on $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ from Section 3.2, because $\mathcal{T}(S_m)$ does not satisfy the i.i.d. assumption. In such a case, one can derive new probabilistic bounds on D^* with learned representation $\mathcal{T}$ via the following standard statistical approach. 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 $\mathcal{T}' \in \Phi$, including the learned $\mathcal{T}$. As in data-dependent approaches (e.g., Shawe-Taylor et al. 1998), one can also consider a sequence of sets $\{\Phi_j\}_j$ such that $\mathcal{T} \in \cup_j \Phi_j$. However, in both cases, the bounds on $D^*[\mathcal{T}_*\mu, \mathcal{T}(S_m)]$ now depend on the set Φ (via its covering number) or the sequence $\{\Phi_j\}_j$ (via the ordering in j and a complexity of Φ_j) that depends on $\hat{y} \neq \hat{y}_{\mathcal{A}(S_m)}$, and hence are not strongly instance-dependent. This is a limitation of our result toward a complete learning theory with strong instance-dependence, and solving it is left to future work.

From a practical viewpoint, this might not be a major limitation, since a probabilistic bound on D^* is unnecessary for the convergence of the generalization gap. The convergence of $V[f]$ to zero (faster than the increase rate of D^*) is sufficient for the convergence of generalization gap, and D^* would also decrease deterministically. Moreover, one can simply measure the empirical error with another dataset $Z_{m'} \neq S_m$ (e.g., held-out validation dataset) to get $D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})] \leq O(\sqrt{d/m'})$ (with high probability), where the representation $\mathcal{T}$ is learned with S_m (since the i.i.d. condition can now be straightforwardly satisfied).

The fact that Theorem 1 is invariant to a hypothesis space $\mathcal{H}$ and certain details of a learning algorithm $\mathcal{A}$ can make it difficult to understand their effects. However, 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, our 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.

Acknowledgements

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 W911NF1410433, as well as support from NSERC, CIFAR and Canada Research Chairs. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Dziugaite, Gintare Karolina and Roy, Daniel M. 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.

Hardy, Godfrey H. 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.

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

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

Hestness, Joel, Narang, Sharan, Ardalani, Newsha, Diamos, Gregory, Jun, Heewoo, Kianinejad, Hassan, Patwary, Md, Ali, Mostofa, Yang, Yang, and Zhou, Yanqi. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

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

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

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

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

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

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

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

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

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

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

Mukherjee, Sayan, Niyogi, Partha, Poggio, Tomaso, and Rifkin, Ryan. 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.

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

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

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

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

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

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

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

Zahavy, Tom, Kang, Bingyi, Sivak, Alex, Feng, Jiashi, Xu, Huan, and Mannor, Shie. Ensemble robustness and generalization of stochastic deep learning algorithms. *arXiv preprint arXiv:1602.02389*, 2016.

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

Appendix

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

A. Additional explanations

Both statistical learning theory and analytical learning theory have relative advantages and disadvantages, because of the difference in the objectives and the sets of assumptions.

To recognize certain differences in analytical learning theory and statistical learning theory, it is good to remember the basics of the mathematical logics such as the difference in "$\forall x, \exists y$" v.s. "$\exists y, \forall x$". Typically in statistical learning theory, some upper bound holds over a fixed $\mathcal{H}$ or a fixed $\mathcal{A}$ with a certain property *with high probability over different datasets*. Usually in analytical learning theory, some upper bound holds *individually for each problem instance*.

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.

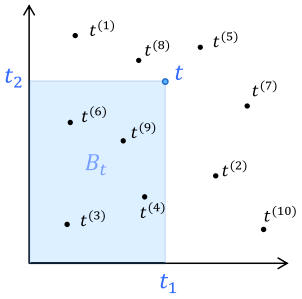
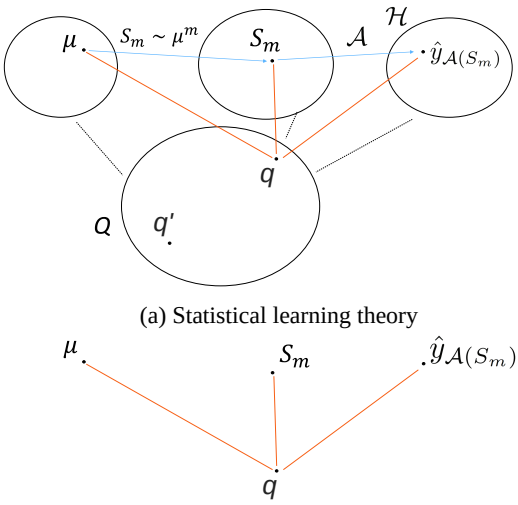


Figure 1. The local discrepancy $D[B_t; T_m, \nu]$ evaluates the difference between the empirical measure of the box B_t (the normalized number of data points in the box B_t, which is $4/10$) and the measure ν of the box B_t (the measure of the blue region)

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 others $q \neq q'$. Intuitively, we are typically analyzing the quality of the set Q, instead of each individual $q \in Q$.



(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 $\mathcal{A}$.

In this view, it becomes clear what is going on in some potentially surprising 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)) ignores the set Q, and 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 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 $q \in Q$. While proving such an assertion seems challenging in general (proving that a upper bound on $\forall q \in Q$ matches a lower bound $\exists q' \in Q$ is not sufficient), one can study it with empirical experiments (e.g., see Hestness et al. 2017).

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.

Several recent studies also consider a stochastic property of learning algorithms (e.g., Zahavy et al. 2016).

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

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

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. Pathological cases for non-flat measures

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.

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

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

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,

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

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,

$$V^{(k)}[f_{j_1\dots j_k}] = \sup_{P\in\mathcal{P}_k} \sum_{i_1=1}^{m_1^P-1}\cdots\sum_{i_k=1}^{m_k^P-1} \left|\partial_{1,\dots,k} f_{j_1\dots j_k}(c^{(i_1)}_{j_1},\dots,c^{(i_k)}_{j_k})\right| \cdot\prod_{l=1}^{k}(t^{(i_k+1)}_{j_k} - t^{(i_k)}_{j_k}).$$

By taking supremum for $\left|\partial_{1,\dots,k} f_{j_1\dots j_k}(c^{(i_1)}_{j_1},\dots,c^{(i_k)}_{j_k})\right|$ and taking it out from the sum, we obtain the first statement. The second statement follows the fact that if $\partial_{1,\dots,k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k})$ is continuous, then $|\partial_{1,\dots,k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k})|$ is continuous and Riemann integrable. Thus, the right hand side on the above equation coincides with the definition of the Riemann integral of $|\partial_{1,\dots,k} f_{j_1\dots j_k}(t^{(i_1)}_{j_1},\dots,t^{(i_k)}_{j_k})|$ 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. Thus, while presenting the proof is necessary for the purpose of this paper, each component of the proof by itself is *not* intended to be an original contribution. 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 & 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 & 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 & 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 & 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,

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

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

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

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, $\left| \int_\mathcal{Z} L\hat{y}_{\mathcal{A}(S_m)}(z) d\mu(z) - \frac{1}{m'} \sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z_i) \right| \leq \inf Q$, 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 $m' \in \mathbb{N}^+$,

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

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 & 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 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

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

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

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

Since higher derivatives exist and are zeros, from Proposition 1, $V^{(k)}[f_{j_1 \dots j_k}] = 0$ for $k = 3, \dots, d$. 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.6. 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 $j_1, j_2 \in \{1, \dots, d_\phi\}$, with $l = j_1$ and $l' = j_2$,

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

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

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

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

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

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,

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

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

B.7. Proof of the inequality in General Example 3

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 General Example 3.