Journal of Machine Learning Research 25 (2024) 1-45 Submitted 1/23; Revised 6/24; Published 6/24

An Entropy-Based Model for Hierarchical Learning

Amir R. Asadi ASADI@STATSLAB.CAM.AC.UK
Statistical Laboratory,
Centre for Mathematical Sciences,
University of Cambridge,
Cambridge CB3 0WA
United Kingdom

Editor: Gabor Lugosi

Abstract

Machine learning, the predominant approach in the field of artificial intelligence, enables computers to learn from data and experience. In the supervised learning framework, accurate and efficient learning of dependencies between data instances and their corresponding labels requires auxiliary information about the data distribution and the target function. This central concept aligns with the notion of regularization in statistical learning theory. Real-world datasets are often characterized by multiscale data instance distributions and well-behaved, smooth target functions. Scale-invariant probability distributions, such as power-law distributions, provide notable examples of multiscale data instance distributions in various contexts. This paper introduces a hierarchical learning model that leverages such a multiscale data structure with a multiscale entropy-based training procedure and explores its statistical and computational advantages. The hierarchical learning model is inspired by the logical progression in human learning from easy to complex tasks and features interpretable levels. In this model, the logarithm of any data instance's norm can be construed as the data instance's complexity, and the allocation of computational resources is tailored to this complexity, resulting in benefits such as increased inference speed. Furthermore, our multiscale analysis of the statistical risk yields stronger guarantees compared to conventional uniform convergence bounds.

Keywords: machine learning, neural network, chaining, information theory, scale-invariant distribution, curriculum learning, logarithmic binning

1. Introduction

This paper introduces a hierarchical learning model with a multiscale entropy-based training mechanism. Designed to exploit the multiscale structure in many real-world datasets, the proposed model aims to provide statistical and computational efficiency while featuring interpretability.

1.1 Background

In contemporary times, machine learning is the predominant approach in artificial intelligence, enabling computers to acquire knowledge from data and experience. Within the supervised learning paradigm, training data is postulated to arise randomly as pairs (X, Y), denoting the data instance and its corresponding label, respectively. A computer is presented with a sequence of such instances and labels independently drawn from the underlying data probability distribution. Its task is to discern the relationship between the data instances and their labels, ultimately predicting the label of new, randomly drawn instances. In this paper, for simplicity, we adopt the *function learning* setting, part of the realizability assumption in statistical learning theory. That is, we assume that the dependency between data instances and their corresponding labels is modeled noiselessly by a deterministic target function $Y = T(X)$, mapping any data instance to its label (Shalev-Shwartz and Ben-David, 2014; Mendelson, 2008).

An insightful observation is that the sequence of training examples typically lacks comprehensive information about the target function or the data domain. Beyond the training data, providing the computer *auxiliary information* through the learning model is crucial. Higher auxiliary information enables the computer to learn the target function more accurately and more efficiently. To illustrate this point further, consider the extreme scenario where no auxiliary information is given to the computer, implying a total lack of knowledge about the target function or the data domain beyond the training data. In such a case, the optimal task for the computer becomes merely memorizing the training examples, leaving it unable to predict the label for any new instance not present in the training data. This approach is very prone to *overfitting*, most likely resulting in poor performance on new and unseen data instances.

This notion of auxiliary information aligns with the idea of *regularization* studied in statistical learning theory; see, for example, (Vapnik, 1999). Regularization manifests in various forms, such as restricting the hypothesis class and explicitly and implicitly regularizing the training mechanism. Moreover, it is intricately linked to the *no-free-lunch theorem*, which underscores the necessity for every learning algorithm to possess some level of prior knowledge about the underlying assumptions of the learning problem to attain success; see, for instance, (Shalev-Shwartz and Ben-David, 2014, Theorem 5.1). Analogously, in most tasks, human knowledge is acquired through a combination of training data, prior knowledge, and intuition.

A prevalent characteristic in many real-world datasets is the multiscale nature of their data instances. In other words, in these datasets, data manifest across different scales of magnitude, exhibiting a diverse range of sizes and complexities. This inherent property is used in applications in various fields, such as wavelet theory, Fourier analysis, and signal processing, as detailed in (E, 2011) and references within. In particular, empirical data distributions across various domains such as physics, biology, medicine, finance, natural language processing, and the social sciences frequently exhibit power-law distributions. Notable instances of power-law probability distributions include the distributions of people's incomes, city populations, stars' brightness, file sizes in computing, earthquake magnitudes, and word frequencies in human languages, among other examples. These phenomena have been extensively studied, as demonstrated in works such as (Newman, 2005; Clauset et al., 2009) and references therein. The generation of power-law distributions in the real world in natural and artificial systems involves diverse mechanisms, and each is believed to apply to specific applications; see (Sornette, 2006, Chapter 14) and (Newman, 2005; Mitzenmacher, 2004). The most important of such mechanisms are considered to be growth with preferential attachment (Yule's process) and critical phenomena (Newman, 2005). Furthermore, the *generalized central limit theorem* posits that the normalized sum of independent and identically distributed random variables with infinite variance can only converge to a stable distribution; see, for example, (Nolan, 2020). All stable distributions with infinite variance exhibit power-law tails, whereas the Gaussian distribution is the sole stable distribution with finite variance (Samoradnitsky and Taqqu, 2017). Power-law distributions are examples of scale-invariant probability distributions.

Additionally, target functions in the real world, relating continuous data instances to their real-valued labels, tend to be smooth and well-behaved. Leveraging such characteristics of real-world data instances and target functions as auxiliary information in a machine-learning model presents an opportunity to yield statistical and computational benefits in real-world applications.

1.2 Overview of the Contributions

In this paper, we introduce a hierarchical learning model to take advantage of the aforementioned multiscale nature of data instances and the smoothness of their target functions. The model comprises a compositional learning architecture with a sequential multiscale training mechanism. First, the training data is partitioned at different scales, and the training mechanism starts by learning from the batch of the smallest data and progressing step by step toward the batches of larger data. The learning process at each batch of data takes advantage of the learned model over smaller data as prior information. This is inspired by the logical learning mechanism observed in humans, who progressively learn different tasks, commencing with small (easy) examples and advancing toward large (complex) examples. Due to the ubiquity of such multiscale data, our proposed learning model holds promise for many applications.

More precisely, we consider the following two assumptions:

(a) Data instances $X_i \in \mathbb{R}^m$ emerge in different scales of magnitude from a distribution μ defined on the data domain $\mathcal{X}$. Here, we assume that for $0 < \varepsilon < R$, the data domain is $\mathcal{X} = \{x \in \mathbb{R}^m : \varepsilon \leq |x| < R\}$, where $|x|$ denotes the Euclidean norm of datum x. Typically, ε is much smaller than R. Later in the paper in Subsection 5.2, we assume that μ is a scale-invariant probability distribution with shape parameter α, whose probability density $q(x)$ satisfies the following condition: For all $x \in \mathcal{X}$ and any $\gamma \geq 1$ such that $x/\gamma \in \mathcal{X}$, we have

$$q\left(\frac{x}{\gamma}\right) = \gamma^{\alpha} q(x). \tag{1}$$

(b) The target function $T : B_R^m \to \mathbb{R}^m$ is well-behaved, where $B_R^m = \{x \in \mathbb{R}^m : |x| < R\}$ denotes the m-dimensional Euclidean ball with radius $R > 0$, centered at the origin. In this paper, we assume that T is differentiable and smooth, and we further assume both the invertibility of T and the Lipschitz continuity of its inverse, T^{-1}. For the special case of $m = 1$, we show (in Theorem 19) that an additional smoothness assumption of T^{-1} leads to a strengthened result. We further assume that from prior knowledge or intuition, the learning model knows the behavior of function T on B_ε^m, data instances at very small scales.

We take advantage of such assumptions on the data in our hierarchical learning model based on the following contributions:

Ladder decompositions of functions The first main result of the paper studies *ladder decompositions* of invertible functions $T : B_R^m \to \mathbb{R}^m$. The decomposition is defined in terms of *dilations* $T_{[\gamma]} : B_R^m \to \mathbb{R}^m$ of T, where for any scale $0 < \gamma \leq 1$ and all $x \in B_R^m$,

$$T_{[\gamma]}(x) := \frac{T(\gamma x)}{\gamma}.$$

The dilation $T_{[\gamma]}$ can be conceptualized as a 'zoomed' version of the original function T into the origin, where the degree of zooming is determined by the scale γ. Given a sequence of dilation scales $0 < \gamma_0 < \cdots < \gamma_d = 1$, the ladder decomposition is defined based on the sequence of dilations $\{T_{[\gamma_k]} : k = 0, \ldots, d\}$ interpolating between $T_{[\gamma_0]}$ and $T_{[1]} = T$. The decomposition is constructed as a sequence of successive compositions of functions $T_k := T_{[\gamma_k]} \circ T^{-1}_{[\gamma_{k-1}]}$ for all $k = 1, \ldots, d$, and writing

$$T_{[\gamma_k]} = T_k \circ \ldots T_1 \circ T_{[\gamma_0]}.$$

Building on an earlier result from (Bartlett et al., 2018), we show that the smaller the value of $\gamma_k - \gamma_{k-1}$ is, then the closer T_k is to the identity mapping, thus being more well-behaved and 'easier' to learn. To elaborate, for all $1 \leq k \leq d$, let $\psi_k(x) := T_k(x) - x$. The first main contribution of the paper, Corollaries 20 and 26 in Section 3, states that if M_2-smooth invertible function $T : B_R^m \to \mathbb{R}^m$ is such that T^{-1} is M_1-Lipschitz, then

$$\|\psi_k\|_{\mathrm{Lip}} \leq C(\gamma_k - \gamma_{k-1}),$$

where $\|\psi_k\|_{\mathrm{Lip}}$ denotes the Lipschitz norm of ψ_k and constant C only depends on M_1, M_2 and R.

The definitions of the ladder decomposition and functions ψ_k yield that, for all $1 \leq k \leq d$, we have

$$T_{[\gamma_k]} = T_k \circ T_{[\gamma_{k-1}]} = T_{[\gamma_{k-1}]} + \psi_k \circ T_{[\gamma_{k-1}]}.$$

This dependency between the levels of the ladder decomposition fits with a hierarchical learning model with residual levels of the following form:

$$h_k(x) := h_{k-1}(x) + f(h_{k-1}(x); w_k). \tag{2}$$

Here, $h_0 := T_{[\gamma_0]}$, and $f(\cdot; w)$ is a learnable model with parameters w (a vector), for example, a neural network. A specific example of f, a two-layer network with step activation functions, is explored in Section 6. Let $\mathsf{w} = (w_1, \ldots, w_d)$ denote the parameters of the whole hierarchical model. For any $1 \leq k \leq d$, define $w_{1:k} := (w_1, \ldots, w_k)$. Level h_k is a function of the input datum x and the weights $w_{1:k}$. Therefore, we sometimes denote h_k with $h_k(x; w_{1:k})$. Parameters $w_1, \ldots, w_d$ will be chosen sequentially by the training mechanism such that $f(\cdot; w_k)$ approximates ψ_k well for all $k = 1, \ldots, d$. Thus, for all $k = 1, \ldots, d$, the kth level of our learning model (h_k) aims to approximate the dilation $T_{[k]}$ of the target function.

The hierarchical learning model By defining $\gamma := (R/\varepsilon)^{1/d}$, we split the data domain $\mathcal{X}$ into d scales:

$$\mathcal{X}_k := \left\{x \in \mathbb{R}^m : \varepsilon\gamma^{k-1} \leq |x| < \varepsilon\gamma^k\right\},$$

for all $1 \leq k \leq d$. This multiscale partitioning of data domains is sometimes referred to as *logarithmic binning* in the literature; see, for instance, (Newman, 2005). We then consider the ladder decomposition of the target function T with respect to $\gamma_k := \gamma^{k-d}$, for $k = 0, \ldots, d$. The logarithm of the norm of a datum $x \in \mathcal{X}$ is viewed as a measure of its complexity; that is, the sets $\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_d$ represent inputs of increasing complexity. We assume that the target function is known at level γ_0, that is, $T_{[\gamma_0]}$ is given. In other words, the behavior of the function T at minimal values of x is known, and we shall sequentially learn $T_{[\gamma_1]}, \ldots, T_{[\gamma_d]}$. For all $1 \leq k \leq d$, since the training mechanism aims to make $h_k(x)$ closely approximate dilation $T_{[\gamma_k]}(x) = T(\gamma_k x)/\gamma_k$, we define the output of the learned model for any $x \in \mathcal{X}_k$ as

$$h_{\mathsf{w}}(x) := \gamma_k h_k\left(\frac{x}{\gamma_k}\right). \tag{3}$$

Given training examples $\mathsf{s} = (x_i, T(x_i))_{i=1}^n$, the training of the model is performed by sampling the parameters $w_1, w_2, \ldots, w_d$ in a sequential manner from a sequence of Gibbs measures. Namely, we start by considering training data examples whose instances belong to the smallest scale $\mathcal{X}_1$, corresponding to the 'easiest' instances. An approximation of ψ_1 is then learned by sampling w_1 from a Gibbs measure, a maximum entropy distribution, defined with the empirical risk over these small-scaled examples with the absolute error loss function and with hyperparameter λ_1. More precisely, w_1 is sampled from a discrete set $\mathcal{W}_1$ according to the Gibbs measure:

$$\mathbb{P}_1(W_1 = w_1) := \frac{\exp\left(-\frac{1}{n\lambda_1}\sum_{x_i \in \mathsf{s}\cap\mathcal{X}_1}\left|\gamma_1 h_1\left(\frac{x_i}{\gamma_1}; w_1\right) - T(x_i)\right|\right)}{\sum_{w_1' \in \mathcal{W}_1}\exp\left(-\frac{1}{n\lambda_1}\sum_{x_i \in \mathsf{s}\cap\mathcal{X}_1}\left|\gamma_1 h_1\left(\frac{x_i}{\gamma_1}; w_k'\right) - T(x_i)\right|\right)}.$$

Subsequently, we observe data at the next scale $\mathcal{X}_2$, characterized by a higher magnitude. Similarly, given the learned approximation $f(\cdot, w_1)$ of ψ_1, an approximation $f(\cdot, w_2)$ of ψ_2 is learned by sampling w_2 from a Gibbs measure with hyperparameter λ_2, and this process is iteratively repeated. That is, for $k = 2, \ldots, d$, we sample $w_k \in \mathcal{W}_k$ conditionally on $w_{1:(k-1)}$ with probability

$$\mathbb{P}_{W_k|W_{1:(k-1)}}\left(w_k\middle|w_{1:(k-1)}\right) := \frac{\exp\left(-\frac{1}{n\lambda_k}\sum_{x_i \in \mathsf{s}\cap\mathcal{X}_k}\left|\gamma_k h_k\left(\frac{x_i}{\gamma_k}; w_{1:k}\right) - T(x_i)\right|\right)}{\sum_{w_k' \in \mathcal{W}_k}\exp\left(-\frac{1}{n\lambda_k}\sum_{x_i \in \mathsf{s}\cap\mathcal{X}_k}\left|\gamma_k h_k\left(\frac{x_i}{\gamma_k}; w_{1:(k-1)}w_k'\right) - T(x_i)\right|\right)}.$$

Consequently, learning the target function for data at smaller scales is inherently 'easier,' serving as a foundational step for tackling more challenging learning tasks at higher scales. Thus, this hierarchical model is also a model for implementing the concept of easy-to-hard learning, commonly known as *curriculum learning* (Bengio et al., 2009). Here, scale serves as a temporal progression—akin to how humans learn a course by starting with simpler examples and gradually advancing to more complex ones. This hierarchical approach to learning the ladder decomposition is metaphorically akin to the step-by-step ascent of a ladder.

The total training is then modeled with the following measure:

$$\mathbb{P}^{\star}_{\mathsf{W}}(\mathsf{w}) := \mathbb{P}_{W_1}(w_1)\mathbb{P}_{W_2|W_1}(w_2|w_1)\dots\mathbb{P}_{W_d|W_{1:(d-1)}}\big(w_d|w_{1:(d-1)}\big).$$

The next main result of the paper, Theorem 29, states that $\mathbb{P}^{\star}_{\mathsf{W}}$ minimizes a multiscale loss regularized by a multiscale form of entropy. That is,

$$\mathbb{P}^{\star}_{\mathsf{W}} = \arg\min_{P_{\mathsf{W}}} \left\{ \mathbb{E}\left[\ell^{(\lambda)}(\mathsf{W},\mathsf{s})\right] - \sum_{k=1}^{d}(\lambda_k - \lambda_{k+1})H(W_{1:k}) \right\}, \tag{4}$$

where $\ell^{(\lambda)}$ is related to the total loss over the multiple scales and $H(W_{1:k})$ is Shannon entropy. The proof's idea, which expands upon the proof technique of (Asadi and Abbe, 2020, Theorem 13), is to show that the functional optimized by $\mathbb{P}^{\star}_{\mathsf{W}}$ can be decomposed into a sum of conditional entropies of W_k given $W_{1:(k-1)}$, for all $1 \le k \le d$.

The statistical risk The paper's final main result bounds the model's statistical risk when the parameters are chosen according to the above multiscale training mechanism. The result proceeds through a 'chaining' argument, where the loss is decomposed over the successive stages of the training procedure, during which the parameters $\mathsf{w} = (w_1, \dots, w_d)$ are chosen.

Under the assumption that the model is *realizable*, that is, there exists a choice of parameters $\hat{w}_1, \dots, \hat{w}_d$ such that $\psi_k(\cdot) = f(\cdot; \hat{w}_k)$, for all $1 \le k \le d$, we define the *chained risk* as

$$L^{(\mathsf{C})}_{\mu}(\mathsf{w}) := \mathbb{E}\left[\sum_{k=1}^{d}\Big(\ell_k(w_{1:k}, X) - \ell_k\big(w_{1:(k-1)}\hat{w}_k, X\big)\Big)\right],$$

where $X \sim \mu$ and, for all $1 \le k \le d$,

$$\ell_k(w_{1:k}, x) := \begin{cases} |h_{\mathsf{w}}(x) - T(x)| = \left|\gamma_k h_k\left(\frac{x}{\gamma_k}, w_{1:k}\right) - T(x)\right| & \text{if } x \in \mathcal{X}_k \\ 0 & \text{if } x \notin \mathcal{X}_k \end{cases}$$

is the loss of the model on data at scale k.

