A Measure-Theoretic Approach to Generalization in Machine Learning

Eugene Han

Yale University

December 12, 2024

Abstract

In the paper *Towards Understanding Generalization via Analytical Learning Theory* [4], Kawaguchi et al. present "analytical learning theory" as a novel framework for understanding the generalization of machine learning models from a non-statistical perspective that is problem instance-dependent rather than data-dependent. In our report, we explore foundational concepts from analytical learning theory, focusing on its implications for the generalization of machine learning models. Key results we will rigorously prove and flesh out include Proposition 1, which defines the variation of functions and establishes an upper bound on this variation through partial derivatives; Theorem 1, which introduces a decomposition of the generalization gap in terms of function variation and dataset discrepancy, offering instance-dependent bounds that enhance the understanding of model generalization beyond traditional statistical learning theory; and Theorem 2, which applies this framework to linear regression, providing tight bounds on the expected error while incorporating structured label assumptions. Collectively, these results underscore the utility of measure-theoretic approaches for analyzing and improving model performance in complex learning scenarios.

1 Introduction

Before presenting and rigorously proving the selected results from Kawaguchi et al.'s work, we will first provide some background on the problem of analyzing model generalization in the machine learning domain, ultimately motivating the introduction of "analytical learning theory." We will also briefly introduce and discuss the concepts of *discrepancy* and *variation*, borrowed from fields like "harmonic analysis, number theory, and numerical analysis" [4], which serve as building blocks for this new learning theory.

1.1 Generalization Gap

A common goal in machine learning problems is to learn some model given a training dataset that minimizes the expected error on unseen data. To formalize this, the authors introduce some notation for these ingredients:

- A model $\hat{y}_{\mathcal{A}(S_m)}$ returned by a learning algorithm $\mathcal{A}$ given a dataset $S_m = \{s^{(1)}, \ldots, s^{(m)}\}$
- The expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ with respect to a true unknown normalized measure μ and some function $L\hat{y}$ which couples a loss function l and a learned model $\hat{y}$

Since $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ requires knowledge of the true data generating process and access to μ, we often estimate it empirically using some subset of observed data $Z_{m'} = \{z^{(1)}, \ldots, z^{(m')}\}$ via

$$\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] = \frac{1}{m'} \sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z^{(i)})$$

For applied machine learning practitioners, $Z_{m'}$ could be our training set S_m or a held out validation/test set, in which $\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$ would represent the training and validation/test errors, respectively. Because we are interested in building models that perform well out of sample, we seek to analyze how $\hat{y}_{\mathcal{A}(S_m)}$ generalizes to unseen data by studying the generalization gap

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$$

which quantifies the difference between the expected error and the empirical error over $Z_{m'}$.

1.2 Discrepancy of a Dataset

We now introduce the concept of *discrepancy* with respect to a dataset. Let $B_t = [0, t_1] \times \cdots \times [0, t_d]$ be a closed box in $[0,1]^d$ for some $t = (t_1, \ldots, t_d) \in [0,1]^d$, let $T_m = \{t^{(1)}, \ldots, t^{(m)}\}$ be a dataset, and let ν be a normalized Borel measure. We define the *local discrepancy* of the dataset T_m with respect to ν on B_t to be

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

which intuitively quantifies the difference between the empirical measure of B_t and the measure ν of B_t. Then, the *star-discrepancy* of T_m with respect to ν on B_t is defined as

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

and quantifies how effectively the dataset T_m represents the measure ν.

1.3 Variation of a Function

Lastly, we introduce the concept of *variation* with respect to a function. Let P be a partition of $[0,1]^k$. If P consists of subsets of sizes $m_1^P, \ldots, m_k^P$, it can be written as a set of sequences $t_l^{(0)}, t_l^{(1)}, \ldots, t^{(m_l^P}$ with $0 = t_l^{(0)} \leq t_l^{(1)} \leq \cdots \leq t^{(m_l^P}$ for all $l \in [k]$. Define Δ_l^P to be the difference operator with respect to P such that for a function g and a point $(t_1, \ldots, t_{l-1}, t_l^{(i)}, t_{l+1}, \ldots, t_k)$ in P,

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

for $i = 0, 1, \ldots, m_l^P - 1$, and let $\Delta_{1,\ldots,k}^P = \Delta_1^P \cdots \Delta_k^P$ be a chain of difference operations. The authors also define $f_{j_1 \cdots j_k}$ to be the restriction of a function f with d variables on $k \leq d$ variables such that $f_{j_1 \cdots j_k}(t_{j_1}, \ldots, t_{j_k}) = f(t_1, \ldots, t_d)$ with $t_l = 1$ for all $l \notin \{j_1, j_2, \ldots, j_k\}$.

If $\mathcal{P}_k$ represents the set of all partitions of $[0,1]^k$, then we define the *Vitali variation* of $f_{j_1 \cdots j_k}$ on $[0,1]^k$ as

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

Based on this, we also define the *Hardy-Krause variation* or *total variation* of f on $[0,1]^d$ as

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

The authors write that the Hardy-Krause variation $V[f]$ "computes how a function f varies in total with respect to each small perturbation of every cross combination of its variables." [4]

Before we discuss the main result of the paper, Theorem 1, we will present the following proposition which relates the variation of a function to the partial derivatives of that function which will be useful when we discuss an application of Theorem 1 to linear regression. The authors denote ∂_l to be the partial derivative operator such that $\partial_l g(t_1, \ldots, t_k) = \frac{\partial g}{\partial x_l}|_{(x_1,\ldots,x_k)=(t_1,\ldots,t_k)}$ is the partial derivative of g with respect to the l-th coordinate evaluated at the point $(t_1, \ldots, t_k)$ and $\partial^k_{1,\ldots,k} = \partial_1 \cdots \partial_k$ to be a chain of partial derivative operations.

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

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

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

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

Proof of Proposition 1. [1] By definition of the difference operator, we have the recursive formulation

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

By definition, we also have

$$\frac{\Delta^P_{j_k} f_{j_1 \cdots j_k}(t^{(i_1)}_{j_1}, \ldots, t^{(i_k)}_{j_k})}{t^{(i_k+1)}_{j_k} - t^{(i_k)}_{j_k}} = \frac{f_{j_1 \cdots j_k}(t^{(i_1)}_{j_1}, \ldots, t^{(i_k+1)}_{j_k}) - f_{j_1 \cdots j_k}(t^{(i_1)}_{j_1}, \ldots, t^{(i_k)}_{j_k}))}{t^{(i_k+1)}_{j_k} - t^{(i_k)}_{j_k}}$$

Recall from single-variable calculus that the Mean Value Theorem states for some function g that is continuous on $[a,b]$ and differentiable on (a,b), there exists some $c \in (a,b)$ such that $g'(c) = \frac{g(b)-g(a)}{b-a}$. Focusing on just t_{j_k} and considering the partial derivative, by MVT there exists a $c^{(i_k)}_{j_k} \in (t^{(i_k)}_{j_k}, t^{(i_k+1)}_{j_k})$ such that

$$\frac{\Delta^P_{j_k} f_{j_1 \cdots j_k}(t^{(i_1)}_{j_1}, \ldots, t^{(i_k)}_{j_k})}{t^{(i_k+1)}_{j_k} - t^{(i_k)}_{j_k}} = \partial_k f_{j_1 \cdots j_k}(t^{(i_1)}_{j_1}, \ldots, c^{(i_k)}_{j_k})$$

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

Using the recursive formulation from the beginning of the proof and applying MVT repeatedly, considering $c^{(i_l)}_{j_l} \in (t^{(i_l)}_{j_l}, t^{(i_l+1)}_{j_l})$ for $l \in [k]$, we obtain

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

[1] There are a handful of typos with respect to variable and operator indices in the proof presented in the original paper. They have been corrected in this report.

Now, recall the definition of Vitali variation. Substituting our result from above yields

$$\begin{aligned} V^{(k)}[f_{j_1\cdots j_k}] &= \sup_{P\in\mathcal{P}_k} \sum_{i_1=1}^{m_1^P-1} \cdots \sum_{i_k=1}^{m_k^P-1} \left|\Delta^P_{j_1,\ldots,j_k} f_{j_1\cdots j_k}(t_{j_1}^{(i_1)},\ldots,t_{j_k}^{(i_k)})\right| \\ &= \sup_{P\in\mathcal{P}_k} \sum_{i_1=1}^{m_1^P-1} \cdots \sum_{i_k=1}^{m_k^P-1} \left|\partial_{1,\ldots,k} f_{j_1\cdots j_k}(c_{j_1}^{(i_1)},\ldots,c_{j_k}^{(i_k)})\right| \prod_{l=1}^{k}(t_{j_l}^{(i_l+1)} - t_{j_l}^{(i_l)}) \end{aligned}$$

To show the first result, we can upper bound $\left|\partial_{1,\ldots,k} f_{j_1\cdots j_k}(c_{j_1}^{(i_1)},\ldots,c_{j_k}^{(i_k)})\right|$ by taking the supremum since

$$\left|\partial_{1,\ldots,k} f_{j_1\cdots j_k}(c_{j_1}^{(i_1)},\ldots,c_{j_k}^{(i_k)})\right| \leq \sup_{(t_{j_1},\ldots,t_{j_k})\in[0,1]^k} |\partial_{1,\ldots,k} f_{j_1\cdots j_k}(t_{j_1},\ldots,t_{j_k})|$$

and then factoring it out of the summation in the expression for $V^{(k)}[f_{j_1\cdots j_k}]$ to get

$$V^{(k)}[f_{j_1\cdots j_k}] \leq \sup_{(t_{j_1},\ldots,t_{j_k})\in[0,1]^k} |\partial_{1,\ldots,k} f_{j_1\cdots j_k}(t_{j_1},\ldots,t_{j_k})| \cdot \sup_{P\in\mathcal{P}_k} \sum_{i_1=1}^{m_1^P-1} \cdots \sum_{i_k=1}^{m_k^P-1} 1 \prod_{l=1}^{k}(t_{j_l}^{(i_l+1)} - t_{j_l}^{(i_l)})$$

We recognize $\sum_{i_1=1}^{m_1^P-1} \cdots \sum_{i_k=1}^{m_k^P-1} 1 \prod_{l=1}^{k}(t_{j_l}^{(i_l+1)} - t_{j_l}^{(i_l)})$ to be the Riemann integral representing the volume of $[0,1]^k$ which is just equal to 1, so we conclude