Theorem 33 then states that when $(\mathsf{S}, \mathsf{W}) \sim \mu^{\otimes n}\mathbb{P}^{\star}_{\mathsf{W}|\mathsf{S}}$, that is, when training data S are n independent samples from distribution μ and the parameters W are the outputs of the multiscale entropy-based training mechanism given random training data S, then the expected chained risk satisfies

$$\mathbb{E}\left[L^{(\mathsf{C})}_{\mu}(\mathsf{W})\right] \le 2\sum_{k=1}^{d}\left((\lambda_k - \lambda_{k+1})\left(\sum_{j=1}^{k}\log|\mathcal{W}_j|\right) + \frac{4\gamma_k^2\rho_k^2}{n(\lambda_k - \lambda_{k+1})}\right),$$

where $\lambda_{d+1} := 0$ and $\rho_k = O(\gamma^k)$ is the maximum output norm of $\psi_k(\cdot)$, for all $1 \le k \le d$. Optimizing the hyperparameters $\lambda_1, \dots, \lambda_d$ of our learning mechanism in the bound above yields that, for all $1 \leq k \leq d$, we have

$$\lambda_k - \lambda_{k+1} = \frac{2\gamma_k \rho_k}{\sqrt{n \sum_{j=1}^{k} \log |\mathcal{W}_j|}}.$$

This, in turn, establishes the following bound on the expected chained risk:

$$\mathbb{E}\Big[L_\mu^{(\mathsf{C})}(\mathsf{W})\Big] \leq \frac{8}{\sqrt{n}} \sum_{k=1}^{d} \gamma_k \rho_k \sqrt{\sum_{j=1}^{k} \log |\mathcal{W}_j|}. \tag{5}$$

Although the true parameter $\hat{\mathsf{w}} = (\hat{w}_1, \ldots, \hat{w}_d)$ achieves zero chained risk, in general it is not clear if a low expected chain risk implies low statistical risk. In fact, the chained risk is always a *lower bound* for the statistical risk. However, for any *scale-invariant* probability distribution satisfying equation (1), with sufficiently large shape parameter α, we demonstrate in Subsection 5.2 an *upper bound* on the statistical risk, based on the chained risk. More precisely, we show that for such scale-invariant μ, there exists a constant $\hat{C} > 0$ independent of w such that

$$\hat{C} L_\mu(\mathsf{w}) \leq L_\mu^{(\mathsf{C})}(\mathsf{w}),$$

yielding an upper bound on the expected statistical risk from (5).

Bounded-norm parameterization example In Section 6, we study a particular example of the function f in (2), where invertible functions that are Lipschitz and smooth are approximated by two-layer neural networks with step activation functions and bounded-norm parameters. We study the approximation error, a bound on the network's output, and a bound on the network parameters' norm. The parameters of the example are discretized, hence the hypothesis set of the whole hierarchical model is finite. We then apply the derived bound on the statistical risk of the hierarchical model from the earlier sections to this example.

1.3 Related Work

This work draws inspiration primarily from integrating concepts presented in (Bartlett et al., 2018) and (Asadi and Abbe, 2020). The paper (Bartlett et al., 2018) establishes that any smooth bi-Lipschitz function T can be represented as a composition of d functions $T_d \circ \cdots \circ T_1$, where each function T_k, for $1 \leq k \leq d$, approximates the identity function, manifested by the Lipschitz norm of $T_k(x) - x$ decreasing inversely with d. Notably, the proof of (Bartlett et al., 2018, Theorem 2) employs the concepts of function dilation and ladder decomposition. On the other hand, the paper (Asadi and Abbe, 2020) addresses the solution to the multiscale entropy regularization problem, an extension of the Gibbs probability distribution to a multiscale context. However, efficient sampling from the optimal probability distribution, the output of the Marginalize-Tilt algorithm of (Asadi and Abbe, 2020), remains challenging due to multiple steps of marginalizations of probability distributions. To alleviate this problem, (Asadi and Loh, 2024) considers finding self-similar approximations to the optimal probability distribution. Developing on the proof technique of (Asadi and Abbe, 2020), this paper introduces the multiscale loss function $\ell^{(\lambda)}$ that, when regularized with multiscale entropy in (4), is minimized by the self-similar and computable distribution $\mathbb{P}^{\star}_{\mathsf{W}}$. Distinct from these works, our approach emphasizes interpretability by hierarchically learning ladder decompositions of target functions and involves reading the output of the multilevel learning model from different levels and depths, contingent on the scale of the input data instance, as in (3). In essence, we align the multiscale architecture of the learning model with the multiscale data domain.

Information-theoretic methods for analyzing the statistical risk and the generalization error of learning algorithms have been pioneered within the framework of PAC-Bayesian bounds (McAllester, 1999). This line of inquiry later evolved, adopting a related form that utilizes mutual information, exemplified by the works (Russo and Zou, 2019), (Xu and Raginsky, 2017) and (Bu et al., 2020). In a related vein, (Raginsky et al., 2016) introduces alternative information-theoretic measures to assess the stability of learning algorithms and bounds on their generalization capabilities. The extension of these information-theoretic methods to multiscale techniques, inspired by the method of 'chaining' in probability theory (see, for example, Talagrand, 2014), is presented in (Audibert and Bousquet, 2007),(Asadi et al., 2018), (Asadi and Abbe, 2020), and (Clerico et al., 2022). Multiscale entropies, a combination of entropies at different scales, play an implicit role in chaining. Notably, Dudley's inequality (Dudley, 1967) can be variationally transformed, expressing the bound as a linear mixture of metric entropies across multiple scales. The work (Xu and Raginsky, 2017) also studies into the expected statistical risk of sampling from the Gibbs distribution, sometimes referred to as maximum-entropy training. A subsequent multiscale extension of this analysis has likewise been derived in (Asadi and Abbe, 2020).

The work (Fletcher and Markovic, 2012) establishes that any diffeomorphism defined on the sphere can be decomposed as the composition of bi-Lipschitz functions with small distortion.

Hierarchical learning models comprising layers that are nearly identity functions find applications in deep learning, evident in the context of residual networks (He et al., 2016) and through dynamical systems approaches; see, for example, (E, 2017). A plausible rationale behind the superior performance of deep neural networks compared to shallow networks is their capability to learn different aspects of the data distribution across various layers. This hypothesis finds support in empirical evidence where pre-trained bottom layers combined with task-specific layers achieve excellent performance in image classification tasks; see, for instance, (Devlin et al., 2018) and (Girshick et al., 2014).

1.4 Organization of This Paper

The rest of the paper is organized as follows: We provide the preliminaries and notation in Section 2. Following that, in Section 3, we introduce the concept of ladder decompositions of functions, examining the Lipschitz continuity and smoothness of its components. Section 4 explains our proposed learning model. Section 5 comprises two parts: Subsection 5.1 demonstrates the efficacy of multiscale entropy-based training in achieving low chained risk, a new analytical tool. Subsection 5.2 establishes that if the data distribution μ is scale-invariant, the chained risk can serve as an upper bound on the statistical risk, thereby providing an overall upper bound on the statistical risk of the learned model. Section 6 gives an example where diffeomorphisms can be represented using a parameterized model with bounded-norm parameters. We then compute the bound on the statistical risk from Section 5 for this example. Finally, Section 7 encapsulates our work's conclusions.

2. Preliminaries and Notation

The sets of real numbers and integers are symbolized by $\mathbb{R}$ and $\mathbb{Z}$, respectively. Throughout the paper, $|x|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^m$, $|x|_1$ represents the ℓ_1-norm of a vector $x \in \mathbb{R}^m$, $|A|_2$ denotes the spectral norm of a matrix $A \in \mathbb{R}^{m\times m}$, and $\|f\|_{\mathrm{Lip}}$ is the Lipschitz norm of a function f. The identity function is denoted as id. For a pair of positive integers $k \leq n$, let $\binom{n}{k}$ represent the 'n choose k' binomial coefficient. For any $x \in \mathbb{R}$, the floor function $\lfloor x \rfloor$ equals the largest integer smaller than or equal to x. Given two matrices $A, B \in \mathbb{R}^{m\times m}$, $A \preceq B$ indicates that $B - A$ is a positive semidefinite matrix.

Random variables and vectors are denoted by capital letters, while lowercase letters represent their specific realizations. The equiprobable (uniform) probability distribution is represented by U, and its support is indicated with a subscript. If X is a random variable and P is a probability measure, the notation $X \sim P$ signifies that X is distributed according to P. The Dirac probability measure on w is denoted as δ_{w}. The n-fold tensor product of a measure μ with itself is represented by $\mu^{\otimes n}$. For two distributions P and Q, $P \ll Q$ means that P is absolutely continuous with respect to Q.

In supervised batch learning, we define $\mathcal{X}$ as the instance domain and $\mathcal{Y}$ as the label domain. A target function $T : \mathcal{X} \to \mathcal{Y}$ exists, which maps any data instance to its label. We also define the hypothesis set $\mathcal{H} = \{h_{\mathsf{w}} : \mathsf{w} \in \mathcal{W}\}$ consisting of hypotheses indexed by the set $\mathcal{W}$. A loss function $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}^+$ is introduced.

A learning algorithm is presented with a random training sequence $\mathsf{S} = (X_1, ..., X_n)$ comprising n data instances along with their corresponding labels $(T(X_1), \ldots, T(X_n))$, where S is drawn i.i.d. from $\mathcal{X}$ with an unknown distribution μ. In other words, $\mathsf{S} \sim \mu^{\otimes n}$. During the training procedure, the algorithm selects $h_{\mathsf{W}} \in \mathcal{H}$, modeled by a random transformation $P_{\mathsf{W}|\mathsf{S}}$.

Definition 1 (Statistical Risk) *For any* $\mathsf{w} \in \mathcal{W}$*, the statistical (or population) risk of* w *is defined as*

$$L_\mu(\mathsf{w}) := \mathbb{E}[\ell(\mathsf{w}, X)], \tag{6}$$

where $X \sim \mu$*.*

A primary goal of statistical learning is to identify computationally efficient learning algorithms for which, given a random training set of size n, the expected statistical risk $\mathbb{E}[L_\mu(\mathsf{W})]$ is small. Here, W is distributed with the marginal distribution of $P_{\mathsf{WS}} = \mu^{\otimes n} P_{\mathsf{W}|\mathsf{S}}$.

Next, we present some preliminary tools later used in the paper.

Definition 2 (Entropy) *The Shannon entropy of a discrete random variable* X*, taking values on set* $\mathcal{A}$*, is defined as*

$$H(X) := -\sum_{x\in\mathcal{A}} P_X(x) \log P_X(x).$$

The relative entropy between two distributions P_X and Q_X, if $P_X \ll Q_X$ is defined as

$$D(P_X \| Q_X) := \sum_{x \in \mathcal{A}} P_X(x) \log\left(\frac{P_X(x)}{Q_X(x)}\right),$$

otherwise, we define $D(P_X \| Q_X) := \infty$. The conditional relative entropy is

$$\begin{aligned} D\big(P_{Y|X} \big\| Q_{Y|X} \big| P_X\big) &:= \sum_{x \in \mathcal{A}} D(P_{Y|X=x} \| Q_{Y|X=x}) P_X(x) \\ &= \mathbb{E}\big[D\big(P_{Y|X}(\cdot|X) \big\| Q_{Y|X}(\cdot|X)\big)\big], \quad X \sim P_X. \end{aligned}$$

The following useful property of entropy is called the 'chain rule'. For proof, see, for example, (Cover and Thomas, 2012, Theorem 2.5.3):

Lemma 3 (Entropy Chain Rule) *Let P_{XY} and Q_{XY} be two distributions. We have*

$$D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X) + D\big(P_{Y|X} \big\| Q_{Y|X} \big| P_X\big).$$

The next definition relates to 'geometric' transformations of probability measures:

Definition 4 (Escort and Tilted Distributions) *Given a discrete probability measure P defined on a set $\mathcal{A}$, and any $\lambda \in [0,1]$, we define the escort distribution $(P)^\lambda$ for all $a \in \mathcal{A}$ as*

$$(P)^\lambda(a) := \frac{P^\lambda(a)}{\sum_{x \in \mathcal{A}} P^\lambda(x)}.$$

Given two discrete probability measures P and Q defined on a set $\mathcal{A}$, and any $\lambda \in [0,1]$, we define the tilted distribution $(P,Q)^\lambda$ as the following geometric mixture:

$$(P,Q)^\lambda(a) := \frac{P^\lambda(a) Q^{1-\lambda}(a)}{\sum_{x \in \mathcal{A}} P^\lambda(x) Q^{1-\lambda}(x)}.$$

Evidently, if U is the equiprobable distribution on $\mathcal{A}$, then

$$(P)^\lambda = (P,U)^\lambda.$$

In our analysis in Section 5, similar to (Asadi and Abbe, 2020), we encounter linear combinations of relative entropies. The next lemma shows the role of tilted distributions in dealing with such linear combinations; see, for example, (van Erven and Harremoës, 2014, Theorem 30):

Lemma 5 (Entropy Combination) *Let $\lambda \in [0,1]$. For any distributions P, Q and R defined on a discrete set $\mathcal{A}$ such that $P \ll Q$ and $P \ll R$, we have*

$$\lambda D(P \| Q) + (1-\lambda) D(P \| R) = D\Big(P \Big\| (Q,R)^\lambda\Big) - \log\left(\sum_{x \in \mathcal{A}} Q^\lambda(x) R^{1-\lambda}(x)\right).$$

Proof Let $Z := \sum_{x\in\mathcal{A}} Q^{\lambda}(x)R^{1-\lambda}(x)$. We have

$$\begin{aligned}\lambda D(P\|Q)+(1-\lambda)D(P\|R) &= \lambda\sum_{x\in\mathcal{A}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)+(1-\lambda)\sum_{x\in\mathcal{A}} P(x)\log\left(\frac{P(x)}{R(x)}\right)\\ &= \sum_{x\in\mathcal{A}} P(x)\log\left(\frac{P(x)}{Q^{\lambda}(x)R^{1-\lambda}(x)}\right)\\ &= \sum_{x\in\mathcal{A}} P(x)\log\left(\frac{P(x)}{\frac{Q^{\lambda}(x)R^{1-\lambda}(x)}{Z}}\right)-\log Z\\ &= D\left(P\middle\|(Q,R)^{\lambda}\right)-\log\left(\sum_{x\in\mathcal{A}} Q^{\lambda}(x)R^{1-\lambda}(x)\right).\end{aligned}$$

■

We provide the following definition to later simplify the notation in the proof of Theorem 29:

Definition 6 (Congruent Functionals) *We call two functionals $\mathcal{L}_1(P)$ and $\mathcal{L}_2(P)$ of a distribution P congruent and write $\mathcal{L}_1 \cong \mathcal{L}_2$ if $\mathcal{L}_1 - \mathcal{L}_2$ does not depend on P.*

For example, Lemma 5 implies that if Q and R are fixed distributions, then as functionals of P, the following congruency holds:

$$\lambda D(P\|Q)+(1-\lambda)D(P\|R)\cong D\left(P\middle\|(Q,R)^{\lambda}\right).$$

Specifically, if U is the equiprobable distribution, then

$$\lambda D(P\|Q)+(1-\lambda)D(P\|U)\cong D\left(P\middle\|(Q)^{\lambda}\right). \tag{7}$$

The following well-known result, sometimes referred to as the Gibbs variational principle, implies that the distribution that minimizes the sum of average energy (loss) and entropy (regularization) is the Gibbs measure:

Lemma 7 (Gibbs Variational Principle) *Let $\mathcal{W}$ be an arbitrary finite set and U_W be the equiprobable distribution on $\mathcal{W}$. Given a function $g:\mathcal{W}\to\mathbb{R}$ and $\lambda>0$, we define the following Gibbs probability distribution for all $w\in\mathcal{W}$:*

$$Q_W(w)\triangleq\frac{\exp\left(-\frac{g(w)}{\lambda}\right)}{\sum_{w'\in\mathcal{W}}\exp\left(-\frac{g(w')}{\lambda}\right)}.$$

Then, for any probability measure P_W defined on $\mathcal{W}$, we have

$$\mathbb{E}[g(W)]+\lambda D(P_W\|U_W)=\lambda D(P_W\|Q_W)-\lambda\log\left(\sum_{w'\in\mathcal{W}}\exp\left(-\frac{g(w')}{\lambda}\right)\mathrm{d}v\right),$$

where $W\sim P_W$.

Particularly, Lemma 7 yields the following congruency identity as functionals of P_W:

$$\mathbb{E}[g(W)] + \lambda D(P_W \| U_W) \cong \lambda D(P_W \| Q_W). \tag{8}$$

We later make use of the congruency relations (7) and (8) iteratively in the proof of Theorem 29.

In Section 5, we require the following well-known result on the Log-Sum-Exp function:

Lemma 8 (Log-Sum-Exp) *For any positive integer N, let $\mathsf{z} = (z_1, z_2, \ldots, z_N) \in \mathbb{R}^N$ be arbitrarily chosen. For any $\lambda > 0$, the Log-Sum-Exp function*

$$G_\lambda(\mathsf{z}) := -\lambda \log\left(\sum_{j=1}^{N} \exp\Big(-\frac{z_j}{\lambda}\Big)\right)$$

satisfies

$$\min_{j=1,\ldots,N} z_j - \lambda \log N \leq G_\lambda(\mathsf{z}) \leq \min_{j=1,\ldots,N} z_j.$$

The proof of Theorem 33 requires some tools on the topic of concentration of measures, as stated next.

Definition 9 (Subgaussian) *A random variable X is called σ-subgaussian if for all $\lambda \in \mathbb{R}$, its cumulant generating function satisfies*

$$\log \mathbb{E}\Big[e^{\lambda(X - \mathbb{E}X)}\Big] \leq \frac{\lambda^2 \sigma^2}{2}.$$

The next result is based on (Xu and Raginsky, 2017, Lemma 1), which itself can be derived from (Boucheron et al., 2013, Lemma 4.18). This result and its variants are key ingredients in information-theoretic generalization bounds.