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

To show the second result, we note that if $\partial^k_{1,\ldots,k} f_{j_1\cdots j_k}$ is continuous on $[0,1]^k$, then $\left|\partial^k_{1,\ldots,k} f_{j_1\cdots j_k}\right|$ is also continuous and consequently Riemann integrable. In a similar fashion to the first result, we recognize

$$\sup_{P\in\mathcal{P}_k} \sum_{i_1=1}^{m_1^P-1} \cdots \sum_{i_k=1}^{m_k^P-1} \left|\partial_{1,\ldots,k} f_{j_1\cdots j_k}(c_{j_1}^{(i_1)},\ldots,c_{j_k}^{(i_k)})\right| \prod_{l=1}^{k}(t_{j_l}^{(i_l+1)} - t_{j_l}^{(i_l)})$$

to be the Riemann integral of $\left|\partial^k_{1,\ldots,k} f_{j_1\cdots j_k}\right|$ over $[0,1]^k$. Therefore, we conclude that if $\partial^k_{1,\ldots,k} f_{j_1\cdots j_k}$ is continuous, then

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

□

2 Decomposition of Expected Error

We are now primed to study the expected error $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}]$ and in turn the generalization gap $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}]$. The instance-dependence of analytical learning theory comes from the authors' introduction of an object called a *problem instance* parameterized as $(\mu, S_m, Z_{m'}, L\hat{y}_{\mathcal{A}(S_m)})$, with the measure μ coming from some unknown measure space $(\mathcal{Z}, \Sigma, \mu)$, which fully characterizes the generalization gap. It follows that $(\mathcal{Z}, \Sigma, \mu)$ defines the expected error as the Lebesgue integral of $L\hat{y}_{\mathcal{A}(S_m)}$; that is,

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] = \int_{\mathcal{Z}} L\hat{y}_{\mathcal{A}(S_m)} d\mu$$

The authors also denote $\mathcal{T}_*\mu$ to be the pushforward measure of μ under a map $\mathcal{T}$ and $|\nu|(E)$ to be the total variation of a measure ν on E. For reference, given measurable spaces (X_1, Σ_1) and (X_2, Σ_2), a measurable mapping $f : X_1 \to X_2$, and a measure $\mu : \Sigma_1 \to [0, +\infty]$, the pushforward measure of μ, $f_*(\mu) : \Sigma_2 \to [0, +\infty]$, is defined as $f_*(\mu)(B) = \mu(f^{-1}(B))$ for $B \in \Sigma_2$. Also, $|\nu|(E) = \sup \sum_i |\nu(E_i)|$ where supremum is taken over all partitions $\cup E_i$ of E into measurable subsets E_i.

We now present the most important result of Kawaguchi et al.'s paper.

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) $\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] \leq \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] + \inf_{(\mathcal{T},f)\in\hat{\mathcal{F}}} V[f] \cdot D^*[\mathcal{T}_*\mu, \mathcal{T}(Z_{m'})]$, 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 left-continuous [2] 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 $|\nu_f|([0,1]^d) = V[f]$.

To show that Theorem 1 holds, we must first discuss several lemmas since the proof for Theorem 1 involves results from multiple different papers and fields. We won't prove most of these and will instead direct the reader to the corresponding literature; the purpose of presenting these intermediate results is to make each of their contributions to Theorem 1 clear and, if relevant, how they relate to one another.

Lemma 1.1 (Corollary 3 in [5]). *Any function of bounded Hardy-Krause variation can be written as the difference of two completely monotone functions.*

Lemma 1.2 (Theorem 3.2 in [2]). *Every completely monotone and real-valued function on $[0,1]^d$ is $([0,1]^d, \mathcal{B}([0,1]^d)) - (\mathbb{R}, \mathcal{B}(\mathbb{R}))$-measurable.*

Lemma 1.3 (Theorem 3.1 in [2]). *Every real-valued function f on $[0,1]^d$ such that $V[f] < \infty$ is Borel measurable.*

Proof of Lemma 1.3. By Lemma 1.1, we can write f as $f = g - h$, where g and h are both completely monotone functions on $[0,1]^d$, since $V[f] < \infty$. By Lemma 1.2, g and h are both Borel measurable. It then follows from the fact that $g - h$ is also Borel measurable that we can conclude f is Borel measurable.

□

Lemma 1.4 (Theorem 1.6.12 in [3]). *For any $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$,*

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

where $\Omega = [0,1]^d$ and $\omega \in \Omega$.

Proof of Lemma 1.4. Since $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$, by hypothesis f is a function on $[0,1]^d$ such that $V[f] < \infty$. By Lemma 1.3, f is Borel measurable. We now proceed by cases.

[2] There is a typo in the original paper that incorrectly enforces right-continuity for (ii).

Case 1: Suppose f is an indicator function for a set A, that is $f = \mathbb{1}_A$. Then, we have that

$$\begin{aligned}
\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) &= \mu(\mathcal{Z} \cap \mathcal{T}^{-1}(A)) && \text{integral will just be measure of indicator [6]} \\
&= (\mathcal{T}_*\mu)(\Omega \cap A) && \text{by definition of pushforward measure} \\
&= \int_{\Omega} f(\omega)d(\mathcal{T}_*\mu)(\omega)
\end{aligned}$$

as desired.

Case 2: Suppose f is a non-negative simple function, that is $f = \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}$. Then,

$$\begin{aligned}
\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) &= \int_{\mathcal{Z}} \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}(\mathcal{T}(z))d\mu(z) \\
&= \sum_{i=1}^n \alpha_i \int_{\mathcal{Z}} \mathbb{1}_{A_i}(\mathcal{T}(z))d\mu(z) \\
&= \sum_{i=1}^n \alpha_i \int_{\Omega} \mathbb{1}_{A_i}(\omega)d(\mathcal{T}_*\mu)(\omega) && \text{using result from Case 1} \\
&= \int_{\Omega} \sum_{i=1}^n \alpha_i \mathbb{1}_{A_i}(\omega)d(\mathcal{T}_*\mu)(\omega) \\
&= \int_{\Omega} f(\omega)d(\mathcal{T}_*\mu)(\omega) && \text{by definition of } f
\end{aligned}$$

as desired.