Lemma 10 *If $g\big(\bar{A}, \bar{B}\big)$ is σ-subgaussian where $\big(\bar{A}, \bar{B}\big) \sim P_A P_B$, then for all $\lambda > 0$,*

$$\mathbb{E}\big[g\big(\bar{A}, \bar{B}\big)\big] - \mathbb{E}[g(A, B)] \leq \lambda(\log|\mathcal{A}| - H(A|B)) + \frac{\sigma^2}{2\lambda}.$$

The Azuma–Hoeffding inequality shows the subgaussianity of the sum of independent and bounded random variables:

Lemma 11 (Azuma–Hoeffding) *Let $X_1, \ldots, X_n$ be independent random variables such that $a \leq X_i \leq b$ for all $1 \leq i \leq n$. Then,*

$$\mathbb{E}\Big[e^{\frac{\lambda}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)}\Big] \leq \frac{\lambda^2}{2n}(b-a)^2.$$

In other words, $\sum_{i=1}^{n} X_i/n$ is $(b-a)/\sqrt{n}$-subgaussian.

Let $g : [a, b] \to \mathbb{R}$ be a differentiable function and assume that n is a positive integer. Let $\hat{\Delta} := (b-a)/n$. We define the Riemann sum at level n as

$$\Sigma_n := \hat{\Delta}\left(\sum_{i=1}^{n} g(a + (i-1)\hat{\Delta})\right).$$

The subsequent well-known lemma, which we use in Section 6, bounds the approximation error of the Riemann sum. For proof, see, for example, (Hughes-Hallett et al., 2020).

Lemma 12 (Riemann Sum) *The approximation error of the Riemann sum is bounded as follows:*

$$\left| \int_a^b g(x)\mathrm{d}x - \Sigma_n \right| \leq \frac{\hat{M}}{2n}(b-a)^2,$$

where $\hat{M}$ is the maximum absolute value of the derivative of g on $[a,b]$.

3. Ladder Decompositions of Functions

In this section, we first precisely define dilations and ladder decompositions of invertible functions. Then, we study Lipschitzness and smoothness of the components of the ladder decompositions of smooth functions $T : B_R^m \to \mathbb{R}^m$ defined on bounded Euclidean balls with radius $R > 0$. For simplicity, we first consider the special case $m = 1$ in Subsection 3.1 and investigate multi-dimensional functions in Subsection 3.2. We first provide the precise definition of dilations of a function.

Definition 13 *For any $0 < \gamma \leq 1$, let the* dilation *of function $T : B_R^m \to \mathbb{R}^m$ at scale γ be defined as*

$$T_{[\gamma]}(x) := \frac{T(\gamma x)}{\gamma}, \text{ for all } x \in B_R^m.$$

If T is differentiable and $T(0) = 0$, as a continuous extension for $\gamma = 0$, we define $T_{[0]}$ as the derivative of T at the origin. Namely, $T_{[0]}(x) := J_T(0)x$, where $J_T(0)$ is the Jacobian matrix of T at the origin.

The next lemma relates the inverse of the dilation with the dilation of the inverse function:

Lemma 14 *Let T be an invertible function. For any $0 < \gamma \leq 1$, we have $\left(T_{[\gamma]}\right)^{-1} = T_{[\gamma]}^{-1}$. Moreover, if T is differentiable and $T(0) = 0$, then $\left(T_{[0]}\right)^{-1} = T_{[0]}^{-1}$.*

Proof If $\gamma > 0$, then $\left(T_{[\gamma]}\right)^{-1}(x) = T^{-1}(\gamma x)/\gamma = T_{[\gamma]}^{-1}(x)$. The case $\gamma = 0$ follows from the well-known inverse function theorem; see, for example, (Baxandall and Liebeck, 1986, Chapter 4). ■

The next definition characterizes the concept of ladder decompositions:

Definition 15 *For all $1 \leq k \leq d$, let $T_k := T_{[\gamma_k]} \circ T_{[\gamma_{k-1}]}^{-1}$. We call the following multiscale decomposition the* ladder decomposition *of T at scale parameters $\{\gamma_k\}_{k=0}^d$:*

$$T = T_d \circ \cdots \circ T_1 \circ T_{[\gamma_0]}. \tag{9}$$

For all $1 \leq k \leq d$, we further define $\psi_k := T_k - \mathrm{id}$.

Clearly, for all $1 \leq k \leq d$, we have $T_{[\gamma_k]} = T_k \circ \cdots \circ T_1 \circ T_{[\gamma_0]}$. Owing to the smoothness of the function T, for all $1 \leq k \leq d$, when consecutive scale parameters γ_k and γ_{k-1} are close, we intuitively expect that the transformation T_k, acting between the dilations of T at scales γ_{k-1} and γ_k, be close to the identity function. In the rest of this section, we make this intuition precise, starting from the simpler case of one-dimensional functions ($m = 1$) and then studying the more general case of multi-dimensional functions ($m \geq 2$).

3.1 One-Dimensional Functions

We begin by providing the definition of smooth functions for one-dimensional functions.

Definition 16 *Let $\mathcal{V}$ be a bounded subset of $\mathbb{R}$. The one-dimensional function $T : \mathcal{V} \to \mathbb{R}$ is M-smooth if it is differentiable and its derivative T' is M-Lipschitz.*

If $T : \mathcal{V} \to \mathbb{R}$ is twice differentiable, then it is M-smooth if $|T''(x)| \leq M$ for all $x \in \mathcal{V}$, where T'' denotes the second derivative of T.

The next definition concerns the concept of diffeomorphisms:

Definition 17 *A function $T : (-R, R) \to \mathbb{R}$ is an (M_1, M_2)-diffeomorphism if it is invertible and both T and its inverse T^{-1} are twice differentiable, M_1-Lipschitz and M_2-smooth.*

If T is smooth, then one may expect $T_{[\gamma]}$ to get closer to $T_{[0]}$ as $\gamma \to 0$. The following preliminary result can be viewed as a formalization of this insight. Its proof is based on a simple extension of the proof of (Bartlett et al., 2018, Theorem 2).

Proposition 18 *Let $T : (-R, R) \to \mathbb{R}$ be a M_2-smooth function and $T(0) = 0$. Then, for any $0 \leq \gamma \leq 1$,*

$$\left\|T_{[\gamma]} - T_{[0]}\right\|_{\mathrm{Lip}} \leq \gamma M_2 R.$$

Proof The statement is trivial if $\gamma = 0$. Assume that $0 < \gamma \leq 1$, and let $x, y \in (-R, R)$ be arbitrarily chosen. Based on the mean value theorem, there exists z between x and y such that $T(\gamma x) - T(\gamma y) = \gamma T'(\gamma z)(x - y)$. We can write

$$\begin{aligned}
\left|(T_{[\gamma]}(x) - T_{[0]}(x)) - (T_{[\gamma]}(y) - T_{[0]}(y))\right| &= \left|\left(\frac{T(\gamma x)}{\gamma} - \frac{T(\gamma y)}{\gamma}\right) - T'(0)(x - y)\right| \\
&= \left|T'(\gamma z)(x - y) - T'(0)(x - y)\right| \\
&= |x - y||T'(\gamma z) - T'(0)| \\
&\leq |x - y||\gamma z|M_2 \\
&\leq |x - y|\gamma M_2 R,
\end{aligned}$$

which implies the statement. ■

Since $T(0) = 0$, then based on Proposition 18 we have

$$\left|(T_{[\gamma]}(x) - T_{[0]}(x))\right| \leq \gamma M_2 R|x|.$$

Thus, the smaller γ_k is in the ladder decomposition, the closer function $T_{[\gamma_k]}$ is to the linear function $T_{[0]}$.

The next result makes precise the intuition that if two subsequent scale parameters γ_k and γ_{k-1} are close, then $T_k := T_{[\gamma_k]} \circ T^{-1}_{[\gamma_{k-1}]}$ is a function close to the identity function:

Theorem 19 *Let $T : (-R, R) \to \mathbb{R}$ be an invertible M_2-smooth function such that T^{-1} is M_1-Lipschitz and $T(0) = 0$. Then, for any $0 \leq \gamma \leq \gamma' \leq 1$, we have*

$$\left\|T_{[\gamma']} \circ T^{-1}_{[\gamma]} - \mathrm{id}\right\|_{\mathrm{Lip}} \leq \left(\gamma' - \gamma\right) M_1 M_2 R.$$

If T is further assumed to be an (M_1, M_2)-diffeomorphism, then $T_{[\gamma']} \circ T^{-1}_{[\gamma]} - \mathrm{id}$ is $(M_1^2 + M_1)M_2$-smooth as well.

Proof The statement is trivial when $\gamma = \gamma'$. It is enough to prove that for any $0 \leq \gamma < \gamma' \leq 1$ and any x, y in the domain of $T_{[\gamma']} \circ T_{[\gamma]}^{-1}$, the following inequality holds:

$$\left|\Big(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(y) - y\Big) - \Big(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(x) - x\Big)\right| \leq \big(\gamma' - \gamma\big) M_1 M_2 R |y - x|.$$

Let $v := T_{[\gamma]}^{-1}(x)$ and $w := T_{[\gamma]}^{-1}(y)$. Based on Lemma 14, $x = T_{[\gamma]}(v)$ and $y = T_{[\gamma]}(w)$. According to Definition 13, the domain of $T_{[\gamma]}$ is $(-R, R)$, thus $\max\{|v|, |w|\} < R$. By defining the function $r := T_{[\gamma']} - T_{[\gamma]}$, we observe that

$$\begin{aligned}\Big(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(y) - y\Big) - \Big(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(x) - x\Big) &= \big(T_{[\gamma']}(w) - T_{[\gamma]}(w)\big) - \big(T_{[\gamma']}(v) - T_{[\gamma]}(v)\big) \\ &= r(w) - r(v).\end{aligned}$$

Note that the derivative of r at point v is equal to $r'(v) = T'(\gamma' v) - T'(\gamma v)$. Based on the mean value theorem, there exists $c \in \mathbb{R}$ between w and v such that

$$\begin{aligned}r(w) - r(v) &= (w - v) r'(c) \\ &= (w - v)(T'(\gamma' c) - T'(\gamma c)).\end{aligned}$$

Hence,

$$\begin{aligned}\left|\Big(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(y) - y\Big) - \Big(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(x) - x\Big)\right| &= |r(w) - r(v)| \\ &= |w - v| \big|T'(\gamma' c) - T'(\gamma c)\big| \\ &\leq |w - v| M_2 |\gamma' c - \gamma c| \qquad (10) \\ &= |w - v| M_2 |c| \big(\gamma' - \gamma\big) \\ &\leq |w - v| M_2 R \big(\gamma' - \gamma\big) \qquad (11) \\ &\leq |y - x| M_1 M_2 R \big(\gamma' - \gamma\big), \qquad (12)\end{aligned}$$

where (10) follows from T being M_2-smooth, (11) follows from $|c| \leq \max\{|v|, |w|\} < R$, and (12) is based on the assumption that T^{-1} is M_1-Lipschitz.

We now prove the smoothness property. Let $g(x) := T^{-1}(x)$ and $\psi := T_{[\gamma']} \circ T_{[\gamma]}^{-1} - \mathrm{id}$. The chain rule of derivatives yields

$$\psi''(x) = \gamma' T''\left(\frac{\gamma'}{\gamma} g(\gamma x)\right) \big(g'(\gamma x)\big)^2 + \gamma T'\left(\frac{\gamma'}{\gamma} g(\gamma x)\right) g''(\gamma).$$

Since T and g are both M_1-Lipschitz and M_2-smooth and $0 \leq \gamma, \gamma' \leq 1$, we deduce

$$|\psi''(x)| \leq (M_1^2 + M_1) M_2.$$

Therefore, ψ is $(M_1^2 + M_1) M_2$-smooth. ∎

Corollary 20 *Theorem 19 implies that for the ladder decomposition of T at scale parameters $\{\gamma_k\}_{k=0}^d$, the following inequality holds for all $1 \leq k \leq d$:*

$$\|\psi_k\|_{\mathrm{Lip}} \leq (\gamma_k - \gamma_{k-1}) M_1 M_2 R. \qquad (13)$$

The following is an example in which the functions ψ_k, $1 \leq k \leq d$, have a closed-form expression:

Example 1 Let $T(x) := \tanh(x)$ be the hyperbolic tangent function where it is known that $T^{-1}(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$ for $|x| < 1$. Assume that $\gamma_k = 2^{k-d}$ for all $0 \leq k \leq d$. For all $|x| < 1$, we can derive

$$\begin{aligned}
\psi_k(x) &= T_k(x) - x \\
&= T_{[\gamma_k]} \circ T^{-1}_{[\gamma_{k-1}]}(x) - x \\
&= \frac{\exp\left(2\gamma_k T^{-1}_{[\gamma_{k-1}]}(x)\right) - 1}{\gamma_k\left(\exp\left(2\gamma_k T^{-1}_{[\gamma_{k-1}]}(x)\right) + 1\right)} - x \\
&= \frac{\exp\left(\frac{\gamma_k}{\gamma_{k-1}}\ln\left(\frac{1+\gamma_{k-1}x}{1-\gamma_{k-1}x}\right)\right) - 1}{\gamma_k\left(\exp\left(\frac{\gamma_k}{\gamma_{k-1}}\ln\left(\frac{1+\gamma_{k-1}x}{1-\gamma_{k-1}x}\right)\right) + 1\right)} - x \\
&= \frac{\left(\frac{1+\gamma_{k-1}x}{1-\gamma_{k-1}x}\right)^{\frac{\gamma_k}{\gamma_{k-1}}} - 1}{\gamma_k\left(\left(\frac{1+\gamma_{k-1}x}{1-\gamma_{k-1}x}\right)^{\frac{\gamma_k}{\gamma_{k-1}}} + 1\right)} - x \\
&= \frac{2\gamma_{k-1}x}{\gamma_k\left(1 + \gamma_{k-1}^2 x^2\right)} - x \\
&= \frac{x}{1 + \gamma_{k-1}^2 x^2} - x \\
&= -\frac{\gamma_{k-1}^2 x^3}{1 + \gamma_{k-1}^2 x^2} \\
&= -\frac{1}{\gamma_{k-1}^{-2} x^{-3} + x^{-1}}.
\end{aligned}$$

Figure 1 depicts the plot of $\psi_k(x)$ for all $1 \leq k \leq d$, where $d = 5$. It clearly demonstrates that the Lipschitz norm of ψ_k increases with k.

3.2 Multi-Dimensional Functions

In this subsection, we extend the first part of Theorem 19 to multi-dimensional functions with $m \geq 2$. We commence by providing the definition of smooth multi-dimensional functions. The well-known extension of Definition 16 to real-valued functions of multiple variables is as follows:

Definition 21 *The scalar-valued function $g : B_R^m \to \mathbb{R}$ is M-smooth if it is differentiable, and its gradient ∇g is M-Lipschitz with respect to the Euclidean distance. Namely, for any $x, y \in B_R^m$, we have*

$$|\nabla g(x) - \nabla g(y)| \leq M|x - y|.$$

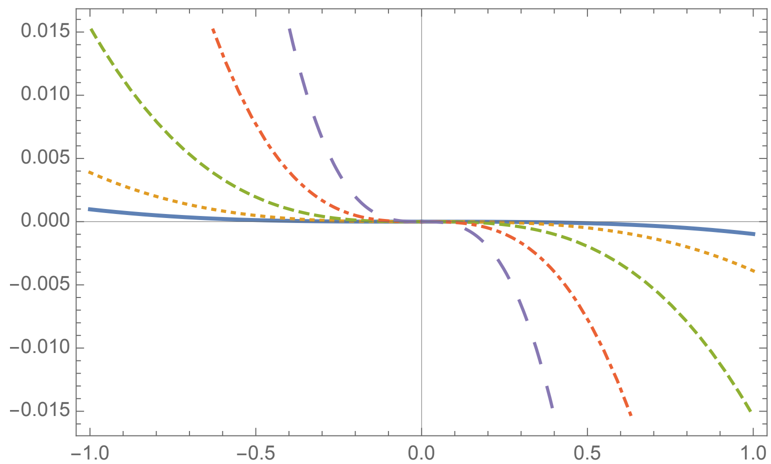

Figure 1: A plot of $\psi_k(x)$ for different values k: $k = 1$ the solid line, $k = 2$ the dotted line, $k = 3$ the dashed line, $k = 4$ the dotted-dashed line and $k = 5$ the large dashed line.

A twice differentiable function $g : B_R^m \to \mathbb{R}$ is M-smooth if the absolute value of any eigenvalues of its Hessian $H_g(x)$ is smaller than or equal to M, for all $x \in B_R^m$. In other words, if I denotes the identity matrix, then $-MI \preceq H_g(x) \preceq MI$ for all $x \in B_R^m$.

We now extend the previous definition to *vector-valued* functions:

Definition 22 *The multi-dimensional function $T : B_R^m \to \mathbb{R}^m$ is M-smooth if it is differentiable and its Jacobian J_T is M-Lipschitz with respect to the spectral distance. In other words, for any $x, y \in B_R^m$, the subsequent inequality holds:*

$$\|J_T(x) - J_T(y)\|_2 \le M|x - y|.$$

The next result shows the relation between Definitions 21 and 22.

Proposition 23 *Assume that $T : B_R^m \to \mathbb{R}^m$ is a function with components*

$$T(x) = (T_1(x), \dots, T_m(x)).$$

Then, the multi-dimensional function T is M-smooth if and only if each scalar-valued function $T_i : B_R^m \to \mathbb{R}$ is M-smooth, where $1 \le i \le m$.