Case 3: Suppose f is any non-negative Borel measurable function. Let $\{f_k\}_{k\in\mathbb{N}}$ be an increasing sequence of simple functions such that they converge pointwise to f as for each $\omega \in \Omega$, $f(\omega) = \lim_{k\to\infty} f_k(\omega)$. By the result of Case 2, we know that for all k we have $\int_{\mathcal{Z}} f_k(\mathcal{T}(z))d\mu(z) = \int_{\Omega} f_k(\omega)d(\mathcal{T}_*\mu)(\omega)$.

Now, recall that Monotone Convergence Theorem states that if $\{f_n : X \to [0,\infty)\}$ is a sequence of measurable functions on a measurable set X such that $f_n \to f$ pointwise almost everywhere and $f_1 \leq f_2 \leq \cdots$, then $\lim_{n\to\infty} \int_X f_n = \int_X f$.

Applying this here yields

$$\begin{aligned}
\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) &= \lim_{k\to\infty} \int_{\mathcal{Z}} f_k(\mathcal{T}(z))d\mu(z) && \text{by MCT} \\
&= \lim_{k\to\infty} \int_{\Omega} f_k(\omega)d(\mathcal{T}_*\mu)(\omega) && \text{using result from Case 2} \\
&= \int_{\Omega} f(\omega)d(\mathcal{T}_*\mu)(\omega) && \text{by MCT}
\end{aligned}$$

as desired.

Case 4: Suppose f is any Borel measurable function. We can write f as $f = f^+ - f^-$ where f^+ and f^- are the positive and negative parts of f and are both non-negative Borel measurable functions. By the result of Case 3, we know that $\int_{\mathcal{Z}} f^+(\mathcal{T}(z))d\mu(z) = \int_{\Omega} f^+(\omega)d(\mathcal{T}_*\mu)(\omega)$ and $\int_{\mathcal{Z}} f^-(\mathcal{T}(z))d\mu(z) = \int_{\Omega} f^-(\omega)d(\mathcal{T}_*\mu)(\omega)$. Then, by definition of Lebesgue integration,

$$\begin{aligned}
\int_{\mathcal{Z}} f(\mathcal{T}(z))d\mu(z) &= \int_{\mathcal{Z}} (f^+ - f^-)(\mathcal{T}(z))d\mu(z) \\
&= \int_{\mathcal{Z}} f^+(\mathcal{T}(z))d\mu(z) - \int_{\mathcal{Z}} f^-(\mathcal{T}(z))d\mu(z)
\end{aligned}$$

$$\begin{aligned}
&= \int_\Omega f^+(\omega)d(\mathcal{T}_*\mu)(\omega) - \int_\Omega f^-(\omega)d(\mathcal{T}_*\mu)(\omega) \\
&= \int_\Omega (f^+ - f^-)(\omega)d(\mathcal{T}_*\mu)(\omega) \\
&= \int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega)
\end{aligned}$$

as desired.

□

Lemma 1.5 (Part (a) of Theorem 3 and Equation (20) in [1])**.** *Let f be a right-continuous function on $[0,1]^d$ which has bounded Hardy-Krause variation. Then there exists a unique signed Borel measure ν on $[0,1]^d$ for which*

$$f(\mathbf{x}) = \nu([\mathbf{0}, \mathbf{x}]), \quad \mathbf{x} \in [0,1]^d.$$

Then we have

$$\mathrm{Var}_{\mathrm{total}}\nu = \mathrm{Var}_{\mathrm{HK}\mathbf{0}}f + |f(\mathbf{0})|$$

where $\mathrm{Var}_{\mathrm{total}}\nu$ refers to the total variation of ν, $|\nu|$. By Equation (20), we also have

$$\mathrm{Var}_{\mathrm{HK}}f = \mathrm{Var}_{\mathrm{HK}\mathbf{0}}g$$

where $g(\mathbf{x}) = f(\mathbf{1} - \mathbf{x})$ for $\mathbf{x} \in [0,1]^d$ (acts as a mirroring of f). Here, $\mathrm{Var}_{\mathrm{HK}}f = V[f]$.

Lemma 1.6. *Let (X, Σ) be a measurable space. Let μ be a signed measure on (X, Σ). Let $|\mu|$ be variation of μ. Then, $|\mu(A)| \leq |\mu|(A)$ for each $A \in \Sigma$.*

Proof of Lemma 1.6. Let (μ^+, μ^-) be the Jordan decomposition of μ. Then, $\mu = \mu^+ - \mu^-$ and $|\mu| = \mu^+ + \mu^-$. Using the Triangle Inequality and the fact that both μ^+ and μ^- are non-negative, we have

$$\begin{aligned}
|\mu(A)| &= |\mu^+(A) - \mu^-(A)| \\
&\leq |\mu^+(A)| + |\mu^-(A)| \\
&= \mu^+(A) + \mu^-(A) \\
&= |\mu|(A)
\end{aligned}$$

□

Lemma 1.7 (Glivenko-Cantelli Theorem)**.** *Let X_i, $i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution function F on $\mathbb{R}$. Define the empirical distribution function $\hat{F}_n$ as $\hat{F}_n(x) = \frac{1}{n}\sum_{1 \leq i \leq n} \mathbb{1}_{\{X_i \leq x\}}$. Then,*

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \to 0 \ a.s.$$

We now prove Theorem 1.

Proof of Theorem 1. [3] Recall that

$$\mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] = \int_{\mathcal{Z}} L\hat{y}_{\mathcal{A}(S_m)}(z)d\mu(z)$$

and

$$\hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] = \frac{1}{m'}\sum_{i=1}^{m'} L\hat{y}_{\mathcal{A}(S_m)}(z^{(i)})$$

[3] The original proof for Theorem 1 has some inconsistent notation that conflicts with several definitions presented earlier in the paper. I've corrected these to the best of my ability to make it easier to follow along with the logic.

By hypothesis, for all $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$ we have $L\hat{y}_{\mathcal{A}(S_m)}(z) = (f \circ \mathcal{T})(z) = f(\mathcal{T}(z))$ for all $z \in \mathcal{Z}$. Using Lemma 1.4, we then have

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

We now prove the results for (i) and (ii). We will first prove (ii) by assuming f is left-continuous. We then drop this assumption for (i), but use the result from (ii) in an intermediate step.

Part (ii): Suppose f is left-continuous. Define $\tilde{f}$ such that $\tilde{f}(\omega) = f(\mathbf{1} - \omega) - f(\mathbf{1})$ for all $\omega \in \Omega$. Since f is left-continuous and has bounded Hardy-Krause variation, $\tilde{f}$ is right-continuous and also has bounded Hardy-Krause variation. By Lemma 1.5, there then exists a signed Borel measure $\mu_{\tilde{f}}$ on $[0,1]^d = \Omega$ such that $\tilde{f}(\omega) = \mu_{\tilde{f}}([\mathbf{0}, \omega])$ for all $\omega \in \Omega$ and $|\mu_{\tilde{f}}|(\Omega) = V[f] + |\tilde{f}(\mathbf{0})| = V[f]$ as $\tilde{f}(\mathbf{0}) = f(\mathbf{1} - \mathbf{0}) - f(\mathbf{1}) = 0$.

Now, let ν_f be another measure defined as $\nu_f(A) = \mu_{\tilde{f}}(\mathbf{1} - A)$ for any Borel set $A \subset \Omega$ such that $\mathbf{1} - A = \{\mathbf{1} - t : t \in A\}$. This is just a reflection of $\mu_{\tilde{f}}$ and is a signed Borel measure with $|\nu_f|(\Omega) = |\mu_{\tilde{f}}|(\Omega) = V[f]$.

We defined $\tilde{f}$ such that $\tilde{f}(\omega) = f(\mathbf{1} - \omega) - f(\mathbf{1})$, so equivalently we may write $f(\omega) = f(\mathbf{1}) + \tilde{f}(\mathbf{1} - \omega)$. Using the definition of ν_f and the fact that $\{\mathbf{1} - t : t \in [\omega, \mathbf{1}]\} = [\mathbf{0}, \mathbf{1} - \omega]$, we can write

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

Consider $\omega = \mathcal{T}(z^{(i)})$. Using the linearity of the Lebesgue integral, we can write

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

Now, consider taking the Lebesgue integral of $f(\omega)$ over Ω with respect to $\mathcal{T}_*\mu$. Using the Fubini-Tonelli Theorem, we can swap the order of integration from ν_f then $\mathcal{T}_*\mu$ to $\mathcal{T}_*\mu$ then ν_f to obtain

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

$$= f(\mathbf{1}) + \int_\Omega (\mathcal{T}_*\mu)([\mathbf{0},t])d\nu_f(t)$$

where the second line comes from the linearity of the Lebesgue integral. It then follows that

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

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

Sub-result: We have that

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

Notice that the integrand on the right side follows the definition of local discrepancy exactly, such that

$$\left|(\mathcal{T}_*\mu)([\mathbf{0},t]) - \frac{1}{m'}\sum_{i=1}^{m'} \mathbb{1}_{[\mathbf{0},t]}(\mathcal{T}(z^{(i)}))\right| = |D[[\mathbf{0},1];\mathcal{T}(Z_{m'}),\mathcal{T}_*\mu]|$$

and so we can bound it from above by taking supremum which is exactly the star-discrepancy $D^*[\mathcal{T}(Z_{m'}),\mathcal{T}_*\mu]$. So, by taking supremum and factoring $D^*[\mathcal{T}(Z_{m'}),\mathcal{T}_*\mu]$ we have

$$\begin{aligned}
\left|\int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{m'}\sum_{i=1}^{m'} f(\mathcal{T}(z^{(i)}))\right| &\le |\nu_f(\Omega)| \cdot D^*[\mathcal{T}(Z_{m'}),\mathcal{T}_*\mu] \\
&\le |\nu_f|(\Omega) \cdot D^*[\mathcal{T}(Z_{m'}),\mathcal{T}_*\mu] \qquad \text{by Lemma 1.6} \\
&= V[f] \cdot D^*[\mathcal{T}(Z_{m'}),\mathcal{T}_*\mu] \qquad |\nu_f|(\Omega) = |\mu_{\tilde{f}}|(\Omega) = V[f]
\end{aligned}$$

Part (i): Suppose we are given some fixed f. For each f by the Strong Law of Large Numbers and a multivariate extension of Lemma 1.7 (Glivenko-Cantelli), for any $\epsilon > 0$ there exists $n \in \mathbb{N}$ and a set $\bar{A}_n = \{\bar{\omega}_i\}_{i=1}^n$ such that $\left|\int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{n}\sum_{i=1}^n f(\bar{\omega}_i)\right| \le \epsilon$ and $D^*[\bar{A}_n,\mathcal{T}_*\mu] \le \epsilon$. Moreover, for each f we can define f_n to be a left-continuous function such that $f_n(\omega) = f(\omega)$ for all $\omega \in \bar{A}_n \cup \mathcal{T}(Z_{m'})$.