Proof Let $x, y \in B_R^m$ be arbitrarily chosen. Assume that each function T_i is M-smooth for all $1 \le i \le m$. For any $u \in \mathbb{R}^m$, we observe

$$|(J_T(x) - J_T(y))u| \le \max_{1 \le i \le m} \left|(\nabla T_i(x) - \nabla T_i(y))^T u\right|.$$

Moreover, given any $1 \le i \le m$, the Cauchy-Schwartz inequality implies

$$\begin{aligned} \left|(\nabla T_i(x) - \nabla T_i(y))^T u\right| &\le |\nabla T_i(x) - \nabla T_i(y)||u| \\ &\le M|x - y||u|. \end{aligned}$$

Thus, for any $u \in \mathbb{R}^m$,

$$|(J_T(x) - J_T(y))u| \leq M|x - y||u|,$$

which implies that

$$\|J_T(x) - J_T(y)\|_2 \leq M|x - y|.$$

This proves the 'if' part. To prove the 'only if' part, assume that T is M-smooth. For any $1 \leq i \leq m$, let e_i be the ith standard unit vector. Then, for all $1 \leq i \leq m$,

$$\begin{aligned} \left|(\nabla T_i(x) - \nabla T_i(y))^T\right| &= |(J_T(x) - J_T(y))e_i| \\ &\leq \|J_T(x) - J_T(y)\|_2 |e_i| \\ &\leq M|x - y|, \end{aligned}$$

which implies that T_i is M-smooth. ■

The proof of Theorem 19 relied upon the mean value theorem. The following result is a multi-dimensional extension of the mean value theorem:

Lemma 24 *Let $r : B_R^m \to \mathbb{R}^m$ be a differentiable function. For any $a, b \in B_R^m$, there exists $\lambda \in (0, 1)$ such that for $c := \lambda b + (1 - \lambda)a \in B_R^m$, we have*

$$|r(b) - r(a)| \leq \|J_r(c)\|_2 |b - a|,$$

where $J_r(c)$ is the Jacobian of r at c.

Proof Define $\xi : [0, 1] \to \mathbb{R}$ as $\xi(t) := (r(b) - r(a))^T r(tb + (1 - t)a)$. Based on the mean value theorem, there exists $\lambda \in (0, 1)$ such that $\xi(1) - \xi(0) = \xi'(\lambda)$. Thus, according to the chain rule of derivatives, we can write

$$(r(b) - r(a))^T (r(b) - r(a)) = (r(b) - r(a))^T J_r(c)(b - a).$$

The Cauchy-Schwartz inequality yields

$$\begin{aligned} |r(b) - r(a)|^2 &\leq |r(b) - r(a)||J_r(c)(b - a)| \\ &\leq |r(b) - r(a)|\|J_r(c)\|_2 |b - a|, \end{aligned}$$

which implies the statement. ■

We now have sufficient tools to extend the first part of Theorem 19 to multi-dimensional functions.

Theorem 25 *Let $T : B_R^m \to \mathbb{R}^m$ be an invertible M_2-smooth function such that T^{-1} is M_1-Lipschitz and $T(0) = 0$. Then, for any $0 \leq \gamma \leq \gamma' \leq 1$, we have*

$$\left\| T_{[\gamma']} \circ T_{[\gamma]}^{-1} - \mathrm{id} \right\|_{\mathrm{Lip}} \leq \left(\gamma' - \gamma\right) M_1 M_2 R.$$

Proof The case $\gamma = \gamma'$ is easy to verify. Thus, it is enough to prove that for any $0 \le \gamma < \gamma' \le 1$ and any x, y in the domain of $T_{[\gamma']} \circ T_{[\gamma]}^{-1}$, we have

$$\left|\left(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(y) - y\right) - \left(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(x) - x\right)\right| \le (\gamma' - \gamma) M_1 M_2 R |y - x|.$$

Let $v := T_{[\gamma]}^{-1}(x)$ and $w := T_{[\gamma]}^{-1}(y)$. Based on Lemma 14, $x = T_{[\gamma]}(v)$ and $y = T_{[\gamma]}(w)$. According to Definition 13, the domain of $T_{[\gamma]}$ is B_R^m, thus $\max\{|v|, |w|\} < R$. Let $r := T_{[\gamma']} - T_{[\gamma]}$. We observe that

$$\begin{aligned}\left(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(y) - y\right) - \left(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(x) - x\right) &= \left(T_{[\gamma']}(w) - T_{[\gamma]}(w)\right) - \left(T_{[\gamma']}(v) - T_{[\gamma]}(v)\right) \\ &= r(w) - r(v).\end{aligned}$$

The Jacobian of function r at point z is equal to $J_r(z) = J_T(\gamma' z) - J_T(\gamma z)$, where J_T is the Jacobian of T, for all $z \in B_R^m$. According to Lemma 24, there exists $c \in \mathbb{R}^m$ as a convex combination of v and w such that

$$|r(w) - r(v)| \le \|J_r(c)\|_2 |w - v|.$$

Hence,

$$\begin{aligned}\left|\left(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(y) - y\right) - \left(T_{[\gamma']} \circ T_{[\gamma]}^{-1}(x) - x\right)\right| &= |r(w) - r(v)| \\ &\le |w - v| \|J_r(c)\|_2 \\ &= |w - v| \|J_T(\gamma' c) - J_T(\gamma c)\|_2 \\ &\le |w - v| M_2 |c| \big(\gamma' - \gamma\big) \qquad (14) \\ &\le |w - v| M_2 R \big(\gamma' - \gamma\big) \qquad (15) \\ &\le |y - x| M_1 M_2 R \big(\gamma' - \gamma\big), \qquad (16)\end{aligned}$$

where (14) follows from T being M_2-smooth, (15) is based on the fact that $|c| \le \max\{|v|, |w|\} < R$, and (16) follows from the assumption that T^{-1} is M_1-Lipschitz. ■

Corollary 26 *Similar to Corollary 20, Theorem 25 implies that for the ladder decomposition of multi-dimensional function T at scale parameters $\{\gamma_k\}_{k=0}^d$, the following inequality holds for all $1 \le k \le d$:*

$$\|\psi_k\|_{\mathrm{Lip}} \le (\gamma_k - \gamma_{k-1}) M_1 M_2 R. \qquad (17)$$

Remark 27 The proof of (Bartlett et al., 2018, Theorem 2) establishes that, for all $1 \le k \le d$, function ψ_k is $C(\gamma_k - \gamma_{k-1})/\gamma_k$-Lipschitz for some constant C. However, this implication is weaker than our result. For example, for each $1 \le k \le d$, when scale parameters are chosen as $\gamma_k = k/d$, then our result yields that $\|\psi_k\|_{\mathrm{Lip}} = O(1/d)$. However, (Bartlett et al., 2018, Theorem 2) states that there exist functions $T_1, \ldots, T_d$ such that $T = T_d \circ \cdots \circ T_1$ and for each $1 \le k \le d$, $\|T_k - \mathrm{id}\|_{\mathrm{Lip}} = O((\log d)/d)$.

Corollaries 20 and 26, when applied to the ladder decomposition of the target function T, are key results underpinning our learning model.

4. The Proposed Learning Model

In this section, we precisely formulate the learning model and discuss its strengths.

Assume that for $0 < \varepsilon < R$, the data instance domain is

$$\mathcal{X} := \{x \in \mathbb{R}^m : \varepsilon \leq |x| < R\}.$$

Typically, ε is much smaller than R. The label set is $\mathcal{Y} = \mathbb{R}^m$, and the target function T is an invertible M_2-smooth function such that T^{-1} is M_1-Lipschitz. Suppose that from previous knowledge or intuition, we know the behavior of function T at data instances in very small scales $0 \leq |x| < \varepsilon$. This is equivalent to knowing $T_{[\gamma_0]}$, where $\gamma_0 := R/\varepsilon$. Without loss of generality, we further assume that $T(0) = 0$.[1] For example, it may be assumed that $T_{[\gamma_0]}$ is a linear function equal to $T_{[0]}$, the derivative of T at the origin. Let $\gamma := (R/\varepsilon)^{1/d}$. We split the data domain $\mathcal{X}$ into d scales. For all $1 \leq k \leq d$, define

$$\mathcal{X}_k := \Big\{x \in \mathbb{R}^m : \varepsilon\gamma^{k-1} \leq |x| < \varepsilon\gamma^k\Big\}.$$

Clearly, $\mathcal{X} = \cup_{k=1}^d \mathcal{X}_k$, indicating that these sets form a partition of $\mathcal{X}$. Each set $\mathcal{X}_k$ is called the domain of instances at scale k. We define the scale parameters $\{\gamma_k\}_{k=0}^d$ to constitute a geometric sequence such that for all $0 \leq k \leq d$,

$$\gamma_k := \gamma^{k-d}. \tag{18}$$

Consider the ladder decomposition of T at scale parameters $\{\gamma_k\}_{k=0}^d$ (Definition 15). For all $1 \leq k \leq d$, we have

$$T_{[\gamma_k]} = T_k \circ T_{[\gamma_{k-1}]} = T_{[\gamma_{k-1}]} + \psi_k \circ T_{[\gamma_{k-1}]}. \tag{19}$$

To progressively learn the target function T on $\mathcal{X}$ through multiple stages, we introduce a d-level hierarchical learning model that aligns with the relation (19). We set $h_0 := T_{[\gamma_0]}$, and for all $1 \leq k \leq d$, define

$$h_k(x) := h_{k-1}(x) + f(h_{k-1}(x); w_k), \tag{20}$$

where f is a function dependent on the parametrization of the learning model. The choice of function f is important for ensuring enough representation power of the model, and we explore an example in Section 6.

Remark 28 One may rely on the universal approximation theorem for neural networks and assume that function f is a two-layer neural network with sufficient width; see a discussion in (Bartlett et al., 2018).

We denote the parameters of the kth level of the learning model with w_k, allowing it to take values from a finite set $\mathcal{W}_k$ during training. The training mechanism aims to make the mapping between the input and each layer h_k approximate $T_{[\gamma_k]}$, the dilated version of the

[1] In general, one can replace $T(\cdot)$ with $\bar{T}(\cdot) := T(\cdot) - T(0)$ and learn this new function. Clearly, $\bar{T}(0) = 0$.

target function T at scale γ_k. For a successfully trained model, $f(\cdot; w_k)$ should effectively approximate $\psi_k(\cdot)$. Let $C_1 := M_1 M_2 R$. According to Theorem 25, we have

$$\|\psi_k\|_{\mathrm{Lip}} = C_1(\gamma_k - \gamma_{k-1}) = C_1(\gamma - 1)\gamma^{k-d-1}. \tag{21}$$

Since $T(0) = 0$, we have $\psi_k(0) = 0$, for all $1 \le k \le d$. Thus, (21) implies that

$$|\psi_k(\cdot)| \le O(\gamma_k - \gamma_{k-1}) = O(\gamma^k). \tag{22}$$

Hence, we regularize our hypothesis set as follows:

$$|f(\cdot; w_k)| \le \rho_k, \text{ for all } w_k \in \mathcal{W}_k, \tag{23}$$

where ρ_k is at least the maximum value of $|\psi_k(\cdot)|$.

In a successfully trained model, each level h_k approximates $T_{[\gamma_k]}$. Knowing $T_{[\gamma_k]}$ is enough to determine the label of each instance $x \in \mathcal{X}_k$ with appropriate rescalings. Thus, we define the output of our model $h_{\mathsf{w}}(x)$ as follows:

$$h_{\mathsf{w}}(x) := \begin{cases} \gamma_1 h_1\left(\frac{x}{\gamma_1}\right) & \text{if} \quad x \in \mathcal{X}_1 \\ \gamma_2 h_2\left(\frac{x}{\gamma_2}\right) & \text{if} \quad x \in \mathcal{X}_2, \\ \quad\vdots & \\ h_d(x) & \text{if} \quad x \in \mathcal{X}_d. \end{cases}$$

Therefore, there is no necessity to propagate every data instance x through *all levels* of our hierarchical model. Rather, the processing steps required to evaluate $h_{\mathsf{w}}(x)$ for an input instance $x \in \mathcal{X}$ are proportional to the logarithm of the instance norm $|x|$, determining in which $\mathcal{X}_k$ does x belong to. Consequently, the logarithm of the norm of each data instance x can be construed as a metric for its difficulty or complexity.

Let the n-tuple of training instances be denoted by $\mathsf{s} = (x_1, \ldots, x_n)$. We assume the training set includes instance-label pairs $(x_i, T(x_i))$ for all $1 \le i \le n$. The training mechanism commences with the simplest training examples whose instances are at the smallest scale $\mathcal{X}_1$, and progressively trains the model's layers by using the larger-scaled (more complex) examples. At each level, corresponding to each scale of the data, training is represented by sampling w_k from a Gibbs measure with the loss (energy) as the empirical risk evaluated for that specific scale of the training data and with hyperparameter (temperature) λ_k. The temperature vector $(\lambda_k)_{k=1}^d$ consists of model hyperparameters that can be chosen based on cross-validation or by Corollary 34, which provides their values that optimize the derived bound on the statistical risk. It is well-known that Gibbs probability measures are maximum-entropy distributions; see (Jaynes, 1957), a property that we later use in the analysis. Precisely, for all $1 \le k \le d$, given trained values for $w_{1:(k-1)}$, we sample the vector value for w_k with the following probability:

$$\mathbb{P}_{W_k|W_{1:(k-1)}}\left(w_k \middle| w_{1:(k-1)}\right) := \frac{\exp\left(-\frac{1}{n\lambda_k}\sum_{x_i \in \mathsf{s}\cap\mathcal{X}_k}\left|\gamma_k h_k\left(\frac{x_i}{\gamma_k}; w_{1:k}\right) - T(x_i)\right|\right)}{\sum_{w'_k \in \mathcal{W}_k}\exp\left(-\frac{1}{n\lambda_k}\sum_{x_i \in \mathsf{s}\cap\mathcal{X}_k}\left|\gamma_k h_k\left(\frac{x_i}{\gamma_k}; w_{1:(k-1)}w'_k\right) - T(x_i)\right|\right)}. \tag{24}$$

This training mechanism is hierarchical and stochastic. Furthermore, it exhibits self-similarity, as at all steps, we sample from Gibbs distributions consisting of loss functions with similar absolute error forms, albeit on data at different scales and with different temperatures. We refer to this training mechanism as *multiscale entropy-based training*, see Algorithm 1.

Algorithm 1 Multiscale entropy-based training

Hyperparameters: Temperature vector $(\lambda_k)_{k=1}^d$.

Input: Training data $(x_i, T(x_i))_{i=1}^n$.

Output: Trained parameters $w_1, \ldots, w_d$.

for $k = 1$ `to` d **do**

Given training data at scale k ($\mathsf{s} \cap \mathcal{X}_k$) and sampled values for $w_{1:(k-1)}$, obtain w_k by sampling from the Gibbs measure:

$$\mathbb{P}_{W_k|W_{1:(k-1)}}\big(w_k \big| w_{1:(k-1)}\big) \propto \exp\Big(-\tfrac{1}{n\lambda_k} \textstyle\sum_{x_i \in \mathsf{s} \cap \mathcal{X}_k} \Big|\gamma_k h_k\Big(\tfrac{x_i}{\gamma_k}; w_{1:k}\Big) - T(x_i)\Big|\Big).$$

The proposed learning model and the forthcoming analysis of its statistical risk possess the following advantages:

1. **Hierarchical learning model with interpretable levels:** The training objective is to ensure each level h_k approximates the dilation of the target function T at scale γ_k, that is, $T_{[\gamma_k]}$. Thus, each level in our hierarchical learning model holds a meaningful interpretation—a departure from black-box hierarchical models such as commonly used neural networks. In other words, the mapping between the input of the whole compositional model

$$(\mathrm{id} + f(\cdot; w_d)) \circ \cdots \circ (\mathrm{id} + f(\cdot; w_1)) \circ T_{[\gamma_0]},$$

with any of its intermediate levels is aimed to be close to dilations of the original target function at different scales.

2. **Measuring the complexity of any data instance with the logarithm of its norm:** For any data instance x, the logarithm of its norm $\log|x|$ can be interpreted as a measure of its complexity. This complexity determines the training and inference stage at which the label of x can be predicted in the learning model. Thus, our learning model can also be observed as a mathematical model for curriculum learning. For illustrative examples of norm-based complexity interpretations, consider the following scenarios: In finance, predicting fraud becomes increasingly challenging as a company's revenue resources grow. In image processing, higher degrees of sparsity lead to lower norms in feature vectors, making their targets easier to predict.

3. **Computational savings in inference:** Assume that $x \in \mathcal{X}_k$ is a new data instance that we want to predict its label using our model. To compute the model's output for x and predict its label, it is sufficient to process x/γ_k only up to the kth level and calculate $\gamma_k h_k(x/\gamma_k)$. In other words, there is no necessity to pass the instance through all d levels of the learning model, and processing only the first k levels is adequate, in contrast to commonly used neural networks. This efficiency stems from

the fact that, on the one hand, h_k approximates the dilation $T_{[\gamma_k]}$. On the other hand, given that $|x| < \gamma_k$ and with proper rescaling, one can compute the value of $T(x)$ just by knowing $T_{[\gamma_k]}$. More precisely, we have $T(x) = \gamma_k T_{[\gamma_k]}(x/\gamma_k)$ whenever $|x| < \gamma_k$. In practical terms, when using the trained model to predict the target of new data, the computational workload for computing the output of the learned model with a particular data instance as input is directly tied to the complexity of that instance. Since, by assumption (such as scale-invariant distributions (1)), data instances are distributed heterogeneously across different scales and difficulties, this characteristic may lead to substantial computational savings and an increased inference speed.

4. **Statistical guarantee stronger than uniform convergence:** The statistical analysis of the risk of the trained compositional model is tailored to the hierarchical training mechanism and takes its multiscale structure into account when deriving the bound on its statistical risk. Consequently, the bound can be much sharper than a uniform convergence bound for the classical empirical-risk-minimizing training algorithm, as shown in the next section.

5. **Robustness to interruptions during training:** The training of current hierarchical learning models (such as neural networks) on massive datasets may take extensive amounts of time and are prone to interruptions. Our sequential training mechanism consists of d stages. If, for any reason, this mechanism terminates prematurely after stage k for any $1 \leq k \leq d-1$, we can still ensure the availability of a useful model. More precisely, this model remains capable of accurately predicting the labels of data instances belonging to $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_k$, which are all $x \in \mathbb{R}^m$ such that $\varepsilon \leq |x| < \varepsilon\gamma^k$.

6. **Computational savings in training for diverse users:** The proposed learning model can provide computational savings when there are d different users, each requiring accurate prediction of the labels of data instances at scale k (set $\mathcal{X}_k$) for $k = 1, \ldots, d$. Instead of training separate models for each of the d users, we streamline the process by using the trained model for user $k-1$ as 'prior information' in training the model for user k, for all $2 \leq k \leq d$.