We now write $\int_\Omega f(\omega)d(\mathcal{T}_*\mu)(\omega) - \frac{1}{m'}\sum_{i=1}^{m'} f(\mathcal{T}(z^{(i)}))$ as follows:

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

Then, by Triangle Inequality we have

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

From earlier, we know the first term on the right side of the inequality

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

by definition of $\bar{A}_n$. Using the sub-result that we derived after our proof for part (ii), we know the second term is upper bounded as

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} f(\bar{\omega}_i) - \int_\Omega f_n(\omega)d(\mathcal{T}_*\mu)(\omega)\right| &\le V[f_n]\cdot D^*[\bar{A}_n, \mathcal{T}_*\mu] \\ &\le \epsilon V[f]\end{aligned}$$

Similarly, using the same sub-result for the third term give us

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

Together, we have

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

If we take ϵ to be arbitrarily small, then we can just write

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

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

Now, consider the set $Q = \left\{ V[f] \cdot D^*[\mathcal{T}(Z_{m'}), \mathcal{T}_*\mu] : (\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}] \right\}$. Clearly from above, we can state that $\left| \mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] \right|$ is a lower bound of Q. By definition, the infimum of a set is the greatest lower bound of that set, so it follows trivially that $\left| \mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] \right| \leq \inf Q$ must hold. If we use $\hat{\mathcal{F}}$ to denote $\mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$ as used in the original statement of the theorem, then we can finally conclude that

$$\begin{aligned}
\left| \mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] \right| &\leq \inf_{(\mathcal{T},f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]} V[f] \cdot D^*[\mathcal{T}(Z_{m'}), \mathcal{T}_*\mu] \\
&= \inf_{(\mathcal{T},f) \in \hat{\mathcal{F}}} V[f] \cdot D^*[\mathcal{T}(Z_{m'}), \mathcal{T}_*\mu] \\
\implies \mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] - \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] &\leq \inf_{(\mathcal{T},f) \in \hat{\mathcal{F}}} V[f] \cdot D^*[\mathcal{T}(Z_{m'}), \mathcal{T}_*\mu] \\
\implies \mathbb{E}_\mu[L\hat{y}_{\mathcal{A}(S_m)}] \leq \hat{\mathbb{E}}_{Z_{m'}}[L\hat{y}_{\mathcal{A}(S_m)}] &+ \inf_{(\mathcal{T},f) \in \hat{\mathcal{F}}} V[f] \cdot D^*[\mathcal{T}(Z_{m'}), \mathcal{T}_*\mu]
\end{aligned}$$

□

3 Application to Linear Regression

To make the results presented in Theorem 1 more tangible, the authors examine the linear regression case. We consider the following setting:

- $S_m = \left\{ s^{(i)} \right\}_{i=1}^m$ is a training dataset with the $s^{(i)} = \left(x^{(i)}, y^{(i)} \right)$ being a collection of input-target pairs
- $\phi : (\mathcal{X}, \Sigma_x) \to \left([0,1]^{d_\phi}, \mathcal{B}\left([0,1]^{d_\phi} \right) \right)$ is any normalized measurable function with dimensionality d_ϕ
- $\hat{y}_{\mathcal{A}(S_m)} = \hat{W}\phi(\cdot)$ is the learned model where $\hat{W} = \mathrm{argmin}_W\, \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2} \|W\phi(x) - y\|_2^2 \right]$ is the typical least-squares solution

For the purposes of this report, we focus on Theorem 2 from the original paper which studies the generalization gap $\mathbb{E}_s\left[\frac{1}{2} \|\hat{W}\phi(x) - y\|_2^2 \right] - \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2} \|\hat{W}\phi(x) - y\|_2^2 \right]$ when y has a Gaussian structure.

We assume that $y = W^*\phi(x) + \xi$ where ξ is a random variable with mean zero that is independent of x. The authors make the following definitions:

- μ_x is the normalized measure for input x with respect to the marginal distribution over (x, y), which is unknown to us
- $X_m = \left\{ x^{(i)} \right\}_{i=1}^m$ is the input part of S_m
- $\tilde{S}_m = \left\{ \left(x^{(i)}, \xi^{(i)} \right) \right\}_{i=1}^m$ is the collection of inputs and noise variables corresponding to S_m
- W_l is the l-th column of W

Applying Theorem 1 to this setting yields the following result.