5. Statistical Analysis of the Learning Model

In this section, we statistically analyze the learning model's performance. In Subsection 5.1, we prove an upper bound on the training mechanism's chained risk. Then, in Subsection 5.2, we show that if the data instance distribution μ is scale-invariant, we can bound the statistical risk from above based on the chained risk.

5.1 Multiscale Entropy-Based Training and the Chained Risk

For all $1 \leq k \leq d$, we define the loss at scale k as

$$\ell_k(w_{1:k}, x) := \begin{cases} |h_{\mathsf{w}}(x) - T(x)| = \left|\gamma_k h_k\left(\frac{x}{\gamma_k}, w_{1:k}\right) - T(x)\right| & \text{if } x \in \mathcal{X}_k \\ 0 & \text{if } x \notin \mathcal{X}_k. \end{cases} \tag{25}$$

Furthermore, suppose that

$$\ell_k(w_{1:k}, \mathsf{s}) := \frac{1}{n}\sum_{i=1}^{n} \ell_k(w_{1:k}, x_i).$$

Equation (24) in the previous section can be written more succinctly as

$$\mathbb{P}_{W_k|W_{1:(k-1)}}\big(w_k|w_{1:(k-1)}\big) = \frac{\exp\Big(-\frac{1}{\lambda_k}\ell_k(w_{1:k}, \mathsf{s})\Big)}{\sum_{w'_k\in\mathcal{W}_k}\exp\Big(-\frac{1}{\lambda_k}\ell_k\big(w_{1:(k-1)}w'_k, \mathsf{s}\big)\Big)}.$$

We denote the whole random parameters as $\mathsf{W} := (W_1, \ldots, W_d)$. The measure

$$\mathbb{P}^\star_{\mathsf{W}} := \mathbb{P}_{W_1}\mathbb{P}_{W_2|W_1}\ldots\mathbb{P}_{W_d|W_{1:(d-1)}}$$

models the total training mechanism. For ease of presentation, define the extra parameter $\lambda_{d+1} := 0$. The next theorem indicates that the self-similar measure $\mathbb{P}^\star_{\mathsf{W}}$ is the minimizing distribution of the sum of a multiscale loss and a multiscale entropy. Consider the definition of the multiscale loss

$$\ell^{(\lambda)}(\mathsf{w}, \mathsf{s}) := \sum_{k=1}^{d}\Big(\ell_k(w_{1:k}, \mathsf{s}) - \bar{\ell}_k\big(w_{1:(k-1)}, \mathsf{s}\big)\Big), \tag{26}$$

where, for all $1 \le k \le d$,

$$\bar{\ell}_k\big(w_{1:(k-1)}, \mathsf{s}\big) := -\lambda_k \log\left(\sum_{w'_k\in\mathcal{W}_k}\exp\left(-\frac{1}{\lambda_k}\ell_k\big(w_{1:(k-1)}w'_k, \mathsf{s}\big)\right)\right) \tag{27}$$

is a Log-Sum-Exp function.

Theorem 29 *We have*

$$\mathbb{P}^\star_{\mathsf{W}} = \arg\min_{P_{\mathsf{W}}}\left\{\mathbb{E}\Big[\ell^{(\lambda)}(\mathsf{W}, \mathsf{s})\Big] - \sum_{k=1}^{d}(\lambda_k - \lambda_{k+1})H(W_{1:k})\right\}. \tag{28}$$

Proof We develop what we call the 'multiscale congruent technique' used in the proof of (Asadi and Abbe, 2020, Theorem 13). Specifically, recalling the definition of congruent functionals (Definition 6), we aim to show that, as a functional of P_{W},

$$\mathbb{E}\Big[\ell^{(\lambda)}(\mathsf{W}, \mathsf{s})\Big] - \sum_{k=1}^{d}(\lambda_k - \lambda_{k+1})H(W_{1:k}) \cong \sum_{k=1}^{d}\lambda_k D\Big(P_{W_k|W_{1:(k-1)}}\Big\|\mathbb{P}_{W_k|W_{1:(k-1)}}\Big|P_{W_{1:(k-1)}}\Big). \tag{29}$$

This will then immediately imply (28), as setting $P_{\mathsf{W}} = \mathbb{P}^\star_{\mathsf{W}}$ makes the entropies in the right side of (29) vanish altogether. For all $1 \le k \le d$, define

$$L_k(w_{1:k}, \mathsf{s}) := \sum_{j=1}^{k}\ell_j(w_{1:j}, \mathsf{s}),$$

and

$$Q^{(k)}_{W_{1:k}}(w_{1:k}) := \frac{\exp\left(-\frac{1}{\lambda_k}L_k(w_{1:k},\mathsf{s})\right)}{\sum_{w'_{1:k}\in\mathcal{W}_1\times\cdots\times\mathcal{W}_k}\exp\left(-\frac{1}{\lambda_k}L_k\left(w'_{1:k},\mathsf{s}\right)\right)}.$$

Recall that $U_{W_{1:d}}$ denotes the equiprobable distribution. We can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{d}\ell_j(w_{1:j},\mathsf{s})\right]+\sum_{k=1}^{d}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})\\
&=\sum_{k=1}^{d-1}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})+\left(\mathbb{E}[L_d(W_{1:d},\mathsf{s})]+\lambda_d D(P_{W_{1:d}}\|U_{W_{1:d}})\right)\\
&\cong\sum_{k=1}^{d-1}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})+\lambda_d D\left(P_{W_{1:d}}\middle\|Q^{(d)}_{W_{1:d}}\right) &(30)\\
&=\sum_{k=1}^{d-1}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})\\
&\quad+\lambda_d D\left(P_{W_{1:(d-1)}}\middle\|Q^{(d)}_{W_{1:(d-1)}}\right)+\lambda_d D\left(P_{W_d|W_{1:(d-1)}}\middle\|Q^{(d)}_{W_d|W_{1:(d-1)}}\middle|P_{W_{1:(d-1)}}\right) &(31)\\
&=\sum_{k=1}^{d-2}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})\\
&\quad+\left((\lambda_{d-1}-\lambda_d)D\left(P_{W_{1:(d-1)}}\middle\|U_{W_{1:(d-1)}}\right)+\lambda_d D\left(P_{W_{1:(d-1)}}\middle\|Q^{(d)}_{W_{1:(d-1)}}\right)\right)\\
&\quad+\lambda_d D\left(P_{W_d|W_{1:(d-1)}}\middle\|Q^{(d)}_{W_d|W_{1:(d-1)}}\middle|P_{W_{1:(d-1)}}\right)\\
&\cong\sum_{k=1}^{d-2}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})+\lambda_{d-1}D\left(P_{W_{1:(d-1)}}\middle\|\left(Q^{(d)}_{W_{1:(d-1)}}\right)^{\frac{\lambda_d}{\lambda_{d-1}}}\right)\\
&\quad+\lambda_d D\left(P_{W_d|W_{1:(d-1)}}\middle\|Q^{(d)}_{W_d|W_{1:(d-1)}}\middle|P_{W_{1:(d-1)}}\right), &(32)
\end{aligned}$$

where (30) is based on the Gibbs variational principle (Lemma 7), (31) is based on the chain rule of entropy (Lemma 3), and (32) is based on Lemma 5 on entropy combination. To reiterate, we have derived

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j=1}^{d}\ell_j(w_{1:j},\mathsf{s})\right]+\sum_{k=1}^{d}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})\\
&\cong\sum_{k=1}^{d-2}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})+\lambda_{d-1}D\left(P_{W_{1:(d-1)}}\middle\|\left(Q^{(d)}_{W_{1:(d-1)}}\right)^{\frac{\lambda_d}{\lambda_{d-1}}}\right)\\
&\quad+\lambda_d D\left(P_{W_d|W_{1:(d-1)}}\middle\|Q^{(d)}_{W_d|W_{1:(d-1)}}\middle|P_{W_{1:(d-1)}}\right). &(33)
\end{aligned}$$

Marginalizing the distribution $Q^{(d)}_{W_{1:d}}$ over W_d yields

$$Q^{(d)}_{W_{1:(d-1)}} \propto \exp\Big(-\frac{1}{\lambda_d} L_{d-1}\big(w_{1:(d-1)}, \mathsf{s}\big)\Big) \left(\sum_{w'_d \in \mathcal{W}_d} \exp\Big(-\frac{1}{\lambda_d} \ell_d\big(w_{1:(d-1)} w'_d, \mathsf{s}\big)\Big)\right).$$

Thus, its escort distribution is

$$\Big(Q^{(d)}_{W_{1:(d-1)}}\Big)^{\frac{\lambda_d}{\lambda_{d-1}}} \propto \exp\Big(-\frac{1}{\lambda_{d-1}} L_{d-1}\big(w_{1:(d-1)}, \mathsf{s}\big)\Big) \left(\sum_{w'_d \in \mathcal{W}_d} \exp\Big(-\frac{1}{\lambda_d} \ell_d\big(w_{1:(d-1)} w'_d, \mathsf{s}\big)\Big)\right)^{\frac{\lambda_d}{\lambda_{d-1}}}.$$

Let Z and $\bar{Z}$ denote the normalizing constants (partition functions) of $\Big(Q^{(d)}_{W_{1:(d-1)}}\Big)^{\frac{\lambda_d}{\lambda_{d-1}}}$ and $Q^{(d-1)}_{W_{1:(d-1)}}$, respectively. We have

$$\begin{aligned}
&D\Big(P_{W_{1:(d-1)}} \Big\| Q^{(d-1)}_{W_{1:(d-1)}}\Big) - D\left(P_{W_{1:(d-1)}} \middle\| \Big(Q^{(d)}_{W_{1:(d-1)}}\Big)^{\frac{\lambda_d}{\lambda_{d-1}}}\right) \\
&= \sum_{w_{1:(d-1)}} \log\left(\frac{\Big(Q^{(d)}_{W_{1:(d-1)}}\Big)^{\frac{\lambda_d}{\lambda_{d-1}}}\big(w_{1:(d-1)}\big)}{Q^{(d-1)}_{W_{1:(d-1)}}\big(w_{1:(d-1)}\big)}\right) P_{W_{1:(d-1)}}\big(w_{1:(d-1)}\big) \\
&= \sum_{w_{1:(d-1)}} \log\left(\frac{\bar{Z}}{Z}\left(\sum_{w'_d \in \mathcal{W}_d} \exp\left(-\frac{\ell_d\big(w_{1:(d-1)} w'_d, \mathsf{s}\big)}{\lambda_d}\right)\right)^{\frac{\lambda_d}{\lambda_{d-1}}}\right) P_{W_{1:(d-1)}}\big(w_{1:(d-1)}\big) \\
&= \frac{\lambda_d}{\lambda_{d-1}} \sum_{w_{1:(d-1)}} \log\left(\sum_{w'_d} \exp\Big(-\frac{1}{\lambda_d} \ell_d\big(w_{1:(d-1)} w'_d, \mathsf{s}\big)\Big)\right) P_{W_{1:(d-1)}}\big(w_{1:(d-1)}\big) \\
&\quad + \sum_{w_{1:(d-1)}} \log\Big(\frac{\bar{Z}}{Z}\Big) P_{W_{1:(d-1)}}\big(w_{1:(d-1)}\big) \\
&= \frac{\lambda_d}{\lambda_{d-1}} \sum_{w_{1:(d-1)}} \log\left(\sum_{w'_d} \exp\Big(-\frac{1}{\lambda_d} \ell_d\big(w_{1:(d-1)} w'_d, \mathsf{s}\big)\Big)\right) P_{W_{1:(d-1)}}\big(w_{1:(d-1)}\big) + \log\Big(\frac{\bar{Z}}{Z}\Big) \\
&\cong \frac{\lambda_d}{\lambda_{d-1}} \sum_{w_{1:(d-1)}} \log\left(\sum_{w'_d} \exp\Big(-\frac{1}{\lambda_d} \ell_d\big(w_{1:(d-1)} w'_d, \mathsf{s}\big)\Big)\right) P_{W_{1:(d-1)}}\big(w_{1:(d-1)}\big) \\
&= \frac{\lambda_d}{\lambda_{d-1}} \mathbb{E}\left[\log\left(\sum_{w'_d} \exp\Big(-\frac{1}{\lambda_d} \ell_d\big(w_{1:(d-1)} w'_d, \mathsf{s}\big)\Big)\right)\right].
\end{aligned}$$

Therefore,

$$\lambda_{d-1}\left(D\Big(P_{W_{1:(d-1)}}\Big\|Q^{(d-1)}_{W_{1:(d-1)}}\Big) - D\left(P_{W_{1:(d-1)}}\middle\|\left(Q^{(d)}_{W_{1:(d-1)}}\right)^{\frac{\lambda_d}{\lambda_{d-1}}}\right)\right)$$
$$\cong \lambda_d\mathbb{E}\left[\log\left(\sum_{w'_d}\exp\Big(-\frac{1}{\lambda_d}\ell_d\big(W_{1:(d-1)}w'_d,\mathsf{s}\big)\Big)\right)\right] \tag{34}$$

By adding both sides of (33) and (34), we obtain

$$\mathbb{E}\left[\sum_{j=1}^{d}\ell_j(w_{1:j},\mathsf{s})\right]+\sum_{k=1}^{d}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})$$
$$+\lambda_d\mathbb{E}\left[\log\left(\sum_{w'_d}\exp\Big(-\frac{1}{\lambda_d}\ell_d\big(W_{1:(d-1)}w'_d,\mathsf{s}\big)\Big)\right)\right]$$
$$\cong\left(\sum_{k=1}^{d-2}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})+\lambda_{d-1}D\Big(P_{W_{1:(d-1)}}\Big\|\Big(Q^{(d-1)}_{W_{1:(d-1)}}\Big)\Big)\right)$$
$$+\lambda_d D\Big(P_{W_d|W_{1:(d-1)}}\Big\|Q^{(d)}_{W_d|W_{1:(d-1)}}\Big|P_{W_{1:(d-1)}}\Big).$$

By iterating this argument for $k=d-1,\ldots,1$, we deduce that

$$\mathbb{E}\Big[\ell^{(\lambda)}(\mathsf{W},\mathsf{s})\Big]+\sum_{k=1}^{d}(\lambda_k-\lambda_{k+1})D(P_{W_{1:k}}\|U_{W_{1:k}})$$
$$\cong\sum_{k=1}^{d}\lambda_k D\Big(P_{W_k|W_{1:(k-1)}}\Big\|Q^{(k)}_{W_k|W_{1:(k-1)}}\Big|P_{W_{1:(k-1)}}\Big). \tag{35}$$

For all $1\leq k\leq d$, it is apparent that

$$D(P_{W_{1:k}}\|U_{W_{1:k}})=\log(|\mathcal{W}_1\times\cdots\times\mathcal{W}_k|)-H(W_{1:k})\cong -H(W_{1:k}).$$

Thus, based on (35), we obtain

$$\mathbb{E}\Big[\ell^{(\lambda)}(\mathsf{W},\mathsf{s})\Big]-\sum_{k=1}^{d}(\lambda_k-\lambda_{k+1})H(W_{1:k})\cong\sum_{k=1}^{d}\lambda_k D\Big(P_{W_k|W_{1:(k-1)}}\Big\|Q^{(k)}_{W_k|W_{1:(k-1)}}\Big|P_{W_{1:(k-1)}}\Big).$$

Since, for all $1\leq k\leq d$,

$$Q^{(k)}_{W_k|W_{1:(k-1)}}\big(w_k\big|w_{1:(k-1)}\big)=\mathbb{P}_{W_k|W_{1:(k-1)}}\big(w_k\big|w_{1:(k-1)}\big),$$

we deduce that

$$\mathbb{E}\Big[\ell^{(\lambda)}(\mathsf{W},\mathsf{s})\Big]-\sum_{k=1}^{d}(\lambda_k-\lambda_{k+1})H(W_{1:k})\cong\sum_{k=1}^{d}\lambda_k D\Big(P_{W_k|W_{1:(k-1)}}\Big\|\mathbb{P}_{W_k|W_{1:(k-1)}}\Big|P_{W_{1:(k-1)}}\Big),$$

as desired. ■

Straightforwardly, the result extends for *random* training sequences $\mathsf{S}:=(X_1,\ldots,X_n)\sim\mu^{\otimes n}$.

Corollary 30 *Assume that for all $1 \leq k \leq d$,*

$$\mathbb{P}^{\star}_{W_k|W_{1:(k-1)}\mathsf{S}}\big(w_k\big|w_{1:(k-1)}\mathsf{s}\big) := \frac{\exp\Big(-\frac{1}{\lambda_k}\ell_k(w_{1:k},\mathsf{s})\Big)}{\sum_{w'_k\in\mathcal{W}_k}\exp\Big(-\frac{1}{\lambda_k}\ell_k\big(w_{1:(k-1)}w'_k,\mathsf{s}\big)\Big)}.$$

Let $\mathbb{P}^{\star}_{\mathsf{W}|\mathsf{S}} = \mathbb{P}^{\star}_{W_1|\mathsf{S}}\mathbb{P}^{\star}_{W_2|W_1\mathsf{S}}\dots\mathbb{P}^{\star}_{W_d|W_{1:(d-1)}\mathsf{S}}$. *Then,*

$$\mathbb{P}^{\star}_{\mathsf{W}|\mathsf{S}} = \arg\min_{P_{\mathsf{W}|\mathsf{S}}}\left\{\mathbb{E}\Big[\ell^{(\lambda)}(\mathsf{W},\mathsf{S})\Big] - \sum_{k=1}^{d}(\lambda_k-\lambda_{k+1})H(W_{1:k}|\mathsf{S})\right\}, \tag{36}$$

where $(\mathsf{S},\mathsf{W}) \sim \mu^{\otimes n}P_{\mathsf{W}|\mathsf{S}}$.

Proof The expression in (36) is linear in $P_{\mathsf{S}} = \mu^{\otimes n}$. Conditioned on any realization $\mathsf{S} = \mathsf{s}$, we can use Theorem 29 to yield the result. ■

Recall the definition of the function ψ_k, for all $1 \leq k \leq d$, in the ladder decomposition (19). We assume that our hypothesis set is *realizable*, meaning that there exist parameters $\hat{\mathsf{w}} = (\hat{w}_1, \dots, \hat{w}_d)$ in our hypothesis index set such that, for all $1 \leq k \leq d$,

$$\psi_k(\cdot) = f(\cdot;\hat{w}_k). \tag{37}$$

In other words, we assume that function T belongs to the hypothesis set and is represented with parameters $\hat{\mathsf{w}}$ as $h_{\hat{\mathsf{w}}} = T$. In case T is represented with different parameterizations; we choose *one* of these and denote it with $(\hat{w}_1, \dots, \hat{w}_d)$. Now, consider the definition of the chained risk:

Definition 31 (Chained Risk) *Let $X \sim \mu$. For any $\mathsf{w} \in \mathcal{W}$, we define the chained risk of w as*

$$L^{(\mathsf{C})}_{\mu}(\mathsf{w}) := \mathbb{E}\left[\sum_{k=1}^{d}\Big(\ell_k(w_{1:k},X) - \ell_k\big(w_{1:(k-1)}\hat{w}_k,X\big)\Big)\right]. \tag{38}$$

The training mechanism of the learning model chooses the values of the parameters $w_1, ..., w_d$ sequentially. At the kth stage, choosing w_k instead of the true target function parameter $\hat{w}_k$ results in the following difference between the risks at scale k:

$$\mathbb{E}\big[\ell_k(w_{1:k},X) - \ell_k\big(w_{1:(k-1)}\hat{w}_k,X\big)\big].$$

The chained risk is equal to the accumulation of these deviations of risk at each of the d stages of the training mechanism. Obviously, we have $L^{(\mathsf{C})}_{\mu}(\hat{\mathsf{w}}) = 0$. In an intuitive and approximate sense, if the chained risk of w is small, then it suggests that h_{w} could be close to the target function $h_{\hat{\mathsf{w}}} = T$.

It is informative to compare the chained risk with the statistical risk (6). Based on the definition of $\ell_k(w_{1:k},x)$ in (25) for all $1 \leq k \leq d$, the loss of the model with parameters w

on any data instance $x \in \mathcal{X}$ is equal to $\ell(\mathsf{w}, x) = \sum_{k=1}^{d} \ell_k(w_{1:k}, x)$. Assume that $X \sim \mu$. The statistical risk of the model with parameters $\mathsf{w} \in \mathcal{W}$ is

$$\begin{aligned} L_\mu(\mathsf{w}) &:= \mathbb{E}[\ell(\mathsf{w}, X)] \\ &= \mathbb{E}\left[\sum_{k=1}^{d} \ell_k(w_{1:k}, X)\right] \\ &= \mathbb{E}\left[\sum_{k=1}^{d} (\ell_k(w_{1:k}, X) - \ell_k(\hat{w}_{1:k}, X))\right], \end{aligned} \tag{39}$$

where (39) is based on the realizability assumption $L_\mu(\hat{\mathsf{w}}) = 0$. The difference between the right sides of (38) and (39) is between the terms $\ell_k(w_{1:(k-1)}\hat{w}_k, X)$ and $\ell_k(\hat{w}_{1:k}, X)$. In fact, we have

$$\begin{aligned} L_\mu^{(\mathsf{C})}(\mathsf{w}) &= \mathbb{E}\left[\sum_{k=1}^{d} \ell_k(w_{1:k}, X)\right] - \mathbb{E}\left[\sum_{k=1}^{d} \ell_k\big(w_{1:(k-1)}\hat{w}_k, X\big)\right] \\ &= L_\mu(\mathsf{w}) - \mathbb{E}\left[\sum_{k=1}^{d} \ell_k\big(w_{1:(k-1)}\hat{w}_k, X\big)\right] \\ &\leq L_\mu(\mathsf{w}). \end{aligned}$$

Hence, the chained risk is always smaller than or equal to the statistical risk. However, in the next subsection, for *scale-invariant* data instance distributions μ with sufficiently large shape parameters, we find a reverse form of this inequality: an *upper bound* on the statistical risk based on the chained risk.

Remark 32 The difference between the chained risk and the statistical risk can be large when the data instance distribution μ is not 'multiscale.' For example, assume that the data instance distribution μ is supported only on $\mathcal{X}_d$. Consider parameters $\mathsf{w} = (w_1, \ldots, w_{d-1}, \hat{w}_d)$. We claim that the chained risk of w is zero. Since the probability that $X \sim \mu$ takes values on $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_{d-1}$ is zero, for any $1 \leq k \leq d-1$ we have

$$\mathbb{E}[\ell_k(w_{1:k}, X)] = \mathbb{E}\big[\ell_k\big(w_{1:(k-1)}\hat{w}_k, X\big)\big] = 0.$$

Therefore,

$$L_\mu^{(\mathsf{C})}(\mathsf{w}) = \mathbb{E}[\ell_d(w_{1:d}, X)] - \mathbb{E}\big[\ell_d\big(w_{1:(d-1)}\hat{w}_d, X\big)\big] = 0.$$

On the other hand, we have

$$L_\mu(\mathsf{w}) = \mathbb{E}\left[\sum_{k=1}^{d} \ell_k(w_{1:k}, X)\right] = \mathbb{E}[\ell_d(w_{1:d}, X)].$$

However, if the parameters $w_1, \ldots, w_{d-1}$ are chosen significantly far from the optimal values $\hat{w}_1, \ldots, \hat{w}_{d-1}$, then h_{w} can be far from the target function T and $\mathbb{E}[\ell_d(w_{1:d}, X)]$ tends to be large, thereby resulting in a high statistical risk.

In the next theorem, we derive an upper bound on the expected value of the chained risk $\mathbb{E}\Big[L_{\mu}^{(\mathsf{C})}(\mathsf{W})\Big]$ for the output of the training mechanism $\mathbb{P}^{\star}_{\mathsf{W}|\mathsf{S}}$. Recall the definition of $\{\rho_k\}_{k=1}^{d}$ when regularizing the hypothesis set in (23).

Theorem 33 *Let* $(\mathsf{S},\mathsf{W}) \sim P_{\mathsf{S}}\mathbb{P}^{\star}_{\mathsf{W}|\mathsf{S}}$. *Then,*

$$\mathbb{E}\Big[L_{\mu}^{(\mathsf{C})}(\mathsf{W})\Big] \leq 2\sum_{k=1}^{d}\left((\lambda_k-\lambda_{k+1})\left(\sum_{j=1}^{k}\log|\mathcal{W}_j|\right)+\frac{4\gamma_k^2\rho_k^2}{n(\lambda_k-\lambda_{k+1})}\right).$$

Proof Let

$$\ell^{(\mathsf{C})}(\mathsf{w},\mathsf{s}) := \sum_{k=1}^{d}\Big(\ell_k(w_{1:k},\mathsf{s})-\ell_k\big(w_{1:(k-1)}\hat{w}_k,\mathsf{s}\big)\Big).$$

Recall from (26) and (27) that

$$\ell^{(\lambda)}(\mathsf{w},\mathsf{s}) := \sum_{k=1}^{d}\Big(\ell_k(w_{1:k},\mathsf{s})-\bar{\ell}_k\big(w_{1:(k-1)},\mathsf{s}\big)\Big),$$

where, for all $1\leq k\leq d$,

$$\bar{\ell}_k\big(w_{1:(k-1)},\mathsf{s}\big) := -\lambda_k\log\left(\sum_{w'_k\in\mathcal{W}_k}\exp\left(-\frac{1}{\lambda_k}\ell_k\big(w_{1:(k-1)}w'_k,\mathsf{s}\big)\right)\right).$$

For all $1\leq k\leq d$, $\bar{\ell}_k\big(w_{1:(k-1)},\mathsf{s}\big)$ is a Log-Sum-Exp function, thus Lemma 8 yields

$$\begin{aligned}
\ell^{(\mathsf{C})}(\mathsf{w},\mathsf{s}) &= \sum_{k=1}^{d}\Big(\ell_k(w_{1:k},\mathsf{s})-\ell_k\big(w_{1:(k-1)}\hat{w}_k,\mathsf{s}\big)\Big)\\
&\leq \sum_{k=1}^{d}\left(\ell_k(w_{1:k},\mathsf{s})-\min_{w'_k\in\mathcal{W}_k}\ell_k\big(w_{1:(k-1)}w'_k,\mathsf{s}\big)\right)\\
&\leq \sum_{k=1}^{d}\Big(\ell_k(w_{1:k},\mathsf{s})-\bar{\ell}_k\big(w_{1:(k-1)},\mathsf{s}\big)\Big)\\
&= \ell^{(\lambda)}(\mathsf{w},\mathsf{s}).
\end{aligned}\tag{40}$$

Lemma 8 also implies that

$$\begin{aligned}
\ell^{(\lambda)}(\hat{\mathsf{w}},\mathsf{s}) &= \sum_{k=1}^{d}\Big(\ell_k(\hat{w}_{1:k},\mathsf{s})-\bar{\ell}_k\big(\hat{w}_{1:(k-1)},\mathsf{s}\big)\Big)\\
&\leq \sum_{k=1}^{d}\left(\ell_k(\hat{w}_{1:k},\mathsf{s})-\min_{w'_k\in\mathcal{W}_k}\ell_k\big(\hat{w}_{1:(k-1)}w'_k,\mathsf{s}\big)+\lambda_k\log|\mathcal{W}_k|\right)\\
&= \sum_{k=1}^{d}\lambda_k\log|\mathcal{W}_k|.
\end{aligned}\tag{41}$$

Let $\bar{\mathsf{W}}$ and $\bar{\mathsf{S}} = \big(\bar{X}_1, \ldots, \bar{X}_n\big)$ be independent copies of W and S, respectively, and assume that $\bar{\mathsf{W}}$ and $\bar{\mathsf{S}}$ are independent from each other. Thus, $\big(\bar{\mathsf{W}}, \bar{\mathsf{S}}\big) \sim P_{\mathsf{W}} P_{\mathsf{S}}$. We have,

$$\mathbb{E}\Big[L_\mu^{(\mathsf{C})}(\mathsf{W})\Big] = \mathbb{E}\Big[\ell^{(\mathsf{C})}(\bar{\mathsf{W}}, \bar{\mathsf{S}})\Big].$$

Based on (20) and (25), for any fixed $\mathsf{w} = (w_1, \ldots, w_d)$ and all $1 \leq k \leq d$, we can write

$$\ell_k(w_{1:k}, x) = \begin{cases} \Big|\gamma_k\Big(h_{k-1}\Big(\frac{x}{\gamma_k}\Big) + f\Big(h_{k-1}\Big(\frac{x}{\gamma_k}\Big), w_k\Big)\Big) - T(x)\Big| & \text{if } x \in \mathcal{X}_k \\ 0 & \text{if } x \notin \mathcal{X}_k, \end{cases}$$

and

$$\ell_k\big(w_{1:(k-1)}\hat{w}_k, x\big) = \begin{cases} \Big|\gamma_k\Big(h_{k-1}\Big(\frac{x}{\gamma_k}\Big) + f\Big(h_{k-1}\Big(\frac{x}{\gamma_k}\Big), \hat{w}_k\Big)\Big) - T(x)\Big| & \text{if } x \in \mathcal{X}_k \\ 0 & \text{if } x \notin \mathcal{X}_k. \end{cases}$$

The triangle inequality implies that, for any $a, b \in \mathbb{R}^m$, we have $||a| - |b|| \leq |a - b|$. Thus, for any $x \in \mathcal{X}$,

$$\begin{aligned} \big|\ell_k(w_{1:k}, x) - \ell_k\big(w_{1:(k-1)}\hat{w}_k, x\big)\big| &\leq \gamma_k \left| f\left(h_{k-1}\left(\frac{x}{\gamma_k}\right), w_k\right) - f\left(h_{k-1}\left(\frac{x}{\gamma_k}\right), \hat{w}_k\right)\right| \\ &\leq \gamma_k\left(\left| f\left(h_{k-1}\left(\frac{x}{\gamma_k}\right), w_k\right)\right| + \left| f\left(h_{k-1}\left(\frac{x}{\gamma_k}\right), \hat{w}_k\right)\right|\right) \\ &\leq 2\gamma_k\rho_k, \end{aligned} \tag{42}$$

where (42) is due to (23). Hence, based on Azuma–Hoeffding's inequality (Lemma 11), for any fixed w and all $1 \leq k \leq d$,

$$\ell_k(w_{1:k}, \mathsf{S}) - \ell_k\big(w_{1:(k-1)}\hat{w}_k, \mathsf{S}\big) = \frac{1}{n}\sum_{i=1}^{n}\big(\ell_k(w_{1:k}, x_i) - \ell_k\big(w_{1:(k-1)}\hat{w}_k, x_i\big)\big)$$

is $4\gamma_k\rho_k/\sqrt{n}$-subgaussian. Since $\bar{W}$ and $\bar{S}$ are independent, $\ell_k\big(\bar{W}_{1:k}, \bar{\mathsf{S}}\big) - \ell_k\big(\bar{W}_{1:(k-1)}\hat{w}_k, \bar{\mathsf{S}}\big)$ is $4\gamma_k\rho_k/\sqrt{n}$-subgaussian as well. Using Lemma 10 for all $1 \leq k \leq d$ and adding up the derived inequalities, we can write

$$\begin{aligned} &\mathbb{E}\Big[\ell^{(\mathsf{C})}(\bar{\mathsf{W}}, \bar{\mathsf{S}})\Big] - \mathbb{E}\Big[\ell^{(\mathsf{C})}(\mathsf{W}, \mathsf{S})\Big] \\ &= \sum_{k=1}^{d}\big(\mathbb{E}\big[\ell_k\big(\bar{W}_{1:k}, \bar{\mathsf{S}}\big) - \ell_k\big(\bar{W}_{1:(k-1)}\hat{w}_k, \bar{\mathsf{S}}\big)\big] - \mathbb{E}\big[\ell_k\big(W_{1:k}, \mathsf{S}\big) - \ell_k\big(W_{1:(k-1)}\hat{w}_k, \mathsf{S}\big)\big]\big) \\ &\leq \sum_{k=1}^{d}\left(\bar{\lambda}_k(\log|\mathcal{W}_1 \times \cdots \times \mathcal{W}_k| - H(W_{1:k}|\mathsf{S})) + \frac{8\gamma_k^2\rho_k^2}{n\bar{\lambda}_k}\right), && (43) \\ &= \sum_{k=1}^{d}\left(\bar{\lambda}_k\left(\sum_{j=1}^{k}\log|\mathcal{W}_j| - H(W_{1:k}|\mathsf{S})\right) + \frac{8\gamma_k^2\rho_k^2}{n\bar{\lambda}_k}\right), && (44) \end{aligned}$$

where we define, for all $1 \le k \le d$,

$$\bar{\lambda}_k := \lambda_k - \lambda_{k+1}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\Big[L_\mu^{(\mathsf{C})}(\mathsf{W})\Big] &= \mathbb{E}\Big[\ell^{(\mathsf{C})}(\bar{\mathsf{W}}, \bar{\mathsf{S}})\Big] \\
&\le \mathbb{E}\Big[\ell^{(\mathsf{C})}(\mathsf{W}, \mathsf{S})\Big] + \sum_{k=1}^{d}\left(\bar{\lambda}_k\left(\sum_{j=1}^{k}\log|\mathcal{W}_j| - H(W_{1:k}|\mathsf{S})\right) + \frac{8\gamma_k^2\rho_k^2}{n\bar{\lambda}_k}\right) && (45)\\
&\le \mathbb{E}\Big[\ell^{(\lambda)}(\mathsf{W}, \mathsf{S})\Big] + \sum_{k=1}^{d}\left(\bar{\lambda}_k\left(\sum_{j=1}^{k}\log|\mathcal{W}_j| - H(W_{1:k}|\mathsf{S})\right) + \frac{8\gamma_k^2\rho_k^2}{n\bar{\lambda}_k}\right) && (46)\\
&\le \mathbb{E}\Big[\ell^{(\lambda)}(\hat{\mathsf{w}}, \mathsf{S})\Big] + \sum_{k=1}^{d}\left(\bar{\lambda}_k\left(\sum_{j=1}^{k}\log|\mathcal{W}_j|\right) + \frac{8\gamma_k^2\rho_k^2}{n\bar{\lambda}_k}\right) && (47)\\
&\le \sum_{k=1}^{d}\left(\bar{\lambda}_k\left(\sum_{j=1}^{k}\log|\mathcal{W}_j|\right) + \lambda_k\log|\mathcal{W}_k| + \frac{8\gamma_k^2\rho_k^2}{n\bar{\lambda}_k}\right) && (48)\\
&= \sum_{k=1}^{d}\left(2\bar{\lambda}_k\left(\sum_{j=1}^{k}\log|\mathcal{W}_j|\right) + \frac{8\gamma_k^2\rho_k^2}{n\bar{\lambda}_k}\right), && (49)
\end{aligned}$$

where (45) is obtained by rewriting (44), (46) is based on (40), (47) is obtained based on Corollary 30 and by replacing $\mathbb{P}^{\star}_{\mathsf{W}|\mathsf{S}}$ with the conditional distribution $P_{\mathsf{W}|\mathsf{S}} = \delta_{\hat{\mathsf{w}}}$ (the Dirac measure on $\hat{\mathsf{w}}$), (48) is based on (41), and (49) is a simple calculation based on summation by parts. ■

Optimizing the bound in (49) over the values of $(\lambda_1, \ldots, \lambda_d)$ gives the following result:

Corollary 34 *Assume that $(\lambda_1, \ldots, \lambda_d)$ are chosen such that for all $1 \le k \le d$,*

$$\bar{\lambda}_k = \lambda_k - \lambda_{k+1} = \frac{2\gamma_k\rho_k}{\sqrt{n\left(\sum_{j=1}^{k}\log|\mathcal{W}_j|\right)}}. \qquad (50)$$

Then, the right side of (49) is minimized with respect to $(\lambda_1, \ldots, \lambda_d)$. In this case, the bound simplifies to the following form:

$$\mathbb{E}\Big[L_\mu^{(\mathsf{C})}(\mathsf{W})\Big] \le \frac{8}{\sqrt{n}}\sum_{k=1}^{d}\gamma_k\rho_k\sqrt{\sum_{j=1}^{k}\log|\mathcal{W}_j|}. \qquad (51)$$

5.2 Bounding the Statistical Risk Based on the Chained Risk

Until now, the analysis did not require any restrictions on the data instance distribution μ. However, we could derive an upper bound only on the expected chained risk of the training mechanism. We also observed that the chained risk is always smaller than or equal to the statistical risk. In this subsection, we show that if μ is scale-invariant, a small chained risk can imply a small statistical risk.