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

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

where $f(t) = \frac{1}{2}\|\hat{W}t - W^*t\|_2^2$, $A_1 = \hat{\mathbb{E}}_{\tilde{S}_m}\left[\xi^T(\hat{W} - W^*)\phi(x)\right]$, $A_2 = \mathbb{E}_\xi\left[\frac{1}{2}\|\xi\|_2^2\right] - \hat{\mathbb{E}}_{\tilde{S}_m}\left[\frac{1}{2}\|\xi\|_2^2\right]$[4], *and*

$$V[f] \leq \sum_{l=1}^{d_\phi}\|(\hat{W}_l - W_l^*)^T(\hat{W} - W^*)\|_1 + \sum_{1 \leq l < l' \leq d_\phi}\left|(\hat{W}_l - W_l^*)^T(\hat{W}_{l'} - W_{l'}^*)\right|.$$

Proof of Theorem 2. We can expand $\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2$ using the definition of the L^2 norm as

$$\begin{aligned}
\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2 &= \frac{1}{2}\|\hat{W}\phi(x) - (W^*\phi(x) + \xi)\|_2^2 \\
&= \frac{1}{2}(\hat{W}\phi(x) - W^*\phi(x) - \xi)^T(\hat{W}\phi(x) - W^*\phi(x) - \xi) \\
&= \frac{1}{2}[(\hat{W}\phi(x))^T(\hat{W}\phi(x)) + (W^*\phi(x))^T(W^*\phi(x)) + \xi^T\xi \\
&\quad - 2(\hat{W}\phi(x))^T(W^*\phi(x)) + 2\xi^T(W^*\phi(x)) - 2\xi^T(\hat{W}\phi(x))] \\
&= \frac{1}{2}[((\hat{W}\phi(x))^T(\hat{W}\phi(x)) - 2(\hat{W}\phi(x))^T(W^*\phi(x)) + (W^*\phi(x))^T(W^*\phi(x))) \\
&\quad + \xi^T\xi + 2\xi^T(W^*\phi(x)) - 2\xi^T(\hat{W}\phi(x))] \\
&= \frac{1}{2}\left[(\hat{W}\phi(x) - W^*\phi(x))^T(\hat{W}\phi(x) - W^*\phi(x)) + \xi^T\xi - 2\xi^T(\hat{W}\phi(x) - W^*\phi(x))\right] \\
&= \frac{1}{2}\left[\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2 + \|\xi\|_2^2 - 2\xi^T(\hat{W} - W^*)\phi(x)\right] \\
&= \frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2 + \frac{1}{2}\|\xi\|_2^2 - \xi^T(\hat{W} - W^*)\phi(x)
\end{aligned}$$

Using this, we can then compute the expectations that make up the generalization gap. Since ξ is a random variable with mean zero that is independent of x, we have by the linearity of expectation

$$\begin{aligned}
\mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] &= \mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2 + \frac{1}{2}\|\xi\|_2^2 - \xi^T(\hat{W}\phi(x) - W^*\phi(x))\right] \\
&= \mathbb{E}_{\mu_x}\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right] + \mathbb{E}_\xi\left[\frac{1}{2}\|\xi\|_2^2\right] - \mathbb{E}_{\mu_x,\xi}\left[\xi^T(\hat{W} - W^*)\phi(x)\right] \\
&= \mathbb{E}_{\mu_x}\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right] + \mathbb{E}_\xi\left[\frac{1}{2}\|\xi\|_2^2\right] - 0 \\
&= \mathbb{E}_{\mu_x}\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right] + \mathbb{E}_\xi\left[\frac{1}{2}\|\xi\|_2^2\right]
\end{aligned}$$

We also have

$$\begin{aligned}
\hat{\mathbb{E}}_{S_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] &= \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2 + \frac{1}{2}\|\xi\|_2^2 - \xi\|_2^2 - \xi^T(\hat{W} - W^*)\phi(x)\right] \\
&= \hat{\mathbb{E}}_{X_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right] + \hat{\mathbb{E}}_{\tilde{S}_m}\left[\frac{1}{2}\|\xi\|_2^2\right] - \hat{\mathbb{E}}_{\tilde{S}_m}\left[\xi^T(\hat{W} - W^*)\phi(x)\right]
\end{aligned}$$

Let $L\hat{y}_{\mathcal{A}(S_m)}(x) = \frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2$, such that $\mathcal{X}$ takes the place of $\mathcal{Z}$ in the definition of $L\hat{y}$ as introduced in Section 2. Moreover, define $A_1 = \hat{\mathbb{E}}_{\tilde{S}_m}\left[\xi^T(\hat{W} - W^*)\phi(x)\right]$ and $A_2 = \mathbb{E}_\xi\left[\frac{1}{2}\|\xi\|_2^2\right] - \hat{\mathbb{E}}_{\tilde{S}_m}\left[\frac{1}{2}\|\xi\|_2^2\right]$. Then,

$$\mathbb{E}_s\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] - \hat{\mathbb{E}}_{S_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - y\|_2^2\right] = \mathbb{E}_{\mu_x}\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right]$$

[4] There is an error in the original paper where the definition of A_2 is missing the $\frac{1}{2}$ factor inside both expectations.

$$
\begin{aligned}
&- \hat{\mathbb{E}}_{X_m}\left[\frac{1}{2}\|\hat{W}\phi(x) - W^*\phi(x)\|_2^2\right] \\
&+ \hat{\mathbb{E}}_{\tilde{S}_m}\left[\xi^T(\hat{W} - W^*)\phi(x)\right] \\
&+ \left(\mathbb{E}_\xi\left[\frac{1}{2}\|\xi\|_2^2\right] - \hat{\mathbb{E}}_{\tilde{S}_m}\left[\frac{1}{2}\|\xi\|_2^2\right]\right) \\
&= \mathbb{E}_{\mu_x}\left[L\hat{y}_{\mathcal{A}(S_m)}(x)\right] - \hat{\mathbb{E}}_{X_m}\left[L\hat{y}_{\mathcal{A}(S_m)}(x)\right] + A_1 + A_2
\end{aligned}
$$

Recall from part (i) of Theorem 1 that

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

and so consequently,

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

if $V[f] < \infty$. If we let $\mu = \mu_x$, $Z_{m'} = X_m$, $\mathcal{T}(x) = \phi(x)$, and $f(t) = \frac{1}{2}\|\hat{W}t - W^*t\|_2^2$ with $t \in \mathbb{R}^{d_\phi}$, then $L\hat{y}_{\mathcal{A}(S_m)} = (f\circ\mathcal{T})(x)$ and $(\mathcal{T}, f) \in \mathcal{F}[L\hat{y}_{\mathcal{A}(S_m)}]$ as required, making application of Theorem 1 very straightforward. So, it is sufficient to show that $V[f] < \infty$ and in particular is bounded by the expression presented in Theorem 2.