Assume that the instance distribution μ is scale-invariant: If $q(x)$ denotes the density function of μ, then there exists a shape parameter $\alpha > 0$ such that for all $x \in \mathcal{X}$ such that $x/\gamma \in \mathcal{X}$,

$$q\left(\frac{x}{\gamma}\right) = \gamma^{\alpha} q(x). \tag{52}$$

Thus, scale-invariant distributions have homogenous density functions. One particular example of such distributions is the following power-law probability density function with shape parameter α:

$$q(x) = \frac{1}{C_q |x|^{\alpha}} \text{ for all } x \in \mathcal{X},$$

where $C_q = \int_{\mathcal{X}} |x|^{-\alpha} \mathrm{d}x$. Given this assumption, in the following result, we show an upper bound on the statistical risk based on the chained risk:

Theorem 35 *If μ is a scale-invariant probability distribution with shape parameter α defined on $\mathcal{X} \subset \mathbb{R}^m$, then for any $\mathsf{w} \in \mathcal{W}$, we have*

$$\left(1 - \gamma^{m+1-\alpha}\left(1 + C_1 R(1-\gamma^{-1})\right)\right) L_{\mu}(\mathsf{w}) \leq L_{\mu}^{(\mathsf{C})}(\mathsf{w}).$$

Proof For ease of presentation, let $\hat{h}_k(x)$ denote the kth level of the model given parameters $\hat{\mathsf{w}}$ and input x, for any $1 \leq k \leq d$. For any $2 \leq k \leq d$ and any $x \in \mathcal{X}_k$, we have

$$\begin{aligned}
\ell_k\left(w_{1:(k-1)}\hat{w}_k, x\right) &= \left|\gamma_k(\mathrm{id} + \psi_k) \circ h_{k-1}\left(\frac{x}{\gamma_k}\right) - T(x)\right| \\
&= \left|\gamma_k(\mathrm{id} + \psi_k) \circ h_{k-1}\left(\frac{x}{\gamma_k}\right) - \gamma_k(\mathrm{id} + \psi_k) \circ \hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right| \\
&= \gamma_k \left| h_{k-1}\left(\frac{x}{\gamma_k}\right) - \hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right) + \psi_k\left(h_{k-1}\left(\frac{x}{\gamma_k}\right)\right) - \psi_k\left(\hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right)\right| \\
&\leq \gamma_k \left| h_{k-1}\left(\frac{x}{\gamma_k}\right) - \hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right| + \gamma_k \left|\psi_k\left(h_{k-1}\left(\frac{x}{\gamma_k}\right)\right) - \psi_k\left(\hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right)\right| \\
&\leq \gamma_k \left| h_{k-1}\left(\frac{x}{\gamma_k}\right) - \hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right| \\
&\quad + C_1(\gamma - 1)\gamma^{k-d-1}\gamma_k \left| h_{k-1}\left(\frac{x}{\gamma_k}\right) - \hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right| \qquad (53) \\
&= \gamma_k\left(1 + C_1(\gamma-1)\gamma^{k-d-1}\right)\left| h_{k-1}\left(\frac{x}{\gamma_k}\right) - \hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right|,
\end{aligned}$$

where in (53), we used the fact that $\|\psi_k\|_{\mathrm{Lip}} \leq C_1(\gamma - 1)\gamma^{k-d-1}$, according to (21). Now, define

$$x' := \frac{\gamma_{k-1}}{\gamma_k} x = \frac{x}{\gamma}.$$

We observe that $x' \in \mathcal{X}_{k-1}$, and that the transformation $x \leftrightarrow x'$ establishes a bijection between $\mathcal{X}_k$ and $\mathcal{X}_{k-1}$. Therefore, we obtain

$$\begin{aligned}
\ell_k\big(w_{1:(k-1)}\hat{w}_k, x\big) &\leq \gamma_k\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\left|h_{k-1}\left(\frac{x}{\gamma_k}\right) - \hat{h}_{k-1}\left(\frac{x}{\gamma_k}\right)\right| \\
&= \gamma_k\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\left|h_{k-1}\left(\frac{x'}{\gamma_{k-1}}\right) - \hat{h}_{k-1}\left(\frac{x'}{\gamma_{k-1}}\right)\right| \\
&= \frac{\gamma_k}{\gamma_{k-1}}\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\left|\gamma_{k-1}h_{k-1}\left(\frac{x'}{\gamma_{k-1}}\right) - \gamma_{k-1}\hat{h}_{k-1}\left(\frac{x'}{\gamma_{k-1}}\right)\right| \\
&= \gamma\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\ell_{k-1}\big(w_{1:(k-1)}, x'\big). \qquad (54)
\end{aligned}$$

Since $x \in \mathbb{R}^m$, the change of variables formula for differentials is $\mathrm{d}x = \gamma^m \mathrm{d}x'$. Let X be distributed according to density $q(x)$. Based on (54), for all $2 \leq k \leq d$, we derive

$$\begin{aligned}
&\mathbb{E}\big[\ell_k\big(w_{1:(k-1)}\hat{w}_k, X\big)\big] \\
&= \int_{x\in\mathcal{X}_k} \ell_k\big(w_{1:(k-1)}\hat{w}_k, x\big)q(x)\mathrm{d}x \\
&\leq \int_{x\in\mathcal{X}_k} \gamma\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\ell_{k-1}\big(w_{1:(k-1)}, x'\big)q(x)\mathrm{d}x \\
&= \int_{x\in\mathcal{X}_k} \gamma\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\ell_{k-1}\big(w_{1:(k-1)}, x'\big)\gamma^{-\alpha}q\big(x'\big)\mathrm{d}x \qquad (55) \\
&= \int_{x'\in\mathcal{X}_{k-1}} \gamma^{m+1-\alpha}\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\ell_{k-1}\big(w_{1:(k-1)}, x'\big)q\big(x'\big)\mathrm{d}x' \\
&= \gamma^{m+1-\alpha}\Big(1 + C_1(\gamma-1)\gamma^{k-d-1}\Big)\mathbb{E}\big[\ell_{k-1}\big(w_{1:(k-1)}, X\big)\big] \\
&\leq \gamma^{m+1-\alpha}\big(1 + C_1(\gamma-1)\gamma^{-1}\big)\mathbb{E}\big[\ell_{k-1}\big(w_{1:(k-1)}, X\big)\big], \qquad (56)
\end{aligned}$$

where (55) is due to the scale invariance property (52). We can extend inequality (56) to the case $k = 1$ by defining $\ell_0 \equiv 0$. Thus,

$$\begin{aligned}
L_\mu^{(\mathsf{C})}(\mathsf{w}) &= \sum_{k=1}^{d}\big(\mathbb{E}[\ell_k(w_{1:k}, X)] - \mathbb{E}\big[\ell_k\big(w_{1:(k-1)}\hat{w}_k, X\big)\big]\big) \\
&\geq \sum_{k=1}^{d}\big(\mathbb{E}[\ell_k(w_{1:k}, X)] - \gamma^{m+1-\alpha}\big(1 + C_1(\gamma-1)\gamma^{-1}\big)\mathbb{E}\big[\ell_{k-1}\big(w_{1:(k-1)}, X\big)\big]\big) \qquad (57) \\
&= \mathbb{E}\left[\sum_{k=1}^{d}\ell_k(w_{1:k}, X)\right] - \gamma^{m+1-\alpha}\big(1 + C_1(\gamma-1)\gamma^{-1}\big)\mathbb{E}\left[\sum_{k=1}^{d}\ell_{k-1}\big(w_{1:(k-1)}, X\big)\right] \\
&\geq \mathbb{E}\left[\sum_{k=1}^{d}\ell_k(w_{1:k}, X)\right] - \gamma^{m+1-\alpha}\big(1 + C_1(\gamma-1)\gamma^{-1}\big)\mathbb{E}\left[\sum_{k=1}^{d}\ell_k(w_{1:k}, X)\right] \\
&= \big(1 - \gamma^{m+1-\alpha}\big(1 + C_1(\gamma-1)\gamma^{-1}\big)\big)L_\mu(\mathsf{w}),
\end{aligned}$$

where (57) is based on (56). ■

Theorem 35 in conjunction with Theorem 33 or Corollary 34 yields an upper bound on the expected statistical risk of multiscale entropy-based training, as follows:

Corollary 36 *If the shape parameter α of density q satisfies*

$$\alpha > m + 1 + \frac{\log\big(1 + C_1(\gamma - 1)\gamma^{-1}\big)}{\log \gamma}, \tag{58}$$

then the following upper bound on the expected statistical risk holds:

$$\mathbb{E}[L_\mu(\mathsf{W})] \leq \frac{2\sum_{k=1}^{d}\Big((\lambda_k - \lambda_{k+1})\Big(\sum_{j=1}^{k}\log|\mathcal{W}_j|\Big) + \frac{4\gamma_k^2\rho_k^2}{n(\lambda_k - \lambda_{k+1})}\Big)}{1 - \gamma^{m+1-\alpha}(1 + C_1(\gamma - 1)\gamma^{-1})}. \tag{59}$$

If the hyperparameters $(\lambda_1, \dots, \lambda_d)$ are chosen as in (50), then (59) reduces to

$$\mathbb{E}[L_\mu(\mathsf{W})] \leq \frac{8}{\sqrt{n}}\left(\frac{\sum_{k=1}^{d}\gamma_k\rho_k\sqrt{\sum_{j=1}^{k}\log|\mathcal{W}_j|}}{1 - \gamma^{m+1-\alpha}(1 + C_1(\gamma - 1)\gamma^{-1})}\right). \tag{60}$$

Proof Inequality (58) is equivalent to $1 - \gamma^{m+1-\alpha}\big(1 + C_1(\gamma - 1)\gamma^{-1}\big) > 0$, which, in conjunction with Theorem 35 and Theorem 33 or Corollary 34, yields the results. ■

Given the realizability assumption on the hypothesis set, the regular union bound applied to the empirical-risk-minimizing hypothesis yields

$$\mathbb{E}[L_\mu(\mathsf{W}_{\mathrm{ERM}})] \leq \frac{\Big(\sum_{k=1}^{d}\rho_k\Big)}{\sqrt{n}}\sqrt{\sum_{j=1}^{d}\log|\mathcal{W}_j|}. \tag{61}$$

Even when ignoring the effect of $\big(1 - \gamma^{m+1-\alpha}\big(1 + C_1(\gamma - 1)\gamma^{-1}\big)\big)/8$, the right side of (60) can be quite smaller than the right side of (61). Consider the following example:

Example 2 *Recall that $\gamma_0 := R/\varepsilon$, $\gamma = (\gamma_0)^{1/d}$ and $\gamma_k = \gamma^{k-d}$, for all $1 \leq k \leq d$, as in* (18). *Assume that $\rho_k = \rho_0\gamma^k$ for all $1 \leq k \leq d$ and $|\mathcal{W}_1| = \cdots = |\mathcal{W}_d|$. We compute the following ratio*

$$\Lambda = \left(\frac{\sum_{k=1}^{d}\gamma_k\rho_k\sqrt{\sum_{j=1}^{k}\log|\mathcal{W}_j|}}{\Big(\sum_{k=1}^{d}\rho_k\Big)\sqrt{\sum_{j=1}^{d}\log|\mathcal{W}_j|}}\right)^2 = \left(\frac{\sum_{k=1}^{d}\gamma^{2k-d}\sqrt{k}}{\sum_{k=1}^{d}\gamma^k\sqrt{d}}\right)^2.$$

The power of two exists to compare the bounds on the required number of samples n. For example, given $\gamma_0 = 10$ and $d = 20$, we obtain $\Lambda \approx 0.2648$.

6. Bounded-Norm Parameterization Example

In this section, for the case of one-dimensional functions ($m = 1$), we show that any (M_1, M_2)-diffeomorphism defined on $(-R, R)$ can be represented with a hierarchical parameterized model with bounded-norm parameters of the form (20). The approach to proving this is by constructing the ladder decomposition of the diffeomorphism. We then proceed to discretize the parameters of this model and apply the bound of Corollary 36 to the corresponding hypothesis set.

Assume that the target function T is a (M_1, M_2)-diffeomorphism with $T(0) = 0$, and consider its ladder decomposition at scale parameters $\{\gamma_k\}_{k=1}^d$. The range of $T_{[\gamma_{k-1}]}$, which is the domain of ψ_k, is an interval subset of $(-M_1R, M_1R)$ that includes 0. Note that for a (M_1, M_2)-diffeomorphism, since the function is M_1-bilipschitz, it is easy to observe that $M_1 \geq 1$. Therefore $(-M_1R, M_1R)$ includes $(-R, R)$.

Let $\Psi : \mathbb{R} \to \mathbb{R}$ be an arbitrary invertible ϕ_1-Lipschitz and ϕ_2-smooth function with support $D_\Psi = (a_1, a_2) \subseteq (-M_1R, M_1R)$ such that $0 \in D_\Psi$ (that is, $a_1 < 0 < a_2$) and $\Psi(0) = 0$. We will later replace Ψ with each function ψ_k of the ladder decomposition of the target function T. For any $x \in D_\Psi$, we can write

$$\begin{aligned}\Psi(x) &= \int_{a_1}^{x} \Psi'(b)\mathrm{d}b + \Psi(a_1)\\ &= \int_{a_1}^{a_2} \Psi'(b)\Theta(x-b)\mathrm{d}b + \Psi(a_1),\end{aligned} \tag{62}$$

where $\Theta(x)$ is the Heaviside (unit) step function. Since Ψ is ϕ_1-Lipschitz, for all $b \in D_\Psi$, we have $|\Psi'(b)| \leq \phi_1$. Moreover, since $\Psi(0) = 0$, we conclude that $|\Psi(a_1)| \leq \phi_1|a_1| \leq \phi_1 M_1 R$.

The idea is now to derive the Riemann sum approximation to the integral representation of $\Psi(x)$ in (62) and to view it as a two-layer neural network. For a given integer $\tau \geq 2$, let a τ-width two-layer network with *continuous* parameters $w := \{w_1, \ldots, w_\tau, w^{(c)}\}$ be defined for any $x \in (-M_1R, M_1R)$ as

$$\bar{\psi}_w^{(\tau)}(x) := \sum_{j=1}^{\tau} w^{(j)}\Theta(x - b_j) + w^{(c)}. \tag{63}$$

For all $1 \leq j \leq \tau$, we set

$$b_j := \left(\frac{2j}{\tau} - 1\right) M_1 R.$$

Clearly, $\{b_j\}_{j=1}^{\tau}$ is an arithmetic progression with common difference $\Delta := 2M_1R/\tau$, where $b_1 \approx -M_1R$ and $b_\tau = M_1R$. We allow the parameters of (63) to take values

$$\begin{cases} w^{(j)} \leftarrow \Delta\Psi'(b_j) & \text{if } b_j \in D_\Psi \\ w^{(j)} \leftarrow 0 & \text{if } b_j \notin D_\Psi \end{cases}$$

and

$$w^{(c)} \leftarrow \Psi(a_1).$$

Using the Riemann sum approximation bound, given as Lemma 12 in Section 2, we deduce the following result:

Lemma 37 *For all $x \in D_\Psi$, we have*

$$\left|\bar{\psi}_w^{(\tau)}(x) - \Psi(x)\right| \leq \frac{6M_1R}{\tau}\phi_1 + \frac{2(M_1R)^2}{\tau}\phi_2.$$

Proof Recall that since Ψ is ϕ_1-Lipschitz, for all $b \in D_\Psi$, we have $|\Psi'(b)| \leq \phi_1$. If $x - a_1 < \Delta$, then $\bar{\psi}_w^{(\tau)}(x) = \Psi(a_1)$. Based on the triangle inequality, we have

$$\left|\bar{\psi}_w^{(\tau)}(x) - \Psi(x)\right| = \left|\int_{a_1}^{x} \Psi'(b)\mathrm{d}b\right| \leq \int_{a_1}^{x} \left|\Psi'(b)\right|\mathrm{d}b < \Delta\phi_1 = \frac{2M_1R}{\tau}\phi_1,$$

which implies the result. Therefore, assume that $x - a_1 \geq \Delta$, and let j_1 and j_2 be the smallest and largest integer j such that $a_1 \leq b_j \leq x$, respectively. We have

$$\begin{aligned}
\left|\bar{\psi}_w^{(\tau)}(x) - \Psi(x)\right| &= \left|\sum_{j=1}^{\tau} w^{(j)}\Theta(x - b_j) - \int_{a_1}^{a_2} \Psi'(b)\Theta(x - b)\mathrm{d}b\right| \\
&= \left|\sum_{j=j_1}^{j_2} \Delta\Psi'(b_j) - \int_{a_1}^{x} \Psi'(b)\mathrm{d}b\right| \\
&= \left|\sum_{j=j_1}^{j_2} \Delta\Psi'(b_j) - \int_{b_{j_1}}^{b_{j_2}} \Psi'(b)\mathrm{d}b - \int_{a_1}^{b_{j_1}} \Psi'(b)\mathrm{d}b - \int_{b_{j_2}}^{x} \Psi'(b)\mathrm{d}b\right| \\
&\leq \left|\sum_{j=j_1}^{j_2} \Delta\Psi'(b_j) - \int_{b_{j_1}}^{b_{j_2}} \Psi'(b)\mathrm{d}b\right| + 2\Delta\phi_1 && (64) \\
&\leq \left|\sum_{j=j_1}^{j_2-1} \Delta\Psi'(b_j) - \int_{b_{j_1}}^{b_{j_2}} \Psi'(b)\mathrm{d}b\right| + 3\Delta\phi_1, && (65)
\end{aligned}$$

where, (64) and (65) are based on the triangle inequality. If $j_2 = j_1$, then

$$\left|\sum_{j=j_1}^{j_2-1} \Delta\Psi'(b_j) - \int_{b_{j_1}}^{b_{j_2}} \Psi'(b)\mathrm{d}b\right| = 0,$$

which implies the result. Otherwise, we have $j_2 > j_1$. In this case, we obtain

$$\begin{aligned}
\left|\sum_{j=j_1}^{j_2-1} \Delta\Psi'(b_j) - \int_{b_{j_1}}^{b_{j_2}} \Psi'(b)\mathrm{d}b\right| &\leq \frac{\phi_2(b_{j_2} - b_{j_1})^2}{2(j_2 - j_1)} && (66) \\
&= \frac{\phi_2(j_2 - j_1)^2\Delta^2}{2(j_2 - j_1)} \\
&= \frac{\phi_2}{2}(j_2 - j_1)\Delta^2 \\
&\leq \frac{\phi_2}{2}\left(\frac{2M_1R}{\Delta}\right)\Delta^2 && (67) \\
&= M_1R\Delta\phi_2 \\
&= \frac{2(M_1R)^2}{\tau}\phi_2, && (68)
\end{aligned}$$

where (66) is based on Lemma 12, and (67) is based on the fact that $(j_2 - j_1)\Delta \le (b_\tau - b_1) \le 2M_1R$. The proof of the statement is now complete. ■

We proceed to discretize the parameters of the network. Let $\eta > 0$ be the precision level of the parameters. For any $y \in \mathbb{R}$, we denote the closest real number in $\eta\mathbb{Z} = \{\dots, -2\eta, -\eta, 0, \eta, 2\eta, \dots\}$ to y with $[y]_\eta$. For any function Ψ, the approximate τ-width two-layer network with *discretized* parameters at precision level η is

$$\psi_w^{(\tau,\eta)}(x) := \sum_{j=1}^{\tau} \Big[w^{(j)}\Big]_\eta \Theta(x - b_j) + \Big[w^{(c)}\Big]_\eta. \tag{69}$$

We define the norm of the network $\psi_w^{(\tau,\eta)}$ as the sum of the absolute values of its parameters:

$$\operatorname{norm}\Big(\psi_w^{(\tau,\eta)}\Big) := \sum_{j=1}^{\tau} \Big|\Big[w^{(j)}\Big]_\eta\Big| + \Big|\Big[w^{(c)}\Big]_\eta\Big|.$$

The next result studies the approximation error, the output norm, and the norm of the network of $\psi_w^{(\tau,\eta)}$.

Proposition 38 *The two-layer network $\psi_w^{(\tau,\eta)}$, approximating function Ψ, satisfies the following properties:*

(a) For all $x \in D_\Psi$,

$$\Big|\psi_w^{(\tau,\eta)}(x) - \Psi(x)\Big| \le (\tau+1)\frac{\eta}{2} + \frac{6M_1R}{\tau}\phi_1 + \frac{2(M_1R)^2}{\tau}\phi_2.$$

(b) For all $x \in (-M_1R, M_1R)$,

$$\Big|\psi_w^{(\tau,\eta)}(x)\Big| \le \operatorname{norm}\Big(\psi_w^{(\tau,\eta)}\Big).$$

(c) We have

$$\operatorname{norm}\Big(\psi_w^{(\tau,\eta)}\Big) \le (\tau+1)\frac{\eta}{2} + \left(3M_1R + \frac{2M_1R}{\tau}\right)\phi_1 + \frac{2(M_1R)^2}{\tau}\phi_2 =: \varrho(\phi_1,\phi_2). \tag{70}$$

Proof

(a) For all $x \in D_\Psi$, we have

$$\begin{aligned}
\Big|\psi_w^{(\tau,\eta)}(x) - \Psi(x)\Big| &= \Big|\psi_w^{(\tau,\eta)}(x) - \bar\psi_w^{(\tau)}(x) + \bar\psi_w^{(\tau)}(x) - \Psi(x)\Big| \\
&\le \Big|\psi_w^{(\tau,\eta)}(x) - \bar\psi_w^{(\tau)}(x)\Big| + \Big|\bar\psi_w^{(\tau)}(x) - \Psi(x)\Big| \\
&\le \sum_{j=1}^{\tau}\Big|w^{(j)} - \Big[w^{(j)}\Big]_\eta\Big| + \Big|w^{(c)} - \Big[w^{(c)}\Big]_\eta\Big| + \frac{6M_1R}{\tau}\phi_1 + \frac{2(M_1R)^2}{\tau}\phi_2 \qquad (71) \\
&\le (\tau+1)\frac{\eta}{2} + \frac{6M_1R}{\tau}\phi_1 + \frac{2(M_1R)^2}{\tau}\phi_2,
\end{aligned}$$

where (71) is based on Lemma 37 and the triangle inequality.

(b) Based on the triangle inequality,

$$\begin{aligned}\left|\psi_w^{(\tau,\eta)}(x)\right| &= \left|\sum_{j=1}^{\tau}\left[w^{(j)}\right]_\eta \Theta(x-b_j) + \left[w^{(c)}\right]_\eta\right| \\ &\leq \sum_{j=1}^{\tau}\left|\left[w^{(j)}\right]_\eta\right| + \left|\left[w^{(c)}\right]_\eta\right| \\ &= \mathrm{norm}\left(\psi_w^{(\tau,\eta)}\right).\end{aligned}$$

(c) We have

$$\begin{aligned}\mathrm{norm}\left(\psi_w^{(\tau,\eta)}\right) &= \sum_{j=1}^{\tau}\left|\left[w^{(j)}\right]_\eta\right| + \left|\left[w^{(c)}\right]_\eta\right| \\ &\leq \sum_{j=1}^{\tau}\left|w^{(j)}\right| + \left|w^{(c)}\right| + (\tau+1)\frac{\eta}{2} \\ &\leq \sum_{j=1}^{\tau}\left|w^{(j)}\right| + M_1R\phi_1 + (\tau+1)\frac{\eta}{2}, \qquad (72)\end{aligned}$$

where (72) is based on $\left|w^{(c)}\right| = |\Psi(a_1)| \leq M_1R\phi_1$. If $a_2 - a_1 < \Delta$, then $\sum_{j=1}^{\tau}\left|w^{(j)}\right| = 0$ and the result is proven. Thus, assume that $a_2 - a_1 \geq \Delta$, and let j_1 and j_2 be the smallest and largest integer j such that $a_1 \leq b_j \leq a_2$, respectively. If $j_1 = j_2$, then $\sum_{j=1}^{\tau}\left|w^{(j)}\right| = |\Delta\Psi'(b_{j_1})| \leq \Delta\phi_1$, which implies the result. Therefore, assume that $j_2 > j_1$. Since $\Psi(x)$ is ϕ_2-smooth, $\Psi'(x)$ is ϕ_2-Lipschitz. Thus, $|\Psi'(x)|$ is ϕ_2-Lipschitz as well. Moreover, due to the fact that Ψ is invertible, $|\Psi'(x)|$ is either identical to $\Psi'(x)$ or $-\Psi'(x)$, thus is differentiable. We derive

$$\begin{aligned}\sum_{j=1}^{\tau}\left|w^{(j)}\right| &= \sum_{j=j_1}^{j_2}\left|w^{(j)}\right| \\ &= \sum_{j=j_1}^{j_2}\left|\Delta\Psi'(b_j)\right| \\ &\leq \sum_{j=j_1}^{j_2-1}\left|\Delta\Psi'(b_j)\right| + \Delta\phi_1 \\ &\leq \int_{b_{j_1}}^{b_{j_2}} |\Psi'(b)|\mathrm{d}b + \frac{\phi_2(b_{j_2}-b_{j_1})^2}{2(j_2-j_1)} + \Delta\phi_1 \qquad (73) \\ &\leq \int_{b_{j_1}}^{b_{j_2}} |\Psi'(b)|\mathrm{d}b + \frac{2(M_1R)^2}{\tau}\phi_2 + \Delta\phi_1 \qquad (74) \\ &\leq (b_{j_2}-b_{j_1})\phi_1 + \frac{2(M_1R)^2}{\tau}\phi_2 + \Delta\phi_1 \\ &\leq (2M_1R+\Delta)\phi_1 + \frac{2(M_1R)^2}{\tau}\phi_2,\end{aligned}$$

where (73) is due to Lemma 12 and (74) is based on the same derivation of (68) from the right side of (66). The result is now proven. ■

Now, consider the ladder decomposition of the target function T at scale parameters $\{\gamma_k\}_{k=1}^{d}$. For any $1 \leq k \leq d$, we replace function ψ_k with the function Ψ in Proposition 38. Let $C_2 := (M_1^2 + M_1)M_2$. Based on Theorem 19, ψ_k is $C_1\gamma^{k-d-1}(\gamma-1)$-Lipschitz and C_2-smooth. Thus, we take $\phi_1 \leftarrow C_1\gamma^{k-d-1}(\gamma-1)$ and $\phi_2 \leftarrow C_2$. For all $1 \leq k \leq d$, define

$$\rho_k := \varrho\Big(C_1\gamma^{k-d-1}(\gamma-1), C_2\Big),$$

where the function ϱ is defined in (70). Note that, as function of k, $\rho_k = O(\gamma^k)$, conforming with (22). Suppose $\mathcal{W}_k$, the set of parameters for our learning model at level k, is

$$\mathcal{W}_k := \Big\{w = \Big(w^{(1)}, \ldots, w^{(\tau)}, w^{(c)}\Big) \in \eta\mathbb{Z}^{\tau+1} : |w|_1 \leq \rho_k\Big\}.$$

In (20), the recursive definition of the model, we define

$$f(h_{k-1}(x); w_k) := \psi_{w_k}^{(\tau,\eta)}(h_{k-1}(x)),$$

for all $1 \leq k \leq d$. Therefore, the kth level of our learning model is the following function:

$$h_k = \Big(\psi_{w_k}^{(\tau,\eta)} + \mathrm{id}\Big) \circ \cdots \circ \Big(\psi_{w_2}^{(\tau,\eta)} + \mathrm{id}\Big) \circ \Big(\psi_{w_1}^{(\tau,\eta)} + \mathrm{id}\Big).$$

Notice that since $\mathcal{W}_k$ is a discretization of the ℓ_1-ball $\{|w|_1 \leq \rho_k\}$, it is a finite set. The next result finds the cardinality of $\mathcal{W}_k$:

Proposition 39 *For all $1 \leq k \leq d$, we have*

$$|\mathcal{W}_k| = \sum_{r=0}^{\tau} 2^{\tau+1-r}\binom{\tau+1}{r}\binom{\lfloor \rho_k/\eta \rfloor}{\tau+1-r}.$$

Proof For all $1 \leq k \leq d$, suppose

$$\mathcal{W}_k' := \Big\{w \in \mathbb{Z}^{\tau+1} : |w|_1 \leq \Big\lfloor \frac{\rho_k}{\eta} \Big\rfloor\Big\}.$$

Based on the one-to-one correspondence $w \leftrightarrow \eta w$, we have $|\mathcal{W}_k| = |\mathcal{W}_k'|$. We will employ a counting argument to find $|\mathcal{W}_k'|$. Assume that r components of $w \in \mathcal{W}_k'$ are equal to zero, and the rest are non-zero, where $0 \leq r \leq \tau$. There are $\binom{\tau+1}{r}$ ways to choose the r zero components. We now count the number of configurations that $\tau+1-r$ *positive* integers have a sum less than or equal to $\lfloor \rho_k/\eta \rfloor$. For each such configuration, there are $2^{\tau+1-r}$ ways to make each component retain its positive value or negate it.

By adding a slack variable, it is easy to observe that the number of configurations that $\tau+1-r$ positive integers have sum less than or equal to $\lfloor \rho_k/\eta \rfloor$ is equal to the number of configurations that $(\tau+1-r)+1$ positive integers have sum *equal* to $\lfloor \rho_k/\eta \rfloor + 1$. Based on a basic result in combinatorics (Niven, 1965, (4.7) in Section 4.2), this number is equal to

$$\binom{\lfloor \rho_k/\eta \rfloor}{\tau+1-r}.$$

Therefore,

$$|\mathcal{W}_k| = |\mathcal{W}'_k| = \sum_{r=0}^{\tau} 2^{\tau+1-r} \binom{\tau+1}{r} \binom{\lfloor \rho_k/\eta \rfloor}{\tau+1-r}.$$

■

Corollary 40 *Based on Proposition 39, the optimized bound on the statistical risk in Corollary 36 for the example studied in this section is as follows:*

$$\begin{aligned}\mathbb{E}[L_\mu(\mathsf{W})] &\leq \frac{8}{\sqrt{n}} \left(\frac{\sum_{k=1}^{d} \gamma_k \rho_k \sqrt{\sum_{j=1}^{k} \log |\mathcal{W}_j|}}{1-\gamma^{m+1-\alpha}(1+C_1(\gamma-1)\gamma^{-1})} \right) \\ &= \frac{8}{\sqrt{n}} \left(\frac{\sum_{k=1}^{d} \gamma^{k-d} \rho_k \sqrt{\sum_{j=1}^{k} \log\Big(\sum_{r=0}^{\tau} 2^{\tau+1-r} \binom{\tau+1}{r} \binom{\lfloor \rho_j/\eta \rfloor}{\tau+1-r} \Big)}}{1-\gamma^{m+1-\alpha}(1+C_1(\gamma-1)\gamma^{-1})} \right),\end{aligned}$$

where $\rho_k = \varrho\big(C_1 \gamma^{k-d-1}(\gamma-1), C_2\big)$*, for all* $1 \leq k \leq d$*.*

7. Conclusions

In this paper, we introduced an entropy-based hierarchical learning model designed to leverage the multiscale structure of data instance distributions and the smoothness inherent in real-world target functions. We started with the definition of ladder decompositions of invertible functions, followed by an examination of Lipschitz continuity and smoothness in the components of this decomposition for smooth functions. Subsequently, we analyzed the effectiveness of the proposed multiscale entropy-based training, demonstrating its capability to achieve low chained risk. Notably, where the data distribution μ is scale-invariant with a sufficiently large shape parameter, this paper showed an upper bound on the statistical risk based on the chained risk. Consequently, we derived a guarantee on the statistical risk of our training mechanism for these data distributions. Finally, for the specific case of one-dimensional functions, the paper provided an illustrative example of a parameterized model featuring bounded-norm parameters to showcase a simple application of our methodology.

Our proposed learning model offers several noteworthy advantages. Firstly, it adopts a hierarchical structure with interpretable levels. Each level in the model is trained to approximate a dilation of the target function, offering interpretable roles for different levels, which contrasts opaque black-box hierarchical models.

Secondly, our model introduces a new perspective on data instance complexity. The logarithm of the norm of data instances serves as a measure of their individual complexity. The training procedure can also be conceptualized as a mathematical model for curriculum learning.

Another merit of the proposed model is the computational point of view. The amount of computation required to compute the output of the learned model for a given data instance is directly proportional to the complexity of that instance. Given the heterogeneous distribution of data instances at different scales and complexities, this characteristic may translate into significant computational savings and faster inference speeds.

Furthermore, our model's statistical analysis accounts for its hierarchical and multiscale structure, leading to sharper bounds on the statistical risk for scale-invariant distributions compared to uniform convergence bounds for empirical-risk-minimizing mechanisms.

Additionally, the multi-staged nature of our training mechanism can be beneficial when dealing with the challenge of extensive training times on massive datasets. Even if the training terminates prematurely, the mechanism ensures a useful model capable of accurately predicting the labels of data instances with norms smaller than a specific threshold, depending on the terminated stage.

Finally, our learning model proves advantageous in scenarios involving multiple users learning the target function at a particular scale of data instances. It provides computational savings by using each user's learned model as prior information to train the model for the following user.

Our current work has limitations in that the derived risk bounds are applied only when the parameters of our model are discrete and the hypothesis set is finite. Section 6 provided an example of parameterization where diffeomorphisms are expressed using parameters with bounded norms. Consequently, the hypothesis set becomes finite when these parameters are discretized due to, for instance, real-world memory constraints. While the multiscale entropy-based training procedure (Algorithm 1) remains well-defined in the scenario where the hypothesis set is infinite, an extension of our technique is needed to analyze the statistical risk comprehensively. It is plausible that leveraging the Lipschitz property of the hypothesis set, as demonstrated in the derivation of the risk bound on the classical Gibbs distribution with infinite hypothesis set in (Xu and Raginsky, 2017), could prove beneficial. We defer the exploration of this direction to our future investigations.

Acknowledgments

I would like to express my sincere appreciation to the anonymous reviewers for their constructive comments and suggestions, which significantly enriched the quality of this manuscript. This research was supported by Leverhulme Trust grant ECF-2023-189 and Isaac Newton Trust grant 23.08(b).

References

Amir R. Asadi and Emmanuel Abbe. Chaining meets chain rule: Multilevel entropic regularization and training of neural networks. *Journal of Machine Learning Research*, 21 (139):1–32, 2020.

Amir R. Asadi and Po-Ling Loh. Entropic regularization of neural networks: Self-similar approximations. *Journal of Statistical Planning and Inference*, 233:106181, 2024.

Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, pages 7234–7243, 2018.

Jean-Yves Audibert and Olivier Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8(Apr):863–889, 2007.

Peter L. Bartlett, Steven N. Evans, and Philip M. Long. Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv preprint arXiv:1804.05012*, 2018.

Peter Baxandall and Hans Liebeck. *Vector Calculus.* New York: Oxford University Press, 1986.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 41–48, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening mutual information based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 2020.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. ISSN 00361445, 10957200.

Eugenio Clerico, Amitis Shidani, George Deligiannidis, and Arnaud Doucet. Chained generalisation bounds. In *Proceedings of Machine Learning Research*, pages 4212 — 4212, 2022.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 2012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Weinan E. *Principles of Multiscale Modeling.* Cambridge University Press, 2011.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Alastair Fletcher and Vladimir Markovic. Decomposing diffeomorphisms of the sphere. *Bulletin of the London Mathematical Society*, 44(3):599–609, 2012.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Deborah Hughes-Hallett, Andrew M. Gleason, and William G. McCallum. *Calculus: Single and Multivariable*. John Wiley & Sons, 2020.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4): 620, 1957.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational Learning Theory*, pages 164–170, 1999.

Shahar Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008.

Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2), 2004.

M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

Ivan Niven. *Mathematics of Choice: Or, How to Count Without Counting*, volume 15. MAA, 1965.

John P. Nolan. *Univariate Stable Distributions*. Springer, 2020.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop*, pages 26–30. IEEE, 2016.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Gennady Samoradnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Routledge, 2017.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Didier Sornette. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools*. Springer Science & Business Media, 2006.

Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, volume 60. Springer Science & Business Media, 2014.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Science & Business Media, 1999.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.