First, observe that

$$
\begin{aligned}
\frac{\partial f}{\partial t_l} &= \frac{\partial}{\partial t_l}\left(\frac{1}{2}\|\hat{W}t - W^*t\|_2^2\right) \\
&= (\hat{W}_l - W_l^*)^T(\hat{W} - W^*)t
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial^2 f}{\partial t_l t_{l'}} &= \frac{\partial}{\partial t_l}\left(\frac{\partial f}{\partial t_{l'}}\right) \\
&= (\hat{W}_l - W_l^*)^T(\hat{W}_{l'} - W_{l'}^*)
\end{aligned}
$$

Second, since $\frac{\partial^2 f}{\partial t_l t_{l'}}$ is a constant with respect to t, any higher order derivatives will be zero. Recall from Proposition 1 that if $\partial^k_{1,\ldots,k} f_{j_1\cdots j_k}$ is continuous on $[0,1]^k$, then

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

We have $V^{(k)}\left[f_{j_1\cdots j_k}\right] = 0$ for all $k > 2$ since those higher order derivatives are zero. Since

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

we only need to compute $\sum_{l=1}^{d_\phi} V^{(1)}[f_l]$ and $\sum_{1\leq l<l'\leq d_\phi} V^{(2)}[f_{ll'}]$. If we let $\tilde{t}_l = (t_1,\ldots,t_{d_\phi})$ where each $t_j = 1$ when $j \neq l$, we can use Proposition 1 and the fact that $\|\tilde{t}_l\|_\infty = \max_j |t_j| = 1$ to obtain

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

$$
\begin{aligned}
&= \sum_{l=1}^{d_\phi} \|(\hat{W}_l - W_l^*)^T(\hat{W} - W^*)\|_1 \int_{[0,1]} 1 dt_l \\
&= \sum_{l=1}^{d_\phi} \|(\hat{W}_l - W_l^*)^T(\hat{W} - W^*)\|_1
\end{aligned}
$$

Recall from Proposition 1 that

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

so we also have

$$
\sum_{1 \leq l < l' \leq d_\phi} V^{(2)}\left[f_{ll'}\right] \leq \sum_{1 \leq l < l' \leq d_\phi} \left|(\hat{W}_l - W_l^*)^T(\hat{W}_{l'} - W_{l'}^*)\right|
$$

Together,

$$
V[f] \leq \sum_{l=1}^{d_\phi} \|(\hat{W}_l - W_l^*)^T(\hat{W} - W^*)\|_1 + \sum_{1 \leq l < l' \leq d_\phi} \left|(\hat{W}_l - W_l^*)^T(\hat{W}_{l'} - W_{l'}^*)\right|
$$

As long as $\|\hat{W} - W^*\| < \infty$ for any norm, since the equivalency of norms states that $C_1\|x\|_b \leq \|x\|_a \leq C_2\|x\|_b$, we have $V[f] < \infty$ and so

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

Therefore, we can finally conclude

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

□

4 Conclusion

This report examined foundational aspects of generalization in machine learning through the lens of analytical learning theory, focusing on the variation of functions, the expected error, and applications to linear regression. To do so, we explored key results, including Proposition 1, which establishes bounds on function variation in terms of partial derivatives, and Theorem 1, which decomposes the generalization gap into two interpretable components: the variation of functions and dataset discrepancy. Building on these results, Theorem 2 applied the framework to the classical linear regression problem, demonstrating that tight, instance-specific bounds on expected error can be achieved even in high-dimensional settings with structured labels.

These findings are significant because they move beyond traditional statistical learning theory, which relies on population-level guarantees, by providing strongly instance-dependent results that are directly applicable to specific learning problems. The shift from statistical assumptions to measure-theoretic analysis not only complements existing theoretical frameworks but also offers new tools for understanding generalization in practical scenarios. From this project, I have gained a deeper appreciation for the interplay between functional analysis, measure theory, and generalization theory, as well as for the critical role of rigorous proofs in advancing our understanding of machine learning principles.

References

[1] C. Aistleitner and J. Dick, *Functions of bounded variation, signed measures, and a general Koksma-Hlawka inequality*, Acta Arithmetica, 167 (2015), pp. 143–171.

[2] C. Aistleitner, F. Pausinger, A. M. Svane, and R. F. Tichy, *On functions of bounded variation*, Mathematical Proceedings of the Cambridge Philosophical Society, 162 (2017), p. 405–418.

[3] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, Harcourt/Academic Press, 2000.

[4] K. Kawaguchi, Y. Benigo, V. Verma, and L. Kaelbling, *Towards Understanding Generalization via Analytical Learning Theory*, 2018.

[5] A. S. Leonov, *On the total variation for functions of several variables and a multidimensional analog of Helly's selection principle*, Mathematical Notes, 63 (1996), pp. 61–71.

[6] D. Pollard, *A User's Guide to Measure Theoretic Probability*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2001